Special Matrices and Operations Useful in Modeling and Data Analysis

In previous chapters, we defined a number of special matrices, such as symmetric matrices, banded matrices, elementary operator matrices, and so on. In this chapter, we will discuss some of these matrices in more detail and also introduce some other special matrices and data structures that are useful in statistics.

There are a number of special kinds of matrices that are useful in statistical applications. In statistical applications in which data analysis is the objective, the initial step is the representation of observational data in some convenient form, which often is a matrix. We discuss the representation of observations using matrices in Section 8.1. The matrices for operating on observational data or summarizing the data often have special structures and properties.

One of the most important properties of many matrices occurring in statistical data analysis is nonnegative or positive definiteness; this is the subject of Sections 8.3 and 8.4. Fitted values of a response variable that are associated with given values of covariates in linear models are often projections of the observations onto a subspace determined by the covariates. Projection matrices and Gramian matrices useful in linear models are considered in Sections 8.5 and 8.6. One of the most important properties of many matrices occurring in statistical modeling over time is irreducible nonnegativeness or positiveness; this is the subject of Section 8.7.

8.1 Data Matrices and Association Matrices

There are several ways that data can be organized for representation in the computer. We distinguish logical structures from computer-storage structures. Data structure in computers is an important concern and can greatly affect the efficiency of computer processing. We discuss some simple aspects of the organization for computer storage in Section 11.1, beginning on page 429. In the present section, we consider some general issues of logical organization and structure.
There are two important aspects of data in applications that we will not address here. One is metadata; that is, data about the data. Metadata includes names or labels associated with data, information about how and when the data were collected, information about how the data are stored in the computer, and so on. Another important concern in applications is missing data. In real-world applications it is common to have incomplete data. If the data are stored in some structure that naturally contains a cell or a region for the missing data, the computer representation of the dataset must contain some indication that the cell is empty. For numeric data, the convenient way of doing this is by using “not-a-number”, or NaN (see page 386). We briefly discuss issues in handling missing data in Part III. We consider some effects of missing data on the estimation of matrices in Section 9.4.6.

8.1.1 Flat Files

If several features or attributes are observed on each of several entities, a convenient way of organizing the data is as a two-dimensional array with each column corresponding to a specific feature and each row corresponding to a specific observational entity. In the field of statistics, data for the features are stored in “variables”, the entities are called “observational units”, and a row of the array is called an “observation” (see Figure 8.1).

\[
\begin{array}{l|cccc}
\text{Var 1} & \text{Var 2} & \ldots & \text{Var m} \\
\hline
\text{Obs 1} & x & x & \ldots & x \\
\text{Obs 2} & x & x & \ldots & x \\
\vdots & \vdots & \vdots & \ldots & \vdots \\
\text{Obs n} & x & x & \ldots & x \\
\end{array}
\]

**Fig. 8.1.** Data Appropriate for Representation in a Flat File

The data may be various types of objects, such as names, real numbers, numbers with associated measurement units, sets, vectors, and so on. If the data are represented as real numbers, the data array is a matrix. (Note again our use of the word “matrix”; not just any rectangular array is a matrix in the sense used in this book.) Other types of data can often be made equivalent to a matrix in an intuitive manner.

The flat file arrangement emphasizes the relationships of the data both within an observational unit or row and within a variable or column. Simple operations on the data matrix may reveal relationships among observational units or among variables.

8.1.2 Graphs and Other Data Structures

If the numbers of measurements on the observational units varies or if the interest is primarily in simple relationships among observational units or among
variables, the flat file structure may not be very useful. Sometimes a graph structure can be used advantageously.

A **graph** is a nonempty set \( V \) of points, called **vertices**, together with a set \( E \) of unordered pairs of elements of \( V \), called **edges**. (Other definitions of "graph" allow the null set to be a graph.) If we let \( G \) be a graph, we represent it as \((V, E)\). We often represent the set of vertices as \( V(G) \) and the set of edges as \( E(G) \). An edge is said to be **incident** on each vertex in the edge. The number of vertices (that is, the cardinality of \( V \)) is the **order of the graph**, and the number of edges, the cardinality of \( E \), is the **size of the graph**.

An edge in which the two vertices are the same is called a **loop**. (A **simple graph** is sometimes defined as a graph with no loops; that is, one in which each edge contains only pairs of distinct elements.)

A **path** or **walk** is a sequence of edges, \( e_1, \ldots, e_n \), such that for \( i \geq 2 \) one vertex in \( e_i \) is a vertex in edge \( e_{i-1} \). Alternatively, a path or walk is defined as a sequence of vertices with common edges.

A graph such that there is a path that includes any pair of vertices is said to be **connected**.

A graph with more than one vertex such that all possible pairs of vertices occur as edges is a **complete graph**.

A closed path or closed walk is a path such that a vertex in the first edge (or the first vertex in the alternate definition) is in the last edge (or the last vertex).

A **cycle** is a closed path in which all vertices occur exactly twice (or in the alternate definition, in which all vertices except the first and the last are distinct). A graph with no cycles is said to be **acyclic**. An acyclic graph is also called a **tree**. Trees are used extensively in statistics to represent clusters.

The number of edges that contain a given vertex (that is, the number of edges incident on the vertex \( v \)) denoted by \( d(v) \) is the **degree of the vertex**.

A vertex with degree 0 is said to be **isolated**.

We see immediately that the sum of the degrees of all vertices equals twice the number of edges, that is,

\[
\sum d(v_i) = 2\#(E).
\]

The sum of the degrees hence must be an even number.

A **regular graph** is one for which \( d(v_i) \) is constant for all vertices \( v_i \); more specifically, a graph is **k-regular** if \( d(v_i) = k \) for all vertices \( v_i \).

The natural data structure for a graph is a pair of lists, but a graph is often represented graphically (no pun!) as in Figure 8.2, which shows a graph with five vertices seven edges. While a matrix is usually not an appropriate structure for representing raw data from a graph, there are various types of matrices that are useful for studying the data represented by the graph, which we will discuss in Section 8.8.7.
If $G$ is the graph represented in Figure 8.2, the vertices are $V(G) = \{a, b, c, d, e\}$ and the edges are $E(G) = \{(a, b), (a, c), (a, d), (a, e), (b, e), (c, d), (d, e)\}$.

The presence of an edge between two vertices can indicate the existence of a relationship between the objects represented by the vertices. The graph represented in Figure 8.2 may represent five observational units for which our primary interest is in their relationships with one another. For example, the observations may be authors of scientific papers, and an edge between two authors may represent the fact that the two have been coauthors on some paper.

The same information represented in the 5-order graph of Figure 8.2 may be represented in a $5 \times 5$ rectangular array, as in Figure 8.3.

In the graph represented in Figure 8.2, there are no isolated vertices and the graph is connected. (Note that a graph with no isolated vertices is not necessarily connected.) The graph represented in Figure 8.2 is not complete because, for example, there is no edge that contains vertices $c$ and $e$. The graph is cyclic because of the closed path (defined by vertices) $(c, d, e, b, a, c)$. Note that the closed path $(c, d, a, e, b, a, c)$ is not a cycle.

This use of a graph immediately suggests various extensions of a basic graph. For example, $E$ may be a multiset, with multiple instances of edges containing the same two vertices, perhaps, in the example above, representing multiple papers in which the two authors are coauthors. A graph in which $E$ is a multiset is called a multigraph. Instead of just the presence or absence of edges between vertices, a weighted graph may be more useful; that is, one in which a real number is associated with a pair of vertices to represent the strength of the relationship, not just presence or absence, between the two
vertices. A degenerate weighted graph (that is, an unweighted graph as discussed above) has weights of 0 or 1 between all vertices. A multigraph is a weighted graph in which the weights are restricted to nonnegative integers. Although the data in a weighted graph carry much more information than a graph with only its edges, or even a multigraph that allows strength to be represented by multiple edges, the simplicity of a graph sometimes recommends its use even when there are varying degrees of strength of relationships. A standard approach in applications is to set a threshold for the strength of relationship and to define an edge only when the threshold is exceeded.

Adjacency Matrix; Connectivity Matrix

The connections between vertices in the graphs shown in Figure 8.2 or in Figure 8.4 can be represented in an association matrix called an adjacency matrix, a connectivity matrix, or an incidence matrix to represent edges between vertices, as shown in equation (8.1). (The terms “adjacency”, “connectivity”, and “incidence” are synonymous. “Adjacency” is perhaps the most commonly used term, but I will naturally use both that term and “connectivity” because of the connotative value of the latter term.) The graph, \( G \), represented in Figure 8.2 has the symmetric adjacency matrix

\[
C(G) = \begin{bmatrix}
0 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 \\
1 & 1 & 0 & 1 & 0 \\
\end{bmatrix}.
\] (8.1)

We often use this type of notation; a symbol represents a particular graph, and other objects that relate to the graph make use of that symbol. There is no difference in the connectivity matrix and a table such as in Figure 8.3 except for the metadata.

The diagonal elements of the adjacency matrix for a simple graph (one with no loops) are all 0s.

The relationship can obviously be defined in the other direction; that is, given an \( n \times n \) symmetric matrix \( A \), we define the graph of the matrix as the graph with \( n \) vertices and edges between vertices \( i \) and \( j \) if \( a_{ij} \neq 0 \). We often denote the graph of the matrix \( A \) by \( G(A) \).

Generally we restrict the elements of the connectivity matrix to be 1 or 0 to indicate only presence or absence of a connection, but not to indicate strength of the connection. In this case, a connectivity matrix is a nonnegative matrix; that is, all of its elements are nonnegative. We indicate that a matrix \( A \) is nonnegative by

\[ A \geq 0. \]

We discuss the notation and properties of nonnegative (and positive) matrices in Section 8.7.
Digraphs

Another extension of a basic graph is one in which the relationship may not be the same in both directions. This yields a *digraph*, or “directed graph”, in which the edges are ordered pairs called directed edges. The vertices in a digraph have two kinds of degree, an *indegree* and an *outdegree*, with the obvious meanings.

Although in this book we are more interested in statistical relationships, the simplest applications of digraphs are for representing networks. Consider, for example, the digraph represented by the network in Figure 8.4. This is a network with five vertices, perhaps representing cities, and directed edges between some of the vertices. The edges could represent airline connections between the cities; for example, there are flights from $x$ to $u$ and from $u$ to $x$, and from $y$ to $z$, but not from $z$ to $y$.

Figure 8.4 represents a digraph with order 5 (there are five vertices) and size 11 (eleven directed edges). A sequence of edges, $e_1, \ldots, e_n$, constituting a path in a digraph must be such that for $i \geq 2$ the first vertex in $e_i$ is the second vertex in edge $e_{i-1}$. For example, the sequence $x, y, z, w, u, x$ in the graph of Figure 8.4 is a path (in fact, a cycle) but the sequence $x, u, w, z, y, x$ is not a path.

The connectivity matrix for the digraph in Figure 8.4 with nodes ordered as $u, w, x, y, z$ is

$$C = \begin{bmatrix}
0 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0
\end{bmatrix}. \quad (8.2)$$

A connectivity matrix for a (nondirected) graph is symmetric, but for a digraph it is not necessarily symmetric. Given an $n \times n$ matrix $A$, we define the *digraph of the matrix* as the digraph with $n$ vertices and edges from vertex $i$ to $j$ if $a_{ij} \neq 0$. We use the same notation for a digraph as we used above for a graph, $\mathcal{G}(A)$. 

![Fig. 8.4. A Simple Digraph](image-url)
In statistical applications, graphs are used for representing symmetric associations. Digraphs are used for representing asymmetric associations or one-way processes such as a stochastic process.

In a simple digraph, the edges only indicate the presence or absence of a relationship, but just as in the case of a simple graph, we can define a weighted digraph by associating nonnegative numbers with each directed edge.

Graphical modeling is useful type for analyzing relationships between elements of a collection of sets. For example, in an analysis of internet traffic, profiles of users may be constructed based on the set of web sites each user visits in relation to the sets visited by other users. For this kind of application, an intersection graph may be useful. An intersection graph, for a given collection of sets $S$, is a graph whose vertices correspond to the sets in $S$ and whose edges between any two sets have a common element.

The word “graph” is often used without qualification to mean any of these types.

**Connectivity of Digraphs**

There are two kinds of connected digraphs. A digraph such that there is a (directed) path that includes any pair of vertices is said to be strongly connected. A digraph such that there is a path without regard to the direction of any edge that includes any pair of vertices is said to be weakly connected. The digraph shown in Figure 8.4 is strongly connected. The digraph shown in Figure 8.5 is weakly connected but not strongly connected.

A digraph that is not weakly connected must have two sets of nodes with no edges between any nodes in one set and any nodes in the other set.

![Fig. 8.5. A Digraph that Is Not Strongly Connected](image)

The connectivity matrix of the digraph in Figure 8.5 is

$$
C = \begin{bmatrix}
0 & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0
\end{bmatrix}.
$$

(8.3)
The matrix of a digraph that is not strongly connected can always be reduced to a special block upper triangular form by row and column permutations; that is, if the digraph \( \mathcal{G} \) is not strongly connected, then there exists a permutation matrix \( E_\pi \) such that

\[
E_\pi^T A(\mathcal{G}) E_\pi = \begin{bmatrix}
B_{11} & B_{12} \\
0 & B_{22}
\end{bmatrix},
\]

(8.4)

where \( B_{11} \) and \( B_{22} \) are square. Such a transformation is called a symmetric permutation.

Later we will formally prove this relationship between strong connectivity and this reduced form of the matrix, but first we consider the matrix in equation (8.3). If we interchange the second and fourth columns and rows, we get the reduced form

\[
E_{24}^T C E_{24} = \begin{bmatrix}
0 & 0 & 1 & 1 \\
1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1
\end{bmatrix}.
\]

**Irreducible Matrices**

Any nonnegative square matrix that can be permuted into the form in equation (8.4) with square diagonal submatrices is said to be reducible; a matrix that cannot be put into that form is irreducible. We also use the terms reducible and irreducible to refer to the graph itself.

**Strong Connectivity of Digraphs and Irreducibility of Matrices**

A nonnegative matrix is irreducible if and only if its digraph is strongly connected. Stated another way, a digraph is not strongly connected if and only if its matrix is reducible.

To see this, first consider a reducible matrix. In its reduced form of equation (8.4), none of the nodes corresponding to the last rows have directed edges leading to any of the nodes corresponding to the first rows; hence, the digraph is not strongly connected.

Now, assume that a given digraph \( \mathcal{G} \) is not strongly connected. In that case, there is some node, say the \( i^{th} \) node, from which there is no directed path to some other node. Assume that there are \( m - 1 \) nodes that can be reached from node \( i \). If \( m = 1 \), then we have a trivial partitioning of the \( n \times n \) connectivity in which \( B_{11} \) of equation (8.4) is \((n - 1) \times (n - 1)\) and \( B_{22} \) is a \( 1 \times 1 \) 0 matrix (that is, \( 0_1 \)). If \( m \geq 1 \), perform symmetric permutations so that the row corresponding to node \( i \) and all other \( m - 1 \) nodes are the last \( m \) rows of the permuted connectivity matrix. In this case, the first \( n - k \) elements in each of those rows must be 0. To see that this must be the case, let \( k > n - m \) and \( j \leq n - m \) and assume that the element in the \( (k, j) \)th
position is nonzero. In that case, there is a path from node $i$ to node $k$ to node $j$, which is in the set of nodes not reachable from node $i$; hence the $(k,j)^{th}$ element (in the permuted matrix) must be 0. The submatrix corresponding to $B_{11}$ is $n - m \times n - m$, and that corresponding to $B_{22}$ is $m \times m$. These properties also hold for connectivity matrices with simple loops (with 1s on the diagonal) and for an augmented connectivity matrix (see page 314).

Reducibility plays an important role in the analysis of Markov chains (see Section 9.7.1).

### 8.1.3 Probability Distribution Models

Many phenomena are best characterized in terms of a probability distribution. Data in rows of flat files are often assumed to be realizations of vector random variables, some elements of which may have a degenerate distribution (that is, the elements in some columns of the data matrix may be considered to be fixed rather than random). The data in one row are often considered independent of the data in another row. Statistical data analysis is generally concerned with studying various models of relationships among the elements of the vector random variables. For example, the familiar linear regression model relates one variable (one column) to a linear combination of other variables plus a translation and random noise.

A random graph of fixed order is a discrete probability space over all possible graphs of that order. For a graph of order $n$, there are $2^{\binom{n}{2}}$ possible graphs. Asymptotic properties of the probability distribution refer to the increase of the order without limit. Occasionally it is useful to consider the order of the graph to be random also. If the order is unrestricted, the sample space for a random graph of random order is infinite but countable. The number of digraphs of order $n$ is $4^{\binom{n}{2}}$.

Random graphs have many uses in the analysis of large systems of interacting objects; for example, a random intersection graph may be used to make inferences about the clustering of internet users based on the web sites they visit.

### 8.1.4 Association Matrices

In data analysis, the interesting questions usually involve the relationships among the variables or among the observational units. Matrices formed from the original data matrix for the purpose of measuring these relationships are called association matrices. There are basically two types: similarity and dissimilarity matrices. The variance-covariance matrix, which we discuss in Section 8.6.3, is an example of an association matrix that measures similarity. We discuss dissimilarity matrices in Section 8.6.6 and in Section 8.8.7 discuss a type of similarity matrix for data represented in graphs.

In addition to the distinction between similarity and dissimilarity association matrices, we may identify two types of association matrices based on
whether the relationships of interest are among the rows (observations) or among the columns (variables or features). In applications, dissimilarity relationships among rows tend to be of more interest, and similarity relationships among columns are usually of more interest. (The applied statistician may think of clustering, multidimensional scaling, or Q factor analysis for the former and correlation analysis, principal components analysis, or factor analysis for the latter.)

8.2 Symmetric Matrices

Most association matrices encountered in applications are real and symmetric. Because real symmetric matrices occur so frequently in statistical applications and because such matrices have so many interesting properties, it is useful to review some of those properties that we have already encountered and to state some additional properties.

First, perhaps, we should iterate a trivial but important fact: the product of symmetric matrices is not, in general, symmetric. A power of a symmetric matrix, however, is symmetric.

We should also emphasize that some of the special matrices we have discussed are assumed to be symmetric because, if they were not, we could define equivalent symmetric matrices. This includes positive definite matrices and more generally the matrices in quadratic forms.

For convenience, here we list some of the important properties of symmetric matrices, many of which concern their eigenvalues. In the following, let $A$ be a symmetric matrix:

- If $A$ is nonsingular, then $A^{-1}$ is also symmetric because $(A^{-1})^T = (A^T)^{-1} = A^{-1}$.
- If $k$ is any positive integer, $A^k$ is symmetric.
- If $A$ is nonsingular (so that $A^k$ is defined for nonpositive integers), $A^k$ is symmetric for any integer $k$.
- All eigenvalues of a (real) symmetric matrix are real (see page 110).
- $A$ is diagonalizable (or simple), and in fact $A$ is orthogonally diagonalizable; that is, it has an orthogonally similar canonical factorization, $A = VCV^T$ (see page 120).
- $A$ has the spectral decomposition $A = \sum_i c_i v_i v_i^T$, where the $c_i$ are the eigenvalues and $v_i$ are the corresponding eigenvectors (see page 121).
- A power of $A$ has the spectral decomposition $A^k = \sum_i c_i^k v_i v_i^T$.
- Any quadratic form $x^T Ax$ can be expressed as $\sum_i b_i^2 c_i$, where the $b_i$ are elements in the vector $V^{-1} x$.
- We have

$$\max_{x \neq 0} \frac{x^T Ax}{x^T x} = \max \{c_i\}$$

(see page 122). If $A$ is nonnegative definite, this is the spectral radius $\rho(A)$. 
• For the $L_2$ norm of the symmetric matrix $A$, we have

$$\|A\|_2 = \rho(A).$$

• For the Frobenius norm of the symmetric matrix $A$, we have

$$\|A\|_F = \sqrt{\sum c_i^2}.$$  

This follows immediately from the fact that $A$ is diagonalizable, as do the facts that

$$\text{tr}(A) = \sum c_i$$

and

$$|A| = \prod c_i$$

(see equations (3.179) and (3.180) on page 110).

### Approximation of Symmetric Matrices and an Important Inequality

In Section 3.10, we considered the problem of approximating a given matrix by another matrix of lower rank. There are other situations in statistics in which we need to approximate one matrix by another one. In data analysis, this may be because our given matrix arises from poor observations and we know the “true” matrix has some special properties not possessed by the given matrix computed from the data. A familiar example is a sample variance-covariance matrix computed from incomplete data (see Section 9.4.6). Other examples in statistical applications occur in the simulation of random matrices (see Gentle, 2003, Section 5.3.3). In most cases of interest, the matrix to be approximated is a symmetric matrix.

Consider the difference of two symmetric $n \times n$ matrices, $A$ and $\tilde{A}$; that is,

$$E = A - \tilde{A}. \quad (8.5)$$

The matrix of the differences, $E$, is also symmetric. We measure the “closeness” of $A$ and $\tilde{A}$ by some norm of $E$.

The Hoffman-Wielandt theorem gives a lower bound on the Frobenius norm of $E$ in terms of the differences of the eigenvalues of $A$ and $\tilde{A}$: if the eigenvalues of $A$ are $c_1, \ldots, c_n$ and the eigenvalues of $\tilde{A}$ are $\tilde{c}_1, \ldots, \tilde{c}_n$, each set being arranged in nonincreasing order, we have

$$\sum_{i=1}^{n} (c_i - \tilde{c}_i)^2 \leq \|E\|_F^2. \quad (8.6)$$

This fact was proved by Hoffman and Wielandt (1953) using techniques from linear programming. Wilkinson (1965) gives a simpler proof (which he attributes to Wallace Givens) along the following lines.
Because $A$, $\tilde{A}$, and $E$ are symmetric, they are all orthogonally diagonalizable. Let the diagonal factorizations of $A$ and $E$, respectively, be $VCV^T$ and $U\text{diag}((e_1,\ldots,e_n))U^T$, where $e_1,\ldots,e_n$ are the eigenvalues of $E$ in nonincreasing order. Hence, we have

$$U\text{diag}((e_1,\ldots,e_n))U^T = U(A - \tilde{A})U^T = U(VCV^T - \tilde{A})U^T = UV(C - V^T\tilde{A}V)V^TU^T.$$ 

Taking norms of both sides, we have

$$\sum_{i=1}^n e_i^2 = \|C - V^T\tilde{A}V\|^2. \quad (8.7)$$

(All norms in the remainder of this section will be the Frobenius norm.) Now, let

$$f(Q) = \|C - Q^T\tilde{A}Q\|^2 \quad (8.8)$$

be a function of any $n \times n$ orthogonal matrix, $Q$. (Equation (8.7) yields $f(V) = \sum e_i^2$.) To arrive at inequality (8.6), we show that this function is bounded below by the sum of the differences in the squares of the elements of $C$ (which are the eigenvalues of $A$) and the eigenvalues of $Q^T\tilde{A}Q$ (which are the eigenvalues of the matrix approximating $A$).

Because the elements of $Q$ are bounded, $f(\cdot)$ is bounded, and because the set of orthogonal matrices is compact (see page 105) and $f(\cdot)$ is continuous, $f(\cdot)$ must attain its lower bound, say $l$. To simplify the notation, let

$$X = Q^T\tilde{A}Q.$$ 

Now suppose that there are $r$ distinct eigenvalues of $A$ (that is, the diagonal element in $C$):

$$d_1 > \cdots > d_r.$$ 

We can write $C$ as $\text{diag}(d_i I_{m_i})$, where $m_i$ is the multiplicity of $d_i$. We now partition $Q^T\tilde{A}Q$ to correspond to the partitioning of $C$ represented by $\text{diag}(d_i I_{m_i})$:

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1r} \\ \vdots & \ddots & \vdots \\ X_{r1} & \cdots & X_{rr} \end{bmatrix}. \quad (8.9)$$

In this partitioning, the diagonal blocks, $X_{ii}$, are $m_i \times m_i$ symmetric matrices. The submatrix $X_{ij}$, is an $m_i \times m_j$ matrix.

We now proceed in two steps to show that in order for $f(Q)$ to attain its lower bound $l$, $X$ must be diagonal. First we will show that when $f(Q) = l$, the submatrix $X_{ij}$ in equation (8.9) must be null if $i \neq j$. To this end, let $Q_{\psi}$ be such that $f(Q_{\psi}) = l$, and assume the contrary regarding the corresponding
\(X_v = Q_v^T \tilde{A} Q_v\); that is, assume that in some submatrix \(X_{ij_v}\) where \(i \neq j\), there is a nonzero element, say \(x_v\). We arrive at a contradiction by showing that in this case there is another \(X_0\) of the form \(Q_0^T \tilde{A} Q_0\), where \(Q_0\) is orthogonal and such that \(f(Q_0) < f(Q_v)\).

To establish some useful notation, let \(p\) and \(q\) be the row and column, respectively, of \(X_v\) where this nonzero element \(x_v\) occurs; that is, \(x_{pq} = x_v \neq 0\) and \(p \neq q\) because \(x_{pq}\) is in \(X_{ij_v}\). (Note the distinction between uppercase letters, which represent submatrices, and lowercase letters, which represent elements of matrices.) Also, because \(X_v\) is symmetric, \(x_{qp} = x_v\). Now let \(a_v = x_{pp}\) and \(b_v = x_{qq}\). We form \(Q_0\) as \(Q_v R\), where \(R\) is an orthogonal rotation matrix of the form \(G_{pq}\) in equation (5.12). We have, therefore, \(\|Q_0^T \tilde{A} Q_0\|_2 = \|R^T Q_v^T \tilde{A} Q_v R\|_2 = \|Q_v^T \tilde{A} Q_v\|_2\). Let \(a_0\), \(b_0\), and \(x_0\) represent the elements of \(Q_0^T \tilde{A} Q_0\) that correspond to \(a_v\), \(b_v\), and \(x_v\) in \(Q_v^T \tilde{A} Q_v\).

From the definition of the Frobenius norm, we have

\[
f(Q_0) - f(Q_v) = 2(a_v - a_0)d_i + 2(b_v - b_0)d_j
\]

because all other terms cancel. If the angle of rotation is \(\theta\), then

\[
a_0 = a_v \cos^2 \theta - 2x_v \cos \theta \sin \theta + b_v \sin^2 \theta,
\]

\[
b_0 = a_v \sin^2 \theta - 2x_v \cos \theta \sin \theta + b_v \cos^2 \theta,
\]

and so for a function \(h\) of \(\theta\) we can write

\[
h(\theta) = f(Q_0) - f(Q_v) = 2d_i((a_v - b_v) \sin^2 \theta + x_v \sin 2\theta) + 2d_j((b_v - b_0) \sin^2 \theta - x_v \sin 2\theta)
\]

\[
= 2d_i((a_v - b_v) + 2d_j(b_v - b_0)) \sin^2 \theta + 2x_v (d_i - d_j) \sin 2\theta,
\]

and so

\[
\frac{d}{d\theta} h(\theta) = 2d_i((a_v - b_v) + 2d_j(b_v - b_0)) \sin 2\theta + 4x_v (d_i - d_j) \cos 2\theta.
\]

The coefficient of \(\cos 2\theta\), \(4x_v (d_i - d_j)\), is nonzero because \(d_i\) and \(d_j\) are distinct, and \(x_v\) is nonzero by the second assumption to be contradicted, and so the derivative at \(\theta = 0\) is nonzero. Hence, by the proper choice of a direction of rotation (which effectively interchanges the roles of \(d_i\) and \(d_j\)), we can make \(f(Q_0) - f(Q_v)\) positive or negative, showing that \(f(Q_v)\) cannot be a minimum if some \(X_{ij}\) in equation (8.9) with \(i \neq j\) is nonnull; that is, if \(Q_v\) is a matrix such that \(f(Q_v)\) is the minimum of \(f(Q)\), then in the partition of \(Q_v^T \tilde{A} Q_v\) only the diagonal submatrices \(X_{ii_v}\) can be nonnull:

\[
Q_v^T \tilde{A} Q_v = \text{diag}(X_{11_v}, \ldots, X_{rr_v}).
\]

The next step is to show that each \(X_{ii_v}\) must be diagonal. Because it is symmetric, we can diagonalize it with an orthogonal matrix \(P_i\) as
Now let $P$ be the direct sum of the $P_i$ and form
\[
P^T C P - P^T Q^T \tilde{A} Q P = \text{diag}(d_1 I, \ldots, d_r I) - \text{diag}(G_1, \ldots, G_r)
\]
\[
= C - P^T Q^T \tilde{A} Q P.
\]
Hence,
\[
f(Q^P) = f(Q^V),
\]
and so the minimum occurs for a matrix $Q^P$ that reduces $\tilde{A}$ to a diagonal form. The elements of the $G_i$ must be the $\tilde{c}_i$ in some order, so the minimum of $f(Q)$, which we have denoted by $f(Q^V)$, is $\sum (c_i - \tilde{c}_{p_i})^2$, where the $p_i$ are a permutation of $1, \ldots, n$. As the final step, we show $p_i = i$. We begin with $p_1$. Suppose $p_1 \neq 1$ but $p_s = 1$; that is, $\tilde{c}_1 \geq \tilde{c}_{p_1}$. Interchange $p_1$ and $p_s$ in the permutation. The change in the sum $\sum (c_i - \tilde{c}_{p_i})^2$ is
\[
(c_1 - \tilde{c}_1)^2 + (c_s - \tilde{c}_{p_s})^2 - (c_1 - \tilde{c}_{p_s})^2 - (c_s - \tilde{c}_1)^2 = -2(c_s - c_1)(\tilde{c}_{p_1} - \tilde{c}_1)
\]
\[
\leq 0;
\]
that is, the interchange reduces the value of the sum. Similarly, we proceed through the $p_i$ to $p_n$, getting $p_i = i$.

We have shown, therefore, that the minimum of $f(Q)$ is $\sum_{i=1}^n (c_i - \tilde{c}_i)^2$, where both sets of eigenvalues are ordered in nonincreasing value. From equation (8.7), which is $f(V)$, we have the inequality (8.6).

While an upper bound may be of more interest in the approximation problem, the lower bound in the Hoffman-Wielandt theorem gives us a measure of the goodness of the approximation of one matrix by another matrix. Chu (1991) describes various extensions and applications of the Hoffman-Wielandt theorem.

**Normal Matrices**

A real square matrix $A$ is said to be normal if $A^T A = AA^T$. (In general, a square matrix is normal if $A^H A = AA^H$.) Normal matrices include symmetric (and Hermitian), skew symmetric (and Hermitian), and square orthogonal (and unitary) matrices.

There are a number of interesting properties possessed by normal matrices. One property, for example, is that eigenvalues of normal matrices are real. (This follows from properties 12 and 13 on page 110.) Another property of a normal matrix is its characterization in terms of orthogonal similarity to a diagonal matrix formed from its eigenvalues; a square matrix is normal if and only if it can be expressed in the form of equation (3.197), $A = VCV^T$, which we derived for symmetric matrices.

The normal matrices of most interest to us are symmetric matrices, and so when we discuss properties of normal matrices, we will generally consider those properties only as they apply to symmetric matrices.
8.3 Nonnegative Definite Matrices; Cholesky Factorization

We defined nonnegative definite and positive definite matrices on page 70, and discussed some of their properties, particularly in Section 3.8.8. We have seen that these matrices have useful factorizations, in particular, the square root and the Cholesky factorization. In this section, we recall those definitions, properties, and factorizations.

A symmetric matrix $A$ such that any quadratic form involving the matrix is nonnegative is called a nonnegative definite matrix. That is, a symmetric matrix $A$ is a nonnegative definite matrix if, for any (conformable) vector $x$,

$$x^T Ax \geq 0. \quad (8.10)$$

(There is a related term, positive semidefinite matrix, that is not used consistently in the literature. We will generally avoid the term “semidefinite”.)

We denote the fact that $A$ is nonnegative definite by

$$A \succeq 0. \quad (8.11)$$

(Some people use the notation $A \geq 0$ to denote a nonnegative definite matrix, but we have decided to use this notation to indicate that each element of $A$ is nonnegative; see page 48.)

There are several properties that follow immediately from the definition.

- The sum of two (conformable) nonnegative matrices is nonnegative definite.
- All diagonal elements of a nonnegative definite matrix are nonnegative. Hence, if $A$ is nonnegative definite, $\text{tr}(A) \geq 0$.
- Any square submatrix whose principal diagonal is a subset of the principal diagonal of a nonnegative definite matrix is nonnegative definite. In particular, any square principal submatrix of a nonnegative definite matrix is nonnegative definite.

It is easy to show that the latter two facts follow from the definition by considering a vector $x$ with zeros in all positions except those corresponding to the submatrix in question. For example, to see that all diagonal elements of a nonnegative definite matrix are nonnegative, assume the $(i, i)$ element is negative, and then consider the vector $x$ to consist of all zeros except for a 1 in the $i^{th}$ position. It is easy to see that the quadratic form is negative, so the assumption that the $(i, i)$ element is negative leads to a contradiction.

- A diagonal matrix is nonnegative definite if and only if all of the diagonal elements are nonnegative.

This must be true because a quadratic form in a diagonal matrix is the sum of the diagonal elements times the squares of the elements of the vector.
If $A$ is nonnegative definite, then $A_{-(i_1,...,i_k)(i_1,...,i_k)}$ is nonnegative definite.

Again, we can see this by selecting an $x$ in the defining inequality (8.10) consisting of 1s in the positions corresponding to the rows and columns of $A$ that are retained and 0s elsewhere.

By considering $x^T C^T AC x$ and $y = Cx$, we see that

- if $A$ is nonnegative definite, and $C$ is conformable for the multiplication, then $C^T AC$ is nonnegative definite.

From equation (3.197) and the fact that the determinant of a product is the product of the determinants, we have that

- the determinant of a nonnegative definite matrix is nonnegative.

Finally, for the nonnegative definite matrix $A$, we have

$$a_{ij}^2 \leq a_{ii} a_{jj},$$

as we see from the definition $x^T Ax \geq 0$ and choosing the vector $x$ to have a variable $y$ in position $i$, a 1 in position $j$, and 0s in all other positions. For a symmetric matrix $A$, this yields the quadratic $a_{ii}y^2 + 2a_{ij}y + a_{jj}$. If this quadratic is to be nonnegative for all $y$, then the discriminant $4a_{ij}^2 - 4a_{ii}a_{jj}$ must be nonpositive; that is, inequality (8.12) must be true.

### Eigenvalues of Nonnegative Definite Matrices

We have seen on page 124 that a real symmetric matrix is nonnegative (positive) definite if and only if all of its eigenvalues are nonnegative (positive).

This fact allows a generalization of the statement above: a triangular matrix is nonnegative (positive) definite if and only if all of the diagonal elements are nonnegative (positive).

### The Square Root and the Cholesky Factorization

Two important factorizations of nonnegative definite matrices are the square root,

$$A = (A^{1/2})^2,$$

discussed in Section 5.9.1, and the Cholesky factorization,

$$A = T^T T,$$

discussed in Section 5.9.2. If $T$ is as in equation (8.14), the symmetric matrix $T + T^T$ is also nonnegative definite, or positive definite if $A$ is. The square root matrix is used often in theoretical developments, such as Exercise 4.5b for example, but the Cholesky factor is more useful in practice.
8.4 Positive Definite Matrices

An important class of nonnegative definite matrices are those that satisfy strict inequalities in the definition involving $x^T Ax$. These matrices are called positive definite matrices and they have all of the properties discussed above for nonnegative definite matrices as well as some additional useful properties.

A symmetric matrix $A$ is called a positive definite matrix if, for any (conformable) vector $x \neq 0$, the quadratic form is positive; that is,

$$x^T Ax > 0.$$  \hfill (8.15)

We denote the fact that $A$ is positive definite by

$$A \succ 0.$$  \hfill (8.16)

(Some people use the notation $A > 0$ to denote a positive definite matrix, but we have decided to use this notation to indicate that each element of $A$ is positive.)

- A positive definite matrix is necessarily nonsingular. (We see this from the fact that no nonzero combination of the columns, or rows, can be 0.) Furthermore, if $A$ is positive definite, then $A^{-1}$ is positive definite. (We showed this is Section 3.8.8, but we can see it in another way: because for any $y \neq 0$ and $x = A^{-1}y$, we have $y^T A^{-1}y = x^T y = x^T Ax > 0$.)

- A diagonally dominant symmetric matrix with positive diagonals is positive definite. The proof of this is Exercise 8.2.

The properties of nonnegative definite matrices noted above hold also for positive definite matrices, generally with strict inequalities. It is obvious that all diagonal elements of a positive definite matrix are positive. Hence, if $A$ is positive definite, $\text{tr}(A) > 0$. Furthermore, as above and for the same reasons, if $A$ is positive definite, then $A_{-\{i_1, \ldots, i_k\}(i_1, \ldots, i_k)}$ is positive definite. In particular, any square submatrix whose principal diagonal is a subset of the principal diagonal of a positive definite matrix is positive definite, and furthermore, any square principal submatrix of a positive definite matrix is positive definite. Because a quadratic form in a diagonal matrix is the sum of the diagonal elements times the squares of the elements of the vector, a diagonal matrix is positive definite if and only if all of the diagonal elements are positive.

The definition yields a slightly stronger statement regarding the sums involving positive definite matrices than what we could conclude about nonnegative definite matrices:

- The sum of a positive definite matrix and a (conformable) nonnegative definite matrix is positive definite.

That is,

$$x^T Ax > 0 \forall x \neq 0 \quad \text{and} \quad y^T By \geq 0 \forall y \implies z^T (A + B) z > 0 \forall z \neq 0.$$  \hfill (8.17)
We cannot conclude that the product of two positive definite matrices is positive definite, but we do have the useful fact that

- if $A$ is positive definite, and $C$ is of full rank and conformable for the multiplication $AC$, then $C^TAC$ is positive definite (see page 89).

From equation (3.197) and the fact that the determinant of a product is the product of the determinants, we have that

- the determinant of a positive definite matrix is positive.

For the positive definite matrix $A$, we have, analogous to inequality (8.12),

$$a_{ij}^2 < a_{ii}a_{jj}, \quad (8.18)$$

which we see using the same argument as for that inequality.

We have seen from the definition of positive definiteness and the distribution of multiplication over addition that the sum of a positive definite matrix and a nonnegative definite matrix is positive definite. We can define an ordinal relationship between positive definite and nonnegative definite matrices of the same size. If $A$ is positive definite and $B$ is nonnegative definite of the same size, we say $A$ is **strictly greater than** $B$ and write

$$A \succ B \quad (8.19)$$

if $A - B$ is positive definite; that is, if $A - B \succ 0$.

We can form a partial ordering of nonnegative definite matrices of the same order based on this additive property. We say $A$ is **greater than** $B$ and write

$$A \succeq B \quad (8.20)$$

if $A - B$ is either the 0 matrix or is nonnegative definite; that is, if $A - B \succeq 0$ (see Exercise 8.1a). The “strictly greater than” relation implies the “greater than” relation. These relations are partial in the sense that they do not apply to all pairs of nonnegative matrices; that is, there are pairs of matrices $A$ and $B$ for which neither $A \succeq B$ nor $B \succeq A$.

If $A \succ B$, we also write $B \prec A$; and if $A \succeq B$, we may write $B \preceq A$.

**Principal Submatrices of Positive Definite Matrices**

A sufficient condition for a symmetric matrix to be positive definite is that the determinant of each of the leading principal submatrices be positive. To see this, first let the $n \times n$ symmetric matrix $A$ be partitioned as

$$A = \begin{bmatrix} A_{n-1} & a \\ a^T & a_{nn} \end{bmatrix},$$

and assume that $A_{n-1}$ is positive definite and that $|A| > 0$. (This is not the same notation that we have used for these submatrices, but the notation is convenient in this context.) From equation (3.147),
Because $A_{n-1}$ is positive definite, $|A_{n-1}| > 0$, and so $(a_{nn} - a^T A_{n-1}^{-1} a) > 0$; hence, the $1 \times 1$ matrix $(a_{nn} - a^T A_{n-1}^{-1} a)$ is positive definite. That any matrix whose leading principal submatrices have positive determinants follows from this by induction, beginning with a $2 \times 2$ matrix.

The Convex Cone of Positive Definite Matrices

The class of all $n \times n$ positive definite matrices is a convex cone in $\mathbb{R}^{n \times n}$ in the same sense as the definition of a convex cone of vectors (see page 14). If $X_1$ and $X_2$ are $n \times n$ positive definite matrices and $a, b \geq 0$, then $aX_1 + bX_2$ is positive definite so long as either $a \neq 0$ or $b \neq 0$.

This class is not closed under Cayley multiplication (that is, in particular, it is not a group with respect to that operation). The product of two positive definite matrices might not even be symmetric.

Inequalities Involving Positive Definite Matrices

Quadratic forms of positive definite matrices and nonnegative matrices occur often in data analysis. There are several useful inequalities involving such quadratic forms.

On page 122, we showed that if $x \neq 0$, for any symmetric matrix $A$ with eigenvalues $c_i$,

$$
\frac{x^T A x}{x^T x} \leq \max \{ c_i \}. 
$$

(8.21)

If $A$ is nonnegative definite, by our convention of labeling the eigenvalues, we have $\max \{ c_i \} = c_1$. If the rank of $A$ is $r$, the minimum nonzero eigenvalue is denoted $c_r$. Letting the eigenvectors associated with $c_1, \ldots, c_r$ be $v_1, \ldots, v_r$ (and recalling that these choices may be arbitrary in the case where some eigenvalues are not simple), by an argument similar to that used on page 122, we have that if $A$ is nonnegative definite of rank $r$,

$$
\frac{v_i^T A v_i}{v_i^T v_i} \geq c_r, 
$$

(8.22)

for $1 \leq i \leq r$.

If $A$ is positive definite and $x$ and $y$ are conformable nonzero vectors, we see that

$$
x^T A^{-1} x \geq \frac{(y^T x)^2}{y^T A y} 
$$

(8.23)

by using the same argument as used in establishing the Cauchy-Schwarz inequality (2.10). We first obtain the Cholesky factor $T$ of $A$ (which is, of course, of full rank) and then observe that for every real number $t$
Matrices with Special Properties

\[(tT_y + T^{-T}x)^T (tT_y + T^{-T}x) \geq 0,\]

and hence the discriminant of the quadratic equation in \(t\) must be nonnegative:

\[4 ((T y)^T T^{-T} x)^2 - 4 (T^{-T} x)^T (T^{-T} - x) (T y)^T T y \leq 0.\]

The inequality (8.23) is used in constructing Scheffé simultaneous confidence intervals in linear models.

The Kantorovich inequality for positive numbers has an immediate extension to an inequality that involves positive definite matrices. The Kantorovich inequality, which finds many uses in optimization problems, states, for positive numbers \(c_1 \geq c_2 \geq \cdots \geq c_n\) and nonnegative numbers \(y_1, \ldots, y_n\) such that \(\sum y_i = 1\), that

\[\left(\sum_{i=1}^{n} y_i c_i\right) \left(\sum_{i=1}^{n} y_i c_i^{-1}\right) \leq \frac{(c_1 + c_2)^2}{4 c_1 c_2}.\]

Now let \(A\) be an \(n \times n\) positive definite matrix with eigenvalues \(c_1 \geq c_2 \geq \cdots \geq c_n > 0\). We substitute \(x^2\) for \(y\), thus removing the nonnegativity restriction, and incorporate the restriction on the sum directly into the inequality. Then, using the similar canonical factorization of \(A\) and \(A^{-1}\), we have

\[
\frac{(x^T A x) (x^T A^{-1} x)}{(x^T x)^2} \leq \frac{(c_1 + c_n)^2}{4 c_1 c_n}.
\]

This Kantorovich matrix inequality likewise has applications in optimization; in particular, for assessing convergence of iterative algorithms.

The left-hand side of the Kantorovich matrix inequality also has a lower bound,

\[
\frac{(x^T A x) (x^T A^{-1} x)}{(x^T x)^2} \geq 1,
\]

which can be seen in a variety of ways, perhaps most easily by using the inequality (8.23). (You were asked to prove this directly in Exercise 3.21.)

All of the inequalities (8.21) through (8.25) are sharp. We know that (8.21) and (8.22) are sharp by using the appropriate eigenvectors. We can see the others are sharp by using \(A = I\).

There are several variations on these inequalities and other similar inequalities that are reviewed by Marshall and Olkin (1990) and Liu and Neudecker (1996).

### 8.5 Idempotent and Projection Matrices

An important class of matrices are those that, like the identity, have the property that raising them to a power leaves them unchanged. A matrix \(A\) such that
Idempotent and Projection Matrices

8.5

\[ AA = A \]  

(8.26)

is called an \textit{idempotent matrix}. An idempotent matrix is square, and it is either singular or the identity matrix. (It must be square in order to be conformable for the indicated multiplication. If it is not singular, we have \( A = (A^{-1}A)A = A^{-1}A = I \); hence, an idempotent matrix is either singular or the identity matrix.) An idempotent matrix that is symmetric is called a \textit{projection matrix}.

8.5.1 Idempotent Matrices

Many matrices encountered in the statistical analysis of linear models are idempotent. One such matrix is \( X^{-X} \) (see page 98 and Section 9.2.2). This matrix exists for any \( n \times m \) matrix \( X \), and it is square. (It is \( m \times m \).)

Because the eigenvalues of \( A^2 \) are the squares of the eigenvalues of \( A \), all eigenvalues of an idempotent matrix must be either 0 or 1.

Any vector in the column space of an idempotent matrix \( A \) is an eigenvector of \( A \). (This follows immediately from \( AA = A \).) More generally, if \( x \) and \( y \) are vectors in \( \text{span}(A) \) and \( a \) is a scalar, then

\[ A(ax + y) = ax + y. \]  

(8.27)

(To see this, we merely represent \( x \) and \( y \) as linear combinations of columns (or rows) of \( A \) and substitute in the equation.)

The number of eigenvalues that are 1 is the rank of an idempotent matrix. (Exercise 8.3 asks why this is the case.) We therefore have, for an idempotent matrix \( A \),

\[ \text{tr}(A) = \text{rank}(A). \]  

(8.28)

Because the eigenvalues of an idempotent matrix are either 0 or 1, a symmetric idempotent matrix is nonnegative definite.

If \( A \) is idempotent and \( n \times n \), then

\[ \text{rank}(I - A) = n - \text{rank}(A). \]  

(8.29)

We showed this in equation (3.155) on page 98. (Although there we were considering the special matrix \( A^{-A} \), the only properties used were the idempotency of \( A^{-A} \) and the fact that \( \text{rank}(A^{-A}) = \text{rank}(A) \).)

Equation (8.29) together with the diagonalizability theorem (equation (3.194)) implies that an idempotent matrix is diagonalizable.

If \( A \) is idempotent and \( V \) is an orthogonal matrix of the same size, then \( V^T AV \) is idempotent (whether or not \( V \) is a matrix that diagonalizes \( A \)) because

\[ (V^T AV)(V^T AV) = V^T AAV = V^T AV. \]  

(8.30)

If \( A \) is idempotent, then \( (I - A) \) is also idempotent, as we see by multiplication. This fact and equation (8.29) have generalizations for sums of
idempotent matrices that are parts of Cochran’s theorem, which we consider below.

Although if \( A \) is idempotent so \( (I - A) \) is also idempotent and hence is not of full rank (unless \( A = 0 \)), for any scalar \( a \neq -1 \), \((I + aA)\) is of full rank, and

\[
(I + aA)^{-1} = I - \frac{a}{a + 1}A,
\]

as we see by multiplication.

On page 114, we saw that similar matrices are equivalent (have the same rank). For idempotent matrices, we have the converse: idempotent matrices of the same rank (and size) are similar (see Exercise 8.4).

If \( A_1 \) and \( A_2 \) are matrices conformable for addition, then \( A_1 + A_2 \) is idempotent if and only if \( A_1A_2 = A_2A_1 = 0 \). It is easy to see that this condition is sufficient by multiplication:

\[
(A_1 + A_2)(A_1 + A_2) = A_1A_1 + A_1A_2 + A_2A_1 + A_2A_2 = A_1 + A_2.
\]

To see that it is necessary, we first observe from the expansion above that \( A_1 + A_2 \) is idempotent only if \( A_1A_2 + A_2A_1 = 0 \). Multiplying this necessary condition on the left by \( A_1 \) yields

\[
A_1A_1A_2 + A_1A_2A_1 = A_1A_2 + A_1A_2A_1 = 0,
\]

and multiplying on the right by \( A_1 \) yields

\[
A_1A_2A_1 + A_2A_1A_1 = A_1A_2A_1 + A_2A_1 = 0.
\]

Subtracting these two equations yields

\[
A_1A_2 = A_2A_1,
\]

and since \( A_1A_2 + A_2A_1 = 0 \), we must have \( A_1A_2 = A_2A_1 = 0 \).

**Symmetric Idempotent Matrices**

Many of the idempotent matrices in statistical applications are symmetric, and such matrices have some useful properties.

Because the eigenvalues of an idempotent matrix are either 0 or 1, the spectral decomposition of a symmetric idempotent matrix \( A \) can be written as

\[
V^T AV = \text{diag}(I_r, 0),
\]

where \( V \) is a square orthogonal matrix and \( r = \text{rank}(A) \). (This is from equation (3.198) on page 120.)

For symmetric matrices, there is a converse to the fact that all eigenvalues of an idempotent matrix are either 0 or 1. If \( A \) is a symmetric matrix all of whose eigenvalues are either 0 or 1, then \( A \) is idempotent. We see this from the
spectral decomposition of $A$, $A = V \text{diag}(I_r, 0)V^T$, and, with $C = \text{diag}(I_r, 0)$, by observing

$$AA = VCV^T VCV^T = VCCV^T = VCV^T = A,$$

because the diagonal matrix of eigenvalues $C$ contains only 0s and 1s.

If $A$ is symmetric and $p$ is any positive integer,

$$A^{p+1} = A^p \implies A \text{ is idempotent.} \quad (8.33)$$

This follows by considering the eigenvalues of $A$, $c_1, \ldots, c_n$. The eigenvalues of $A^{p+1}$ are $c_1^{p+1}, \ldots, c_n^{p+1}$ and the eigenvalues of $A^p$ are $c_1^p, \ldots, c_n^p$, but since $A^{p+1} = A^p$, it must be the case that $c_i^{p+1} = c_i^p$ for each $i = 1, \ldots, n$. The only way this is possible is for each eigenvalue to be 0 or 1, and in this case the symmetric matrix must be idempotent.

There are bounds on the elements of a symmetric idempotent matrix. Because $A$ is symmetric and $A^T A = A$,

$$a_{ii} = \sum_{j=1}^{n} a_{ij}^2; \quad (8.34)$$

hence, $0 \leq a_{ii}$. Rearranging equation (8.34), we have

$$a_{ii} = a_{ii}^2 + \sum_{j \neq i} a_{ij}^2, \quad (8.35)$$

so $a_{ii}^2 \leq a_{ii}$ or $0 \leq a_{ii}(1 - a_{ii})$; that is, $a_{ii} \leq 1$. Now, if $a_{ii} = 0$ or $a_{ii} = 1$, then equation (8.35) implies

$$\sum_{j \neq i} a_{ij}^2 = 0,$$

and the only way this can happen is if $a_{ij} = 0$ for all $j \neq i$. So, in summary, if $A$ is an $n \times n$ symmetric idempotent matrix, then

$$0 \leq a_{ii} \leq 1 \text{ for } i = 1, \ldots, m, \quad (8.36)$$

and

$$\text{if } a_{ii} = 0 \text{ or } a_{ii} = 1, \text{ then } a_{ij} = a_{ji} = 0 \text{ for all } j \neq i. \quad (8.37)$$

**Cochran’s Theorem**

There are various facts that are sometimes called **Cochran’s theorem**. The simplest one concerns $k$ symmetric idempotent $n \times n$ matrices, $A_1, \ldots, A_k$, such that

$$I_n = A_1 + \cdots + A_k. \quad (8.38)$$

Under these conditions, we have
We see this by the following argument. For an arbitrary \( j \), as in equation (8.32), for some matrix \( V \), we have
\[
V^T A_j V = \text{diag}(I_r, 0),
\]
where \( r = \text{rank}(A_j) \). Now
\[
I_n = V^T I_n V = \sum_{i=1}^{k} V^T A_i V = \text{diag}(I_r, 0) + \sum_{i \neq j} V^T A_i V,
\]
which implies
\[
\sum_{i \neq j} V^T A_i V = \text{diag}(0, I_{n-r}). \tag{8.40}
\]
Now, from equation (8.30), for each \( i \), \( V^T A_i V \) is idempotent, and so from equation (8.36) the diagonal elements are all nonnegative, and hence equation (8.40) implies that for each \( i \neq j \), the first \( r \) diagonal elements are 0. Furthermore, since these diagonal elements are 0, equation (8.37) implies that all elements in the first \( r \) rows and columns are 0. We have, therefore, for each \( i \neq j \),
\[
V^T A_i V = \text{diag}(0, B_i)
\]
for some \((n-r) \times (n-r)\) symmetric idempotent matrix \( B_i \). Now, for any \( i \neq j \), consider \( A_i A_j \) and form \( V^T A_i A_j V \). We have
\[
V^T A_i A_j V = (V^T A_i V)(V^T A_j V) = \text{diag}(0, B_i) \text{diag}(I_r, 0) = 0.
\]
Because \( V \) is nonsingular, this implies the desired conclusion; that is, that \( A_i A_j = 0 \) for any \( i \neq j \).

We can now extend this result to an idempotent matrix in place of \( I \); that is, for an idempotent matrix \( A \) with \( A = A_1 + \cdots + A_k \). Rather than stating it simply as in equation (8.39), however, we will state the implications differently.

Let \( A_1, \ldots, A_k \) be \( n \times n \) symmetric matrices and let
\[
A = A_1 + \cdots + A_k. \tag{8.41}
\]
Then any two of the following conditions imply the third one:
(a). $A$ is idempotent.
(b). $A_i$ is idempotent for $i = 1, \ldots, k$.
(c). $A_iA_j = 0$ for all $i \neq j$.

This is also called Cochran’s theorem. (The theorem also applies to non-symmetric matrices if condition (c) is augmented with the requirement that \(\text{rank}(A_i^2) = \text{rank}(A_i)\) for all $i$. We will restrict our attention to symmetric matrices, however, because in most applications of these results, the matrices are symmetric.)

First, if we assume properties (a) and (b), we can show that property (c) follows using an argument similar to that used to establish equation (8.39) for the special case $A = I$. The formal steps are left as an exercise.

Now, let us assume properties (b) and (c) and show that property (a) holds. With properties (b) and (c), we have

$$AA = (A_1 + \cdots + A_k)(A_1 + \cdots + A_k)$$
$$= \sum_{i=1}^{k} A_i A_i + \sum_{i\neq j}^{k} A_i A_j$$
$$= \sum_{i=1}^{k} A_i$$
$$= A.$$

Hence, we have property (a); that is, $A$ is idempotent.

Finally, let us assume properties (a) and (c). Property (b) follows immediately from

$$A_i^2 = A_i A_i = A_i A = A_i AA = A_i^2 A = A_i^3$$

and the implication (8.33).

Any two of the properties (a) through (c) also imply a fourth property for $A = A_1 + \cdots + A_k$ when the $A_i$ are symmetric:

(d). $\text{rank}(A) = \text{rank}(A_1) + \cdots + \text{rank}(A_k)$.

We first note that any two of properties (a) through (c) imply the third one, so we will just use properties (a) and (b). Property (a) gives

$$\text{rank}(A) = \text{tr}(A) = \text{tr}(A_1 + \cdots + A_k) = \text{tr}(A_1) + \cdots + \text{tr}(A_k),$$

and property (b) states that the latter expression is $\text{rank}(A_1) + \cdots + \text{rank}(A_k)$, thus yielding property (d).

There is also a partial converse: properties (a) and (d) imply the other properties.

One of the most important special cases of Cochran’s theorem is when $A = I$ in the sum (8.41):

$$I_n = A_1 + \cdots + A_k.$$
The identity matrix is idempotent, so if \( \text{rank}(A_1) + \cdots + \text{rank}(A_k) = n \), all the properties above hold.

The most important statistical application of Cochran’s theorem is for the distribution of quadratic forms of normally distributed random vectors. These distribution results are also called Cochran’s theorem. We briefly discuss it in Section 9.1.3.

**Drazin Inverses**

A Drazin inverse of an operator \( T \) is an operator \( S \) such that \( TS = ST \), \( STS = S \), and \( T^{k+1}S = T^k \) for any positive integer \( k \).

It is clear that, as an operator, an idempotent matrix is its own Drazin inverse. Interestingly, if \( A \) is any square matrix, its Drazin inverse is the matrix \( A^k(A^{2k+1})+A^k \), which is unique for any positive integer \( k \). See Campbell and Meyer (1991) for discussions of properties and applications of Drazin inverses and more on their relationship to the Moore-Penrose inverse.

**8.5.2 Projection Matrices: Symmetric Idempotent Matrices**

For a given vector space \( V \), a symmetric idempotent matrix \( A \) whose columns span \( V \) is said to be a projection matrix onto \( V \); in other words, a matrix \( A \) is a projection matrix onto \( \text{span}(A) \) if and only if \( A \) is symmetric and idempotent. (Some authors do not require a projection matrix to be symmetric. In that case, the terms “idempotent” and “projection” are synonymous.)

It is easy to see that, for any vector \( x \), if \( A \) is a projection matrix onto \( V \), the vector \( Ax \) is in \( V \), and the vector \( x - Ax \) is in \( V^\perp \) (the vectors \( Ax \) and \( x - Ax \) are orthogonal). For this reason, a projection matrix is sometimes called an “orthogonal projection matrix”. Note that an orthogonal projection matrix is not an orthogonal matrix, however, unless it is the identity matrix. Stating this in alternative notation, if \( A \) is a projection matrix and \( A \in \mathbb{R}^{n \times n} \), then \( A \) maps \( \mathbb{R}^n \) onto \( V(A) \) and \( I - A \) is also a projection matrix (called the complementary projection matrix of \( A \)), and it maps \( \mathbb{R}^n \) onto the orthogonal complement, \( N(A) \). These spaces are such that \( V(A) \oplus N(A) = \mathbb{R}^n \).

In this text, we use the term “projection” to mean “orthogonal projection”, but we should note that in some literature “projection” can include “oblique projection”. In the less restrictive definition, for vector spaces \( \mathcal{V} \), \( \mathcal{X} \), and \( \mathcal{Y} \), if \( \mathcal{V} = \mathcal{X} \oplus \mathcal{Y} \) and \( v = x + y \) with \( x \in \mathcal{X} \) and \( y \in \mathcal{Y} \), then the vector \( x \) is called the projection of \( v \) onto \( \mathcal{X} \) along \( \mathcal{Y} \). In this text, to use the unqualified term “projection”, we require that \( \mathcal{X} \) and \( \mathcal{Y} \) be orthogonal; if they are not, then we call \( x \) the oblique projection of \( v \) onto \( \mathcal{X} \) along \( \mathcal{Y} \). The choice of the more restrictive definition is because of the overwhelming importance of orthogonal projections in statistical applications. The restriction is also consistent with the definition in equation (2.29) of the projection of a vector onto another vector (as opposed to the projection onto a vector space).
Because a projection matrix is idempotent, the matrix projects any of its columns onto itself, and of course it projects the full matrix onto itself: $AA = A$ (see equation (8.27)).

If $x$ is a general vector in $\mathbb{R}^n$, that is, if $x$ has order $n$ and belongs to an $n$-dimensional space, and $A$ is a projection matrix of rank $r \leq n$, then $Ax$ has order $n$ and belongs to $\text{span}(A)$, which is an $r$-dimensional space. Thus, we say projections are dimension reductions.

Useful projection matrices often encountered in statistical linear models are $X^+X$ and $XX^+$. (Recall that $X^-X$ is an idempotent matrix.) The matrix $X^+$ exists for any $n \times m$ matrix $X$, and $X^+X$ is square ($m \times m$) and symmetric.

**Projections onto Linear Combinations of Vectors**

On page 25, we gave the projection of a vector $y$ onto a vector $x$ as

$$
\frac{x^Ty}{x^Tx}x.
$$

The projection matrix to accomplish this is the “outer/inner products matrix”,

$$\frac{1}{x^Tx}xx^T. \tag{8.42}
$$

The outer/inner products matrix has rank 1. It is useful in a variety of matrix transformations. If $x$ is normalized, the projection matrix for projecting a vector on $x$ is just $xx^T$. The projection matrix for projecting a vector onto a unit vector $e_i$ is $e_ie_i^T$, and $e_ie_i^Ty = (0, \ldots, y_i, \ldots, 0)$.

This idea can be used to project $y$ onto the plane formed by two vectors, $x_1$ and $x_2$, by forming a projection matrix in a similar manner and replacing $x$ in equation (8.42) with the matrix $X = [x_1|x_2]$. On page 331, we will view linear regression fitting as a projection onto the space spanned by the independent variables.

The angle between vectors we defined on page 26 can be generalized to the angle between a vector and a plane or any linear subspace by defining it as the angle between the vector and the projection of the vector onto the subspace. By applying the definition (2.32) to the projection, we see that the angle $\theta$ between the vector $y$ and the subspace spanned by the columns of a projection matrix $A$ is determined by the cosine

$$
\cos(\theta) = \frac{y^TAy}{y^Ty}. \tag{8.43}
$$

**8.6 Special Matrices Occurring in Data Analysis**

Some of the most useful applications of matrices are in the representation of observational data, as in Figure 8.1 on page 262. If the data are represented as
real numbers, the array is a matrix, say \( X \). The rows of the \( n \times m \) data matrix \( X \) are “observations” and correspond to a vector of measurements on a single \emph{observational unit}, and the columns of \( X \) correspond to \( n \) measurements of a single variable or feature. In data analysis we may form various association matrices that measure relationships among the variables or the observations that correspond to the columns or the rows of \( X \). Many summary statistics arise from a matrix of the form \( X^T X \). (If the data in \( X \) are incomplete—that is, if some elements are missing—problems may arise in the analysis. We discuss some of these issues in Section 9.4.6.)

8.6.1 Gramian Matrices

A (real) matrix \( A \) such that for some (real) matrix \( B \), \( A = B^T B \), is called a \emph{Gramian matrix}. Any nonnegative definite matrix is Gramian (from equation (8.14) and Section 5.9.2 on page 194).

Sums of Squares and Cross Products

Although the properties of Gramian matrices are of interest, our starting point is usually the data matrix \( X \), which we may analyze by forming a Gramian matrix \( X^T X \) or \( XX^T \) (or a related matrix). These Gramian matrices are also called \emph{sums of squares and cross products matrices}. (The term “cross product” does not refer to the cross product of vectors defined on page 33, but rather to the presence of sums over \( i \) of the products \( x_{ij} x_{ik} \) along with sums of squares \( x_{ij}^2 \).) These matrices and other similar ones are useful association matrices in statistical applications.

Some Immediate Properties of Gramian Matrices

Some interesting properties of a Gramian matrix \( X^T X \) are:

\begin{itemize}
  \item \( X^T X \) is symmetric.
  \item \( X^T X \) is of full rank if and only if \( X \) is of full column rank or, more generally,
    \[
    \text{rank}(X^T X) = \text{rank}(X). \tag{8.44}
    \]
  \item \( X^T X \) is nonnegative definite and is positive definite if and only if \( X \) is of full column rank.
  \item \( X^T X = 0 \implies X = 0 \).
\end{itemize}

These properties (except the first one, which is Exercise 8.7) were shown in the discussion in Section 3.3.7 on page 90.

Each element of a Gramian matrix is the dot products of the columns of the constituent matrix. If \( x_{*i} \) and \( x_{*j} \) are the \( i^\text{th} \) and \( j^\text{th} \) columns of the matrix \( X \), then

\[
(X^T X)_{ij} = x_{*i}^T x_{*j}. \tag{8.45}
\]
A Gramian matrix is also the sum of the outer products of the rows of the constituent matrix. If \( x_{is} \) is the \( i \)th row of the \( n \times m \) matrix \( X \), then

\[
X^T X = \sum_{i=1}^{n} x_{is} x_{is}^T.
\]  

(8.46)

This is generally the way a Gramian matrix is computed.

By equation (8.14), we see that any Gramian matrix formed from a general matrix \( X \) is the same as a Gramian matrix formed from a square upper triangular matrix \( T \):

\[
X^T X = T^T T.
\]

Another interesting property of a Gramian matrix is that, for any matrices \( B \) and \( C \) (that are conformable for the operations indicated),

\[
BX^T X = CX^T X \iff BX^T = CX^T.
\]

(8.47)

The implication from right to left is obvious, and we can see the left to right implication by writing

\[
(BX^T X - CX^T X)(B^T - C^T) = (BX^T - CX^T)(BX^T - CX^T)^T,
\]

and then observing that if the left-hand side is null, then so is the right-hand side, and if the right-hand side is null, then \( BX^T - CX^T = 0 \) because \( X^T X = 0 \implies X = 0 \), as above. Similarly, we have

\[
X^T XB = X^T XC \iff X^T B = X^T C.
\]

(8.48)

**Generalized Inverses of Gramian Matrices**

The generalized inverses of \( X^T X \) have useful properties. First, we see from the definition, for any generalized inverse \((X^T X)^{-}\), that \((X^T X)^{-}\) is also a generalized inverse of \( X^T X \). (Note that \((X^T X)^{-}\) is not necessarily symmetric.) Also, we have, from equation (8.47),

\[
X(X^T X)^{-} X^T X = X.
\]

(8.49)

This means that \((X^T X)^{-} X^T\) is a generalized inverse of \( X \).

The Moore-Penrose inverse of \( X \) has an interesting relationship with a generalized inverse of \( X^T X \):

\[
XX^+ = X(X^T X)^{-} X^T.
\]

(8.50)

This can be established directly from the definition of the Moore-Penrose inverse.

An important property of \( X(X^T X)^{-} X^T \) is its invariance to the choice of the generalized inverse of \( X^T X \). Suppose \( G \) is any generalized inverse of \( X^T X \). Then, from equation (8.49), we have \( X(X^T X)^{-} X^T X = XGX^T X \), and from the implication (8.47), we have

\[
XGX^T = X(X^T X)^{-} X^T;
\]

(8.51)

that is, \( X(X^T X)^{-} X^T \) is invariant to the choice of generalized inverse.
Eigenvalues of Gramian Matrices

If the singular value decomposition of $X$ is $UDV^T$ (page 127), then the similar canonical factorization of $X^TX$ (equation (3.197)) is $VD^TDV^T$. Hence, we see that the nonzero singular values of $X$ are the square roots of the nonzero eigenvalues of the symmetric matrix $X^TX$. By using $DD^T$ similarly, we see that they are also the square roots of the nonzero eigenvalues of $XX^T$.

8.6.2 Projection and Smoothing Matrices

It is often of interest to approximate an arbitrary $n$-vector in a given $m$-dimensional vector space, where $m < n$. An $n \times n$ projection matrix of rank $m$ clearly does this.

A Projection Matrix Formed from a Gramian Matrix

An important matrix that arises in analysis of a linear model of the form

$$y = X\beta + \epsilon$$  \hspace{1cm} (8.52)

is

$$H = X(X^TX)^{-}X^T,$$  \hspace{1cm} (8.53)

where $(X^TX)^{-}$ is any generalized inverse. From equation (8.51), $H$ is invariant to the choice of generalized inverse. By equation (8.50), this matrix can be obtained from the pseudoinverse and so

$$H = XX^+.$$  \hspace{1cm} (8.54)

In the full rank case, this is uniquely

$$H = X(X^TX)^{-1}X^T.$$  \hspace{1cm} (8.55)

Whether or not $X$ is of full rank, $H$ is a projection matrix onto span($X$). It is called the “hat matrix” because it projects the observed response vector, often denoted by $y$, onto a predicted response vector, often denoted by $\hat{y}$ in span($X$):

$$\hat{y} = Hy.$$  \hspace{1cm} (8.56)

Because $H$ is invariant, this projection is invariant to the choice of generalized inverse. (In the nonfull rank case, however, we generally refrain from referring to the vector $Hy$ as the “predicted response”; rather, we may call it the “fitted response”.)

The rank of $H$ is the same as the rank of $X$, and its trace is the same as its rank (because it is idempotent). When $X$ is of full column rank, we have

$$\text{tr}(H) = \text{number of columns of } X.$$  \hspace{1cm} (8.57)
(This can also be seen by using the invariance of the trace to permutations of the factors in a product as in equation (3.55).)

In linear models, \( \text{tr}(H) \) is the model degrees of freedom, and the sum of squares due to the model is just \( y^T H y \).

The complementary projection matrix,
\[
I - H,
\]
also has interesting properties that relate to linear regression analysis. In geometrical terms, this matrix projects a vector onto \( \mathcal{N}(X^T) \), the orthogonal complement of \( \text{span}(X) \). We have
\[
y = H y + (I - H) y = \hat{y} + r, \quad (8.59)
\]
where \( r = (I - H) y \in \mathcal{N}(X^T) \). The orthogonal complement is called the residual vector space, and \( r \) is called the residual vector. Both the rank and the trace of the orthogonal complement are the number of rows in \( X \) (that is, the number of observations) minus the regression degrees of freedom. This quantity is the “residual degrees of freedom” (unadjusted).

These two projection matrices (8.53) or (8.55) and (8.58) partition the total sums of squares:
\[
y^T y = y^T H y + y^T (I - H) y. \quad (8.60)
\]
Note that the second term in this partitioning is the Schur complement of \( X^T X \) in \( [X \ y]^T [X \ y] \) (see equation (3.146) on page 96).

### Smoothing Matrices

The hat matrix, either from a full rank \( X \) as in equation (8.55) or formed by a generalized inverse as in equation (8.53), smooths the vector \( y \) onto the hyperplane defined by the column space of \( X \). It is therefore a smoothing matrix. (Note that the rank of the column space of \( X \) is the same as the rank of \( X^T X \).)

A useful variation of the cross products matrix \( X^T X \) is the matrix formed by adding a nonnegative (positive) definite matrix \( A \) to it. Because \( X^T X \) is nonnegative (positive) definite, \( X^T X + A \) is nonnegative definite, as we have seen (page 277), and hence \( X^T X + A \) is a Gramian matrix.

Because the square root of the nonnegative definite \( A \) exists, we can express the sum of the matrices as
\[
X^T X + A = \begin{bmatrix} X \\ A^{1/2} \end{bmatrix}^T \begin{bmatrix} X \\ A^{1/2} \end{bmatrix}. \quad (8.61)
\]

In a common application, a positive definite matrix \( \lambda I \), with \( \lambda > 0 \), is added to \( X^T X \), and this new matrix is used as a smoothing matrix. The analogue of the hat matrix (8.55) is
Matrices with Special Properties

\[ H_\lambda = X(X^T X + \lambda I)^{-1} X^T, \]  

(8.62)

and the analogue of the fitted response is

\[ \hat{y}_\lambda = H_\lambda y. \]  

(8.63)

This has the effect of shrinking the \( \hat{y} \) of equation (8.56) toward 0. (In regression analysis, this is called “ridge regression”.)

Any matrix such as \( H_\lambda \) that is used to transform the observed vector \( y \) onto a given subspace is called a smoothing matrix.

Effective Degrees of Freedom

Because of the shrinkage in ridge regression (that is, because the fitted model is less dependent just on the data in \( X \)) we say the “effective” degrees of freedom of a ridge regression model decreases with increasing \( \lambda \). We can formally define the effective model degrees of freedom of any linear fit \( \hat{y} = H_\lambda y \) as

\[ \text{tr}(H_\lambda), \]  

(8.64)

analogous to the model degrees of freedom in linear regression above. This definition of effective degrees of freedom applies generally in data smoothing. In fact, many smoothing matrices used in applications depend on a single smoothing parameter such as the \( \lambda \) in ridge regression, and so the same notation \( H_\lambda \) is often used for a general smoothing matrix.

To evaluate the effective degrees of freedom in the ridge regression model for a given \( \lambda \) and \( X \), for example, using singular value decomposition of \( X \), \( X = UDV^T \), we have

\[
\text{tr}(X(X^T X + \lambda I)^{-1} X^T) \\
= \text{tr} \left( UDV^T (VD^2 V^T + \lambda VV^T)^{-1} VDU^T \right) \\
= \text{tr} \left( UDV^T (V(D^2 + \lambda I)V^T)^{-1} VDU^T \right) \\
= \text{tr} \left( UD(D^2 + \lambda I)^{-1} DU^T \right) \\
= \text{tr} \left( D^2(D^2 + \lambda I)^{-1} \right) \\
= \sum \frac{d_i^2}{d_i^2 + \lambda}.
\]  

(8.65)

When \( \lambda = 0 \), this is the same as the ordinary model degrees of freedom, and when \( \lambda \) is positive, this quantity is smaller, as we would want it to be by the argument above. The \( d_i^2/(d_i^2 + \lambda) \) are called shrinkage factors.

If \( X^T X \) is not of full rank, the addition of \( \lambda I \) to it also has the effect of yielding a full rank matrix, if \( \lambda > 0 \), and so the inverse of \( X^T X + \lambda I \) exists even when that of \( X^T X \) does not. In any event, the addition of \( \lambda I \) to \( X^T X \) yields a matrix with a better “condition number”, which we define in Section 6.1. (On page 206, we return to this model and show that the condition number of \( X^T X + \lambda I \) is better than that of \( X^T X \).)
Residuals from Smoothed Data

Just as in equation (8.59), we can write

\[ y = \hat{y}_\lambda + r_\lambda. \]  

(8.66)

Notice, however, that \( H_\lambda \) is not in general a projection matrix. Unless \( H_\lambda \)

is a projection matrix, however, \( \hat{y}_\lambda \) and \( r_\lambda \) are not orthogonal as are \( \hat{y} \) and \( r \), and we do not have the additive partitioning of the sum of squares as in equation (8.60).

The rank of \( H_\lambda \) is the same as the number of columns of \( X \), but the trace, and hence the model degrees of freedom, is less than this number.

8.6.3 Centered Matrices and Variance-Covariance Matrices

In Section 2.3, we defined the variance of a vector and the covariance of two vectors. These are the same as the “sample variance” and “sample covariance” in statistical data analysis and are related to the variance and covariance of random variables in probability theory. We now consider the variance-covariance matrix associated with a data matrix. We occasionally refer to the variance-covariance matrix simply as the “variance matrix” or just as the “variance”.

First, we consider centering and scaling data matrices.

Centering and Scaling of Data Matrices

When the elements in a vector represent similar measurements or observational data on a given phenomenon, summing or averaging the elements in the vector may yield meaningful statistics. In statistical applications, the columns in a matrix often represent measurements on the same feature or on the same variable over different observational units as in Figure 8.1, and so the mean of a column may be of interest.

We may center the column by subtracting its mean from each element in the same manner as we centered vectors on page 34. The matrix formed by centering all of the columns of a given matrix is called a centered matrix, and if the original matrix is \( X \), we represent the centered matrix as \( X_c \) in a notation analogous to what we introduced for centered vectors. If we represent the matrix whose \( i \)th column is the constant mean of the \( i \)th column of \( X \) as \( \bar{X} \),

\[ X_c = X - \bar{X}. \]  

(8.67)

Here is an R statement to compute this:

\[ Xc <- X-rep(1,n)\%\%t(apply(X,2,mean)) \]
If the unit of a measurement is changed, all elements in a column of the data matrix in which the measurement is used will change. The amount of variation of elements within a column or the relative variation among different columns ideally should not be measured in terms of the basic units of measurement, which can differ irreconcilably from one column to another. (One column could represent scores on an exam and another column could represent weight, for example.)

In analyzing data, it is usually important to scale the variables so that their variations are comparable. We do this by using the standard deviation of the column. If we have also centered the columns, the column vectors are the centered and scaled vectors of the form of those in equation (2.51),

\[ x_{cs} = \frac{x_c}{s_x}, \]

where \( s_x \) is the standard deviation of \( x \),

\[ s_x = \frac{\|x_c\|}{\sqrt{n-1}}. \]

If all columns of the data matrix \( X \) are centered and scaled, we denote the resulting matrix as \( X_{cs} \). If \( s_i \) represents the standard deviation of the \( i^{th} \) column, this matrix is formed as

\[ X_{cs} = X_c \text{diag}(1/s_i). \] (8.68)

Here is an R statement to compute this:

\texttt{Xcs <- Xc%*%diag(1/apply(X,2,sd))}

If the rows of \( X \) are taken as representative of a population of similar vectors, it is often useful to center and scale any vector from that population in the manner of equation (8.68):

\[ \tilde{x} = \text{diag}(1/s_i)x_c. \] (8.69)

(Note that \( x_c \) is a vector of the same order as a row of \( X_c \).)

**Gramian Matrices Formed from Centered Matrices; Covariance Matrices**

An important Gramian matrix is formed as the sums of squares and cross products matrix from a centered matrix and scaled by \((n - 1)\), where \( n \) is the number of rows of the original matrix:

\[ S_X = \frac{1}{n-1}X_c^T X_c = (s_{ij}). \] (8.70)
This matrix is called the **variance-covariance matrix** associated with the given matrix $X$, and we denote it by $S_X$ or just $S$. If $x_{si}$ and $x_{sj}$ are the vectors corresponding to the $i^{th}$ and $j^{th}$ columns of $X$, then $s_{ij} = \text{Cov}(x_{si}, x_{sj})$; that is, the off-diagonal elements are the covariances between the column vectors, as in equation (2.55), and the diagonal elements are **variances** of the column vectors.

This matrix and others formed from it, such as $R_X$ in equation (8.72) below, are called **association matrices** because they are based on measures of association (covariance or correlation) among the columns of $X$. We could likewise define a Gramian association matrix based on measures of association among the rows of $X$.

A transformation using the Cholesky factor of $S_X$ or the square root of $S_X$ (assuming $S_X$ is full rank) results in a matrix whose associated variance-covariance is the identity. We call this a spered matrix:

$$X_{\text{sphered}} = X_cS_X^{-\frac{1}{2}}.$$  \hspace{1cm} (8.71)

The matrix $S_X$ is a measure of the anisometry of the space of vectors represented by the rows of $X$ as mentioned in Section 3.2.8. The inverse, $S_X^{-1}$, in some sense evens out the anisometry. Properties of vectors in the space represented by the rows of $X$ are best assessed following a transformation as in equation (8.69). For example, rather than orthogonality of two vectors $u$ and $v$, a more interesting relationship would be $S_X^{-1}$-conjugacy (see equation (3.65)):

$$u^T S_X^{-1} v = 0.$$  

Also, the Mahalanobis distance, \(\sqrt{(u - v)^T S_X^{-1} (u - v)}\), may be more relevant for measuring the difference in two vectors than the standard Euclidean distance.

**Gramian Matrices Formed from Scaled Centered Matrices; Correlation Matrices**

If the columns of a centered matrix are standardized (that is, divided by their standard deviations, assuming that each is nonconstant, so that the standard deviation is positive), the scaled cross products matrix is the **correlation matrix**, often denoted by $R_X$ or just $R$,

$$R_X = \frac{1}{n-1}X_{cs}^T X_{cs}$$

$$= (r_{ij}),$$  \hspace{1cm} (8.72)

where if $x_{si}$ and $x_{sj}$ are the vectors corresponding to the $i^{th}$ and $j^{th}$ columns of $X$, $r_{ij} = \text{Corr}(x_{si}, x_{sj})$. The correlation matrix can also be expressed as $R_X =$
\[ X^TDX_c, \] where \( D \) is the diagonal matrix whose \( k \)th diagonal is \( 1/\sqrt{V(x_{*k})} \), where \( V(x_{*k}) \) is the sample variance of the \( k \)th column; that is, \( V(x_{*k}) = \sum_i (x_{ik} - \bar{x}_{*k})^2/(n - 1) \). This Gramian matrix \( R_X \) is based on measures of association among the columns of \( X \).

The elements along the diagonal of the correlation matrix are all 1, and the off-diagonals are between \(-1\) and 1, each being the correlation between a pair of column vectors, as in equation (2.57). The correlation matrix is nonnegative definite because it is a Gramian matrix.

The trace of an \( n \times n \) correlation matrix is \( n \), and therefore the eigenvalues, which are all nonnegative, sum to \( n \).

Without reference to a data matrix, any nonnegative definite matrix with 1s on the diagonal and with all elements less than or equal to 1 in absolute value is called a correlation matrix.

### 8.6.4 The Generalized Variance

The diagonal elements of the variance-covariance matrix \( S \) associated with the \( n \times m \) data matrix \( X \) are the second moments of the centered columns of \( X \), and the off-diagonal elements are pairwise second central moments of the columns. Each element of the matrix provides some information about the spread of a single column or of two columns of \( X \). The determinant of \( S \) provides a single overall measure of the spread of the columns of \( X \). This measure, \( |S| \), is called the generalized variance, or generalized sample variance, to distinguish it from an analogous measure of a distributional model.

On page 57, we discussed the equivalence of a determinant and the volume of a parallelotope. The generalized variance captures this, and when the columns or rows of \( S \) are more orthogonal to each other, the volume of the parallelotope determined by the columns or rows of \( S \) is greater, as shown in Figure 8.6 for \( m = 3 \).

![Fig. 8.6. Generalized Variances in Terms of the Columns of S](image)

The columns or rows of \( S \) are generally not of much interest in themselves. Our interest is in the relationship of the centered columns of the \( n \times m \) data
matrix $X$. Let us consider the case of $m = 2$. Let $z_{*1}$ and $z_{*2}$ represent the centered column vectors of $X$; that is, for $z_{*1}$, we have $z_{*1i} = x_{*1i} - \bar{x}_1$. Now, as in equation (3.33), consider the parallelogram formed by $z_{*1}$ and $z_{*2}$. For computing the area, consider $z_{*1}$ as forming the base. The length of the base is

$$\|z_{*1}\| = \sqrt{(n - 1)s_{11}},$$

and the height is

$$\|z_{*2}\|\sin(\theta) = \sqrt{(n - 1)s_{22}(1 - r_{12}^2)}.$$

(Recall the relationship between the angle and the correlation from equation (2.58).)

The area of the parallelogram therefore is

$$\text{area} = (n - 1)\sqrt{s_{11}s_{22}(1 - r_{12}^2)}.$$

Now, consider $S$:

$$S = \begin{bmatrix} s_{11} & s_{21} \\ s_{12} & s_{22} \end{bmatrix} = \begin{bmatrix} s_{11} & \sqrt{s_{11}s_{22}r_{12}} \\ \sqrt{s_{11}s_{22}r_{12}} & s_{22} \end{bmatrix}. $$

The determinant of $S$ is therefore

$$s_{11}s_{22}(1 - r_{12}^2),$$

that is,

$$|S| = \frac{1}{(n - 1)m}\text{volume}^2. \quad (8.73)$$

Although we considered only the case $m = 2$, equation (8.73) holds generally, as can be seen by induction on $m$ (see Anderson, 2003).

### Comparing Variance-Covariance Matrices

Many standard statistical procedures for comparing groups of data rely on the assumption that the population variance-covariance matrices of the groups are all the same. (The simplest example of this is the two-sample $t$-test, in which the concern is just that the population variances of the two groups be equal.) Occasionally, the data analyst wishes to test this assumption of homogeneity of variances.

On page 138, we considered the problem of comparing two matrices of the same size. There we defined a metric based on a matrix norm. For the
problem of comparing variance-covariance matrices, a measure based on the
generalized variances is more commonly used.

In the typical situation, we have an \( n \times m \) data matrix \( X \) in which the first
\( n_1 \) rows represent observations from one group, the next \( n_2 \) rows represent
observations from another group, and the last \( n_g \) rows represent observations
from the \( g^{th} \) group. For each group, we form a sample variance-covariance
matrix \( S_i \) as in equation (8.70) using the \( i^{th} \) submatrix of \( X \). Whenever we
have individual variance-covariance matrices in situations similar to this, we
define the pooled variance-covariance matrix:

\[
S_p = \frac{1}{(n - g)} \sum_{i=1}^{g} (n_i - 1)S_i.
\] (8.74)

Bartlett suggested a test based on the determinants of \((n_i - 1)S_i\). (From
equation (3.27), \(|(n_i - 1)S_i| = (n_i - 1)^m|S_i|\).) A similar test suggested by
Box also uses the generalized variances. One form of the Box \( M \) statistic for
testing for homogeneity of variances is

\[
M = (n - g) \log(|S_p|) - \sum_{i=1}^{g} (n_i - 1)S_i.
\] (8.75)

8.6.5 Similarity Matrices

Covariance and correlation matrices are examples of similarity association ma-
trices: they measure the similarity or closeness of the columns of the matrices
from which they are formed.

The cosine of the angle between two vectors is related to the correlation
between the vectors, so a matrix of the cosine of the angle between the columns
of a given matrix would also be a similarity matrix. (The angle is exactly the
same as the correlation if the vectors are centered; see equation (2.57).)

Similarity matrices can be formed in many ways, and some are more useful
in particular applications than in others. They may not even arise from a
standard dataset in the familiar form in statistical applications. For example,
we may be interested in comparing text documents. We might form a vector-
like object whose elements are the words in the document. The similarity
between two such ordered tuples, generally of different lengths, may be a
count of the number of words, word pairs, or more general phrases in common
between the two documents.

It is generally reasonable that similarity between two objects be symmet-
ic; that is, the first object is as close to the second as the second is to the first.
We reserve the term similarity matrix for matrices formed from such measures
and, hence, that themselves are symmetric. Occasionally, for example in psy-
chometric applications, the similarities are measured relative to rank order
closeness within a set. In such a case, the measure of closeness may not be
symmetric. A matrix formed from such measures is called a directed similarity
matrix.
8.6.6 Dissimilarity Matrices

The elements of similarity generally increase with increasing “closeness”. We may also be interested in the dissimilarity. Clearly, any decreasing function of similarity could be taken as a reasonable measure of dissimilarity. There are, however, other measures of dissimilarity that often seem more appropriate. In particular, the properties of a metric (see page 22) may suggest that a dissimilarity be defined in terms of a metric.

In considering either similarity or dissimilarity for a data matrix, we could work with either rows or columns, but the common applications make one or the other more natural for the specific application. Because of the types of the common applications, we will discuss dissimilarities and distances in terms of rows instead of columns; thus, in this section, we will consider the dissimilarity of \( x_{i*} \) and \( x_{j*} \), the vectors corresponding to the \( i^{th} \) and \( j^{th} \) rows of \( X \).

Dissimilarity matrices are also association matrices in the general sense of Section 8.1.4.

Dissimilarity or distance can be measured in various ways. A metric is the most obvious measure, although in certain applications other measures are appropriate. The measures may be based on some kind of ranking, for example. If the dissimilarity is based on a metric, the association matrix is often called a distance matrix. In most applications, the Euclidean distance, \( \|x_{i*} - x_{j*}\|_2 \), is the most commonly used metric, but others, especially ones based on \( L_1 \) or \( L_\infty \) norms, are often useful.

Any hollow matrix with nonnegative elements is a directed dissimilarity matrix. A directed dissimilarity matrix is also called a cost matrix. If the matrix is symmetric, it is a dissimilarity matrix.

An \( n \times n \) matrix \( D = (d_{ij}) \) is an \( m \)-dimensional distance matrix if there exists \( m \)-vectors \( x_1, \ldots, x_n \) such that, for some metric \( \Delta \), \( d_{ij} = \Delta(x_i, x_j) \). A distance matrix is necessarily a dissimilarity matrix. If the metric is the Euclidean distance, the matrix \( D \) is called a Euclidean distance matrix.

The matrix whose rows are the vectors \( x_1^T, \ldots, x_n^T \) is called an associated configuration matrix of the given distance matrix. In metric multidimensional scaling, we are given a dissimilarity matrix and seek to find a configuration matrix whose associated distance matrix is closest to the dissimilarity matrix, usually in terms of the Frobenius norm of the difference of the matrices (see Trosset, 2002, for basic definitions and extensions).

8.7 Nonnegative and Positive Matrices

A nonnegative matrix, as the name suggests, is a real matrix all of whose elements are nonnegative, and a positive matrix is a real matrix all of whose elements are positive. In some other literature, the latter type of matrix is
called strictly positive, and a nonnegative matrix with a positive element is called positive.

Many useful matrices are nonnegative. We have already considered various kinds of nonnegative matrices. The adjacency or connectivity of a graph is nonnegative. Dissimilarity matrices, including distance matrices, are nonnegative. Matrices used in modeling stochastic processes are nonnegative.

If $A$ is nonnegative, we write

$$A \geq 0,$$  \hfill (8.76)

and if it is positive, we write

$$A > 0.$$ \hfill (8.77)

Notice that $A \geq 0$ and $A \neq 0$ together do not imply $A > 0$.

We write

$$A \geq B$$

to mean $(A - B) \geq 0$ and

$$A > B$$

to mean $(A - B) > 0$. (Recall the definitions of nonnegative definite and positive definite matrices, and, from equations (8.11) and (8.16), the notation used to indicate those properties, $A \succeq 0$ and $A > 0$. Furthermore, notice that these definitions and this notation for nonnegative and positive matrices are consistent with analogous definitions and notation involving vectors on page 13. Some authors, however, use the notation of equations (8.76) and (8.77) to mean “nonnegative definite” and “positive definite”. We should also note that some authors use somewhat different terms for these and related properties. “Positive” for these authors means nonnegative with at least one positive element, and “strictly positive” means positive as we have defined it.)

Notice that positiveness (nonnegativeness) has nothing to do with positive (nonnegative) definiteness. A positive or nonnegative matrix need not be symmetric or even square, although most such matrices useful in applications are square. A square positive matrix, unlike a positive definite matrix, need not be of full rank.

The following properties are easily verified.

1. If $A \geq 0$ and $u \geq v \geq 0$, then $Au \geq Av$.
2. If $A \geq 0$, $A \neq 0$, and $u > v > 0$, then $Au > Av$.
3. If $A > 0$ and $v \geq 0$, then $Av \geq 0$.
4. If $A > 0$ and $A$ is square, then $\rho(A) > 0$.

Whereas most of the important matrices arising in the analysis of linear models are symmetric, and thus have the properties listed on page 270, many important nonnegative matrices, such as those used in studying stochastic processes, are not necessarily symmetric. The eigenvalues of real symmetric matrices are real, but the eigenvalues of real nonsymmetric matrices may have
an imaginary component. In the following discussion, we must be careful to remember the meaning of the spectral radius. The definition in equation (3.185) for the spectral radius of the matrix $A$ with eigenvalues $c_i$,

$$\rho(A) = \max |c_i|,$$

is still correct, but the operator “$|\cdot|$” must be interpreted as the modulus of a complex number.

### 8.7.1 Properties of Square Positive Matrices

We have the following important properties for square positive matrices. These properties collectively are the conclusions of the Perron theorem.

Let $A$ be a square positive matrix and let $r = \rho(A)$. Then:

1. $r$ is an eigenvalue of $A$. The eigenvalue $r$ is called the Perron root. Note that the Perron root is real (although other eigenvalues of $A$ may not be).
2. There is an eigenvector $v$ associated with $r$ such that $v > 0$.
3. The Perron root is simple. (That is, the algebraic multiplicity of the Perron root is 1.)
4. The dimension of the eigenspace of the Perron root is 1. (That is, the geometric multiplicity of $\rho(A)$ is 1.) Hence, if $v$ is an eigenvector associated with $r$, it is unique except for scaling. This associated eigenvector is called the Perron vector. Note that the Perron vector is real (although other eigenvectors of $A$ may not be). The elements of the Perron vector all have the same sign, which we usually take to be positive; that is, $v > 0$.
5. If $c_i$ is any other eigenvalue of $A$, then $|c_i| < r$. (That is, $r$ is the only eigenvalue on the spectral circle of $A$.)

We will give proofs only of properties 1 and 2 as examples. Proofs of all of these facts are available in Horn and Johnson (1991).

To see properties 1 and 2, first observe that a positive matrix must have at least one nonzero eigenvalue because the coefficients and the constant in the characteristic equation must all be positive. Now scale the matrix so that its spectral radius is 1 (see page 111). So without loss of generality, let $A$ be a scaled positive matrix with $\rho(A) = 1$. Now let $(c, x)$ be some eigenpair of $A$ such that $|c| = 1$. First, we want to show, for some such $c$, that $c = \rho(A)$.

Because all elements of $A$ are positive,

$$|x| = |Ax| \leq A|x|,$$

and so

$$A|x| - |x| \geq 0. \quad (8.78)$$

An eigenvector must be nonzero, so we also have
\[ A|x| > 0. \]

Now we want to show that \( A|x| - |x| = 0. \) To that end, suppose the contrary; that is, suppose \( A|x| - |x| \neq 0. \) In that case, \( A(A|x| - |x|) > 0 \) from equation (8.78), and so there must be a positive number \( \epsilon \) such that

\[ \frac{A}{1 + \epsilon} A|x| > A|x| \]
or

\[ By > y, \]

where \( B = A/(1 + \epsilon) \) and \( y = A|x|. \) Now successively multiplying both sides of this inequality by the positive matrix \( B, \) we have

\[ B^k y > y \quad \text{for all} \quad k = 1, 2, \ldots. \]

Because \( \rho(B) = \rho(A)/(1 + \epsilon) < 1, \) from equation (3.247) on page 136, we have \( \lim_{k \to \infty} B^k = 0; \) that is, \( \lim_{k \to \infty} B^k y = 0 > y. \) This contradicts the fact that \( y > 0. \) Because the supposition \( A|x| - |x| \neq 0 \) led to this contradiction, we must have \( A|x| - |x| = 0. \) Therefore \( 1 = \rho(A) \) must be an eigenvalue of \( A, \) and \( |x| \) must be an associated eigenvector; hence, with \( v = |x|, \) \( (\rho(A), v) \) is an eigenpair of \( A \) and \( v > 0, \) and this is the statement made in properties 1 and 2.

The Perron-Frobenius theorem, which we consider below, extends these results to a special class of square nonnegative matrices. (This class includes all positive matrices, so the Perron-Frobenius theorem is an extension of the Perron theorem.)

### 8.7.2 Irreducible Square Nonnegative Matrices

Nonnegativity of a matrix is not a very strong property. First of all, note that it includes the zero matrix; hence, clearly none of the properties of the Perron theorem can hold. Even a nondegenerate, full rank nonnegative matrix does not necessarily possess those properties. A small full rank nonnegative matrix provides a counterexample for properties 2, 3, and 5:

\[
A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.
\]

The eigenvalues are 1 and 1; that is, 1 with an algebraic multiplicity of 2 (so property 3 does not hold). There is only one nonnull eigenvector, \((1, -1), \) (so property 2 does not hold, but property 4 holds), but the eigenvector is not positive (or even nonnegative). Of course property 5 cannot hold if property 3 does not hold.

We now consider irreducible square nonnegative matrices. This class includes positive matrices. On page 268, we defined reducibility of a nonnegative
square matrix and we saw that a matrix is irreducible if and only if its digraph is strongly connected.

To recall the definition, a nonnegative matrix is said to be reducible if by symmetric permutations it can be put into a block upper triangular matrix with square blocks along the diagonal; that is, the nonnegative matrix $A$ is reducible if and only if there is a permutation matrix $E_\pi$ such that

$$E_\pi^T AE_\pi = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}, \quad (8.79)$$

where $B_{11}$ and $B_{22}$ are square. A matrix that cannot be put into that form is irreducible. An alternate term for reducible is decomposable, with the associated term indecomposable. (There is an alternate meaning for the term “reducible” applied to a matrix. This alternate use of the term means that the matrix is capable of being expressed by a similarity transformation as the sum of two matrices whose columns are mutually orthogonal.)

We see from the definition in equation (8.79) that a positive matrix is irreducible.

Irreducible matrices have several interesting properties. An $n \times n$ nonnegative matrix $A$ is irreducible if and only if $(I + A)^{n-1}$ is a positive matrix; that is,

$$A \text{ is irreducible } \iff (I + A)^{n-1} > 0. \quad (8.80)$$

To see this, first assume $(I + A)^{n-1} > 0$; thus, $(I + A)^{n-1}$ clearly is irreducible. If $A$ is reducible, then there exists a permutation matrix $E_\pi$ such that

$$E_\pi^T AE_\pi = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix},$$

and so

$$E_\pi^T (I + A)^{n-1} E_\pi = (E_\pi^T (I + A) E_\pi)^{n-1}$$

$$= (I + E_\pi^T A E_\pi)^{n-1}$$

$$= \begin{bmatrix} I_{n_1} + B_{11} & B_{12} \\ 0 & I_{n_2} + B_{22} \end{bmatrix}. \quad (8.80)$$

This decomposition of $(I + A)^{n-1}$ cannot exist because it is irreducible; hence we conclude $A$ is irreducible if $(I + A)^{n-1} > 0$.

Now, if $A$ is irreducible, we can see that $(I + A)^{n-1}$ must be a positive matrix either by a strictly linear-algebraic approach or by couching the argument in terms of the digraph $G(A)$ formed by the matrix, as in the discussion on page 268 that showed that a digraph is strongly connected if (and only if) it is irreducible. We will use the latter approach in the spirit of applications of irreducibility in stochastic processes.

For either approach, we first observe that the $(i, j)^{th}$ element of $(I + A)^{n-1}$ can be expressed as
\[(I + A)^{n-1})_{ij} = \left( \sum_{k=0}^{n-1} \binom{n-1}{k} A^k \right)_{ij} \quad (8.81)\]

Hence, for \(k = 1, \ldots, n - 1\), we consider the \((i, j)^{\text{th}}\) entry of \(A^k\). Let \(a_{ij}^{(k)}\) represent this quantity.

Given any pair \((i, j)\), for some \(l_1, l_2, \ldots, l_{k-1}\), we have

\[a_{ij}^{(k)} = \sum_{l_1, l_2, \ldots, l_{k-1}} a_{l_1i}a_{l_1l_2} \cdots a_{l_{k-1}j}.\]

Now \(a_{ij}^{(k)} > 0\) if and only if \(a_{l_1i}, a_{l_1l_2}, \ldots, a_{l_{k-1}j}\) are all positive; that is, if there is a path \(v_1, v_1, \ldots, v_{l_{k-1}}, v_j\) in \(G(A)\). If \(A\) is irreducible, then \(G(A)\) is strongly connected, and hence the path exists. So, for any pair \((i, j)\), we have from equation (8.81) \(((I + A)^{n-1})_{ij} > 0\); that is, \((I + A)^{n-1} > 0\).

The positivity of \((I + A)^{n-1}\) for an irreducible nonnegative matrix \(A\) is a very useful property because it allows us to extend some conclusions of the Perron theorem to irreducible nonnegative matrices.

**Properties of Square Irreducible Nonnegative Matrices; the Perron-Frobenius Theorem**

If \(A\) is a square irreducible nonnegative matrix, then we have the following properties, which are similar to properties 1 through 4 on page 301 for positive matrices. These following properties are the conclusions of the Perron-Frobenius theorem.

1. \(\rho(A)\) is an eigenvalue of \(A\). This eigenvalue is called the Perron root, as before.
2. The Perron root \(\rho(A)\) is simple. (That is, the algebraic multiplicity of the Perron root is 1.)
3. The dimension of the eigenspace of the Perron root is 1. (That is, the geometric multiplicity of \(\rho(A)\) is 1.)
4. The eigenvector associated with \(\rho(A)\) is positive. This eigenvector is called the Perron vector, as before.

The relationship (8.80) allows us to prove properties 1 and 4 in a method similar to the proofs of properties 1 and 2 for positive matrices. (This is Exercise 8.9.) Complete proofs of all of these facts are available in Horn and Johnson (1991). See also the solution to Exercise 8.10b on page 498 for a special case.

The one property of square positive matrices that does not carry over to square irreducible nonnegative matrices is property 5: \(r = \rho(A)\) is the only eigenvalue on the spectral circle of \(A\). For example, the small irreducible nonnegative matrix

\[
A = \begin{bmatrix}
0 & 1 \\
1 & 0
\end{bmatrix}
\]
has eigenvalues 1 and $-1$, and so both are on the spectral circle.

It turns out, however, that square irreducible nonnegative matrices that have only one eigenvalue on the spectral circle also have other interesting properties that are important, for example, in Markov chains. We therefore give a name to the property:

A square irreducible nonnegative matrix is said to be \textit{primitive} if it has only one eigenvalue on the spectral circle.

In modeling with Markov chains and other applications, the limiting behavior of $A^k$ is an important property. On page 135, we saw that $\lim_{k \to \infty} A^k = 0$ if $\rho(A) < 1$. For a primitive matrix, we also have a limit for $A^k$ if $\rho(A) = 1$. (As we have done above, we can scale any matrix with a nonzero eigenvalue to a matrix with a spectral radius of 1.)

If $A$ is a primitive matrix, then we have the useful result

$$\lim_{k \to \infty} \left( \frac{A}{\rho(A)} \right)^k = vw^T,$$

(8.82)

where $v$ is an eigenvector of $A$ associated with $\rho(A)$ and $w$ is an eigenvector of $A^T$ associated with $\rho(A)$, and $w$ and $v$ are scaled so that $w^Tv = 1$. (As we mentioned on page 123, such eigenvectors exist because $\rho(A)$ is a simple eigenvalue. They also exist because of property 4; they are both positive. Note that $A$ is not necessarily symmetric, and so its eigenvectors may include imaginary components; however, the eigenvectors associated with $\rho(A)$ are real, and so we can write $w^T$ instead of $w^H$.)

To see equation (8.82), we consider $(A - \rho(A)vw^T)$. First, if $(c_i, v_i)$ is an eigenpair of $(A - \rho(A)vw^T)$ and $c_i \neq 0$, then $(c_i, v_i)$ is an eigenpair of $A$. We can see this by multiplying both sides of the eigen-equation by $vw^T$:

$$c_i vw^T v_i = vw^T (A - \rho(A)vw^T) v_i$$

$$= (vw^T A - \rho(A)vw^T vw^T) v_i$$

$$= (\rho(A)vw^T - \rho(A)vw^T) v_i$$

$$= 0;$$

hence,

$$Av_i = (A - \rho(A)vw^T) v_i$$

$$= c_i v_i.$$ 

Next, we show that

$$\rho \left( A - \rho(A)vw^T \right) < \rho(A).$$

(8.83)

If $\rho(A)$ were an eigenvalue of $(A - \rho(A)vw^T)$, then its associated eigenvector, say $w$, would also have to be an eigenvector of $A$, as we saw above. But since
as an eigenvalue of $A$ the geometric multiplicity of $\rho(A)$ is 1, for some scalar $s$, $w = sv$. But this is impossible because that would yield

$$\rho(A)sv = (A - \rho(A)vw^T) sv = sAv - s\rho(A)v = 0,$$

and neither $\rho(A)$ nor $sv$ is zero. But as we saw above, any eigenvalue of $(A - \rho(A)vw^T)$ is an eigenvalue of $A$ and no eigenvalue of $(A - \rho(A)vw^T)$ can be as large as $\rho(A)$ in modulus; therefore we have inequality (8.83).

Finally, we recall equation (3.212), with $w$ and $v$ as defined above, and with the eigenvalue $\rho(A)$,

$$(A - \rho(A)vw^T)^k = A^k - (\rho(A))^k vw^T, \quad (8.84)$$

for $k = 1, 2, \ldots$. Dividing both sides of equation (8.84) by $(\rho(A))^k$ and rearranging terms, we have

$$\left( \frac{A}{\rho(A)} \right)^k = vw^T + \frac{(A - \rho(A)vw^T)}{\rho(A)}.$$ \quad (8.85)

Now

$$\rho \left( \frac{(A - \rho(A)vw^T)}{\rho(A)} \right) = \frac{\rho(A - \rho(A)vw^T)}{\rho(A)},$$

which is less than 1; hence, from equation (3.245) on page 135, we have

$$\lim_{k \to \infty} \left( \frac{(A - \rho(A)vw^T)}{\rho(A)} \right)^k = 0;$$

so, taking the limit in equation (8.85), we have equation (8.82).

Applications of the Perron-Frobenius theorem are far-ranging. It has implications for the convergence of some iterative algorithms, such as the power method discussed in Section 7.2. The most important applications in statistics are in the analysis of Markov chains, which we discuss in Section 9.7.1.

### 8.7.3 Stochastic Matrices

A nonnegative matrix $A$ such that

$$P1 = 1 \quad (8.86)$$

is called a *stochastic matrix*. The definition means that $(1, 1)$ is an eigenpair of any stochastic matrix. It is also clear that if $P$ is a stochastic matrix, then $\|P\|_\infty = 1$ (see page 130), and because $\rho(P) \leq \|P\|$ for any norm (see page 134) and 1 is an eigenvalue of $P$, we have $\rho(P) = 1$. 
A stochastic matrix may not be positive, and it may be reducible or irreducible. (Hence, \((1, 1)\) may not be the Perron root and Perron eigenvector.) If \(P\) is a stochastic matrix such that
\[
1^T P = 1^T,
\]
(8.87)
it is called a doubly stochastic matrix. If \(P\) is a doubly stochastic matrix, \(\|P\|_1 = 1\), and, of course, \(\|P\|_\infty = 1\) and \(\rho(P) = 1\).

A permutation matrix is a doubly stochastic matrix; in fact, it is the simplest and one of the most commonly encountered doubly stochastic matrices. A permutation matrix is clearly reducible.

Stochastic matrices are particularly interesting because of their use in defining a discrete homogeneous Markov chain. In that application, a stochastic matrix and distribution vectors play key roles. A distribution vector is a nonnegative matrix whose elements sum to 1; that is, a vector \(v\) such that \(1^T v = 1\). In Markov chain models, the stochastic matrix is a probability transition matrix from a distribution at time \(t\), \(\pi_t\), to the distribution at time \(t + 1\),
\[
\pi_{t+1} = P\pi_t.
\]
In Section 9.7.1, we define some basic properties of Markov chains. Those properties depend in large measure on whether the transition matrix is reducible or not.

8.7.4 Leslie Matrices

Another type of nonnegative transition matrix, often used in population studies, is a Leslie matrix, after P. H. Leslie, who used it in models in demography. A Leslie matrix is a matrix of the form
\[
\begin{bmatrix}
\alpha_1 & \alpha_2 & \cdots & \alpha_{m-1} & \alpha_m \\
\sigma_1 & 0 & \cdots & 0 & 0 \\
0 & \sigma_2 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_{m-1} & 0
\end{bmatrix},
\]
(8.88)
where all elements are nonnegative, and additionally \(\sigma_i \leq 1\).

A Leslie matrix is clearly reducible. Furthermore, a Leslie matrix has a single unique positive eigenvalue (see Exercise 8.10), which leads to some interesting properties (see Section 9.7.2).

8.8 Other Matrices with Special Structures

Matrices of a variety of special forms arise in statistical analyses and other applications. For some matrices with special structure, specialized algorithms
can increase the speed of performing a given task considerably. Many tasks involving matrices require a number of computations of the order of $n^3$, where $n$ is the number of rows or columns of the matrix. For some of the matrices discussed in this section, because of their special structure, the order of computations may be $n^2$. The improvement from $O(n^3)$ to $O(n^2)$ is enough to make some tasks feasible that would otherwise be infeasible because of the time required to complete them. The collection of papers in Olshevsky (2003) describe various specialized algorithms for the kinds of matrices discussed in this section.

### 8.8.1 Helmert Matrices

A Helmert matrix is a square orthogonal matrix that partitions sums of squares. Its main use in statistics is in defining contrasts in general linear models to compare the second level of a factor with the first level, the third level with the average of the first two, and so on. (There is another meaning of “Helmert matrix” that arises from so-called Helmert transformations used in geodesy.)

For example, a partition of the sum $\sum_{i=1}^{n} y_i^2$ into orthogonal sums each involving $\bar{y}_k^2$ and $\sum_{i=1}^{k} (y_i - \bar{y}_k)^2$ is

$$\begin{align*}
\bar{y}_i &= (i(i + 1))^{-1/2} \left( \sum_{j=1}^{i+1} y_j - (i + 1)\bar{y}_{i+1} \right) \quad \text{for } i = 1, \ldots, n-1, \\
\bar{y}_n &= n^{-1/2} \sum_{j=1}^{n} y_j.
\end{align*}$$

(8.89)

These expressions lead to a computationally stable one-pass algorithm for computing the sample variance (see equation (10.7) on page 411).

The Helmert matrix that corresponds to this partitioning has the form

$$H_n = \begin{bmatrix}
\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\
\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -2/\sqrt{6} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & -\frac{(n-1)}{\sqrt{n(n-1)}}
\end{bmatrix}$$

(8.90)

$$H_n = \begin{bmatrix}
\frac{1}{\sqrt{n}} \mathbf{1}_n^T \\
\mathbf{K}_{n-1}
\end{bmatrix},$$

where $\mathbf{K}_{n-1}$ is the $(n - 1) \times n$ matrix below the first row. For the full $n$-vector $y$, we have
The rows of the matrix in equation (8.90) correspond to orthogonal contrasts in the analysis of linear models (see Section 9.2.2).

Obviously, the sums of squares are never computed by forming the Helmert matrix explicitly and then computing the quadratic form, but the computations in partitioned Helmert matrices are performed indirectly in analysis of variance, and representation of the computations in terms of the matrix is often useful in the analysis of the computations.

**8.8.2 Vandermonde Matrices**

A Vandermonde matrix is an $n \times m$ matrix with columns that are defined by monomials,

$$
V_{n \times m} = \begin{bmatrix}
1 & x_1 & x_1^2 & \cdots & x_1^{m-1} \\
1 & x_2 & x_2^2 & \cdots & x_2^{m-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_n & x_n^2 & \cdots & x_n^{m-1}
\end{bmatrix},
$$

where $x_i \neq x_j$ if $i \neq j$. The Vandermonde matrix arises in polynomial regression analysis. For the model equation $y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \epsilon_i$, given observations on $y$ and $x$, a Vandermonde matrix is the matrix in the standard representation $y = X\beta + \epsilon$.

Because of the relationships among the columns of a Vandermonde matrix, computations for polynomial regression analysis can be subject to numerical errors, and so sometimes we make transformations based on orthogonal polynomials. (The "condition number", which we define in Section 6.1, for a Vandermonde matrix is large.) A Vandermonde matrix, however, can be used to form simple orthogonal vectors that correspond to orthogonal polynomials. For example, if the $x$s are chosen over a grid on $[-1, 1]$, a QR factorization (see Section 5.7 on page 188) yields orthogonal vectors that correspond to Legendre polynomials. These vectors are called discrete Legendre polynomials. Although not used in regression analysis so often now, orthogonal vectors are useful in selecting settings in designed experiments.

Vandermonde matrices also arise in the representation or approximation of a probability distribution in terms of its moments.

The determinant of a square Vandermonde matrix has a particularly simple form (see Exercise 8.11).
8.8.3 Hadamard Matrices and Orthogonal Arrays

In a wide range of applications, including experimental design, cryptology, and other areas of combinatorics, we often encounter matrices whose elements are chosen from a set of only a few different elements. In experimental design, the elements may correspond to the levels of the factors; in cryptology, they may represent the letters of an alphabet. In two-level factorial designs, the entries may be either 0 or 1. Matrices all of whose entries are either 1 or \(-1\) can represent the same layouts, and such matrices may have interesting mathematical properties.

An \(n \times n\) matrix with \(-1, 1\) entries whose determinant is \(n^{n/2}\) is called a Hadamard matrix. The name comes from the bound derived by Hadamard for the determinant of any matrix \(A\) with \(|a_{ij}| \leq 1\) for all \(i, j\): \(|\det(A)| \leq n^{n/2}\). A Hadamard matrix achieves this upper bound. A maximal determinant is often used as a criterion for a good experimental design. One row and one column of an \(n \times n\) Hadamard matrix consist of all 1s; all \(n - 1\) other rows and columns consist of \(n/2\) 1s and \(n/2\) \(-1\)s. We often denote an \(n \times n\) Hadamard matrix by \(H_n\), which is the same notation often used for a Helmert matrix, but in the case of Hadamard matrices, the matrix is not unique. All rows are orthogonal and so are all columns. The norm of each row or column is \(n\), so \(H_n^T H_n = nI\).

A Hadamard matrix is often represented as a mosaic of black and white squares, as in Figure 8.7.

\[
\begin{bmatrix}
1 & -1 & -1 & 1 \\
1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 \\
1 & 1 & 1 & 1 \\
\end{bmatrix}
\]

Fig. 8.7. A \(4 \times 4\) Hadamard Matrix

Hadamard matrices do not exist for all \(n\). Clearly, \(n\) must be even because \(|H_n| = n^{n/2}\), but some experimentation (or an exhaustive search) quickly shows that there is no Hadamard matrix for \(n = 6\). It has been conjectured, but not proven, that Hadamard matrices exist for any \(n\) divisible by 4. Given any \(n \times n\) Hadamard matrix, \(H_n\), and any \(m \times m\) Hadamard matrix, \(H_m\), an \(nm \times nm\) Hadamard matrix can be formed as a partitioned matrix in which each 1 in \(H_n\) is replaced by the block submatrix \(H_m\) and each \(-1\) is replaced by the block submatrix \(-H_m\). For example, the \(4 \times 4\) Hadamard matrix shown in Figure 8.7 is formed using the \(2 \times 2\) Hadamard matrix

\[
\begin{bmatrix}
1 & -1 \\
1 & 1 \\
\end{bmatrix}
\]
as both $H_n$ and $H_m$. Not all Hadamard matrices can be formed from other Hadamard matrices in this way, however.

A somewhat more general type of matrix corresponds to an $n \times m$ array with the elements in the $j$th column being members of a set of $k_j$ elements and such that, for some fixed $p \leq m$, in every $n \times p$ submatrix all possible combinations of the elements of the $m$ sets occur equally often as a row. (I make a distinction between the matrix and the array because often in applications the elements in the array are treated merely as symbols without the assumptions of an algebra of a field. A terminology for orthogonal arrays has evolved that is different from the terminology for matrices; for example, a symmetric orthogonal array is one in which $k_1 = \cdots = k_m$. On the other hand, treating the orthogonal arrays as matrices with real elements may provide solutions to combinatorial problems such as may arise in optimal design.)

The $4 \times 4$ Hadamard matrix shown in Figure 8.7 is a symmetric orthogonal array with $k_1 = \cdots = k_4 = 2$ and $p = 4$, so in the array each of the possible combinations of elements occurs exactly once. This array is a member of a simple class of symmetric orthogonal arrays that has the property that in any two rows each ordered pair of elements occurs exactly once.

Orthogonal arrays are particularly useful in developing fractional factorial plans. (The robust designs of Taguchi correspond to orthogonal arrays.) Dey and Mukerjee (1999) discuss orthogonal arrays with an emphasis on the applications in experimental design, and Hedayat, Sloane, and Stufken (1999) provide extensive an discussion of the properties of orthogonal arrays.

### 8.8.4 Toeplitz Matrices

If the elements of the matrix $A$ are such that $a_{i,i+c_k} = d_{c_k}$, where $d_{c_k}$ is constant for fixed $c_k$, then $A$ is called a Toeplitz matrix,

$$
\begin{bmatrix}
  d_0 & d_1 & d_2 & \cdots & d_{n-1} \\
  d_{-1} & d_0 & d_1 & \cdots & d_{n-2} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  d_{-n+2} & d_{-n+3} & d_{-n+4} & \cdots & d_1 \\
  d_{-n+1} & d_{-n+2} & d_{-n+3} & \cdots & d_0 
\end{bmatrix},
$$

that is, a Toeplitz matrix is a matrix with constant codiagonals. A Toeplitz matrix may or may not be a band matrix (i.e., have many 0 codiagonals) and it may or may not be symmetric.

Banded Toeplitz matrices arise frequently in time series studies. The covariance matrix in an ARMA($p, q$) process, for example, is a symmetric Toeplitz matrix with $2 \max(p, q)$ nonzero off-diagonal bands. See page 364 for an example and further discussion.
Inverses of Toeplitz Matrices and Other Banded Matrices

A Toeplitz matrix that occurs often in stationary time series is the $n \times n$ variance-covariance matrix of the form

\[
V = \sigma^2 \begin{bmatrix}
1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\
\rho & 1 & \rho & \cdots & \rho^{n-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 
\end{bmatrix}.
\]

It is easy to see that $V^{-1}$ exists if $\sigma \neq 0$ and $\rho \neq 1$, and that it is the type 2 matrix

\[
V^{-1} = \frac{1}{(1 - \rho^2)\sigma^2} \begin{bmatrix}
1 & -\rho & 0 & \cdots & 0 \\
-\rho & 1 + \rho^2 & -\rho & \cdots & 0 \\
0 & -\rho & 1 + \rho^2 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 
\end{bmatrix}.
\]

Type 2 matrices also occur as the inverses of other matrices with special patterns that arise in other common statistical applications (see Graybill, 1983, for examples).

The inverses of all banded invertible matrices have off-diagonal submatrices that are zero or have low rank, depending on the bandwidth of the original matrix (see Strang and Nguyen, 2004, for further discussion and examples).

8.8.5 Hankel Matrices

A Hankel matrix is an $n \times m$ matrix $H(c, r)$ generated by an $n$-vector $c$ and an $m$-vector $r$ such that the $(i, j)$ element is

\[
\begin{cases}
c_{i+j-1} & \text{if } i+j-1 \leq n, \\
r_{i+j-n} & \text{otherwise}.
\end{cases}
\]

A common form of Hankel matrix is an $n \times n$ skew upper triangular matrix, and it is formed from the $c$ vector only. This kind of matrix occurs in the spectral analysis of time series. If $f(t)$ is a (discrete) time series, for $t = 0, 1, 2, \ldots$, the Hankel matrix of the time series has as the $(i, j)$ element

\[
f(i + j - 2) \text{ if } i + j - 1 \leq n, \\
0 \quad \text{otherwise}.
\]

The $L_2$ norm of the Hankel matrix of the time series (of the “impulse function”, $f$) is called the Hankel norm of the filter frequency response (the Fourier transform).

The simplest form of the square skew upper triangular Hankel matrix is formed from the vector $c = (1, 2, \ldots, n)$:
8.8.6 Cauchy Matrices

Another type of special $n \times m$ matrix whose elements are determined by a few $n$-vectors and $m$-vectors is a Cauchy-type matrix. The standard Cauchy matrix is built from two vectors, $x$ and $y$. The more general form defined below uses two additional vectors.

A Cauchy matrix is an $n \times m$ matrix $C(x, y, v, w)$ generated by $n$-vectors $x$ and $v$ and $m$-vectors $y$ and $w$ of the form

$$
C(x, y, v, w) = \begin{bmatrix}
v_1w_1 & \cdots & v_1w_m \\
 v_iw_1 & \cdots & v_iw_m \\
 x_n - y_1 & x_n - y_m \\
\end{bmatrix}.
$$

(8.92)

Cauchy-type matrices often arise in the numerical solution of partial differential equations (PDEs). For Cauchy matrices, the order of the number of computations for factorization or solutions of linear systems can be reduced from a power of three to a power of two. This is a very significant improvement for large matrices. In the PDE applications, the matrices are generally not large, but nevertheless, even in those applications, it is worthwhile to use algorithms that take advantage of the special structure. Fasino and Gemignani (2003) describe such an algorithm.

8.8.7 Matrices Useful in Graph Theory

Many problems in statistics and applied mathematics can be posed as graphs, and various methods of graph theory can be used in their solution.

Graph theory is particularly useful in cluster analysis or classification. These involve the analysis of relationships of objects for the purpose of identifying similar groups of objects. The objects are associated with vertices of the graph, and an edge is generated if the relationship (measured somehow) between two objects is sufficiently great. For example, suppose the question of interest is the authorship of some text documents. Each document is a vertex, and an edge between two vertices exists if there are enough words in common between the two documents. A similar application could be the determination of which computer user is associated with a given computer session. The vertices would correspond to login sessions, and the edges would be established based on the commonality of programs invoked or files accessed. In applications such as these, there would typically be a training dataset consisting of
text documents with known authors or consisting of session logs with known
users. In both of these types of applications, decisions would have to be made
about the extent of commonality of words, phrases, programs invoked, or files
accessed in order to establish an edge between two documents or sessions.

Unfortunately, as is often the case for an area of mathematics or statistics
that developed from applications in diverse areas or through the efforts of ap-
plied mathematicians somewhat outside of the mainstream of mathematics,
there are major inconsistencies in the notation and terminology employed in
graph theory. Thus, we often find different terms for the same object; for ex-
ample, adjacency matrix and connectivity matrix. This unpleasant situation,
however, is not so disagreeable as a one-to-many inconsistency, such as the
designation of the eigenvalues of a graph to be the eigenvalues of one type
of matrix in some of the literature and the eigenvalues of different types of
matrices in other literature.

**Adjacency Matrix; Connectivity Matrix**

We discussed adjacency or connectivity matrices on page 265. A matrix, such
as an adjacency matrix, that consists of only 1s and 0s is called a *Boolean
matrix*.

Two vertices that are not connected and hence correspond to a 0 in a
connectivity matrix are said to be *independent*.

If no edges connect a vertex with itself, the adjacency matrix is a hollow
matrix.

Because the 1s in a connectivity matrix indicate a strong association, and
we would naturally think of a vertex as having a strong association with
itself, we sometimes modify the connectivity matrix so as to have 1s along the
diagonal. Such a matrix is sometimes called an *augmented connectivity matrix*
or *augmented associativity matrix*.

The eigenvalues of the adjacency matrix reveal some interesting properties
of the graph and are sometimes called the *eigenvalues of the graph*. The eigen-
values of another matrix, which we discuss below, are more useful, however,
and we will refer to them as the eigenvalues of the graph.

**Digraphs**

The digraph represented in Figure 8.4 on page 266 is a network with five
vertices, perhaps representing cities, and directed edges between some of the
vertices. The edges could represent airline connections between the cities; for
example, there are flights from $x$ to $u$ and from $u$ to $x$, and from $y$ to $z$, but
not from $z$ to $y$.

In a digraph, the relationships are directional. (An example of a directional
relationship that might be of interest is when each observational unit has a
different number of measured features, and a relationship exists from $v_i$ to $v_j$
if a majority of the features of $v_i$ are identical to measured features of $v_j$.)
Use of the Connectivity Matrix

The analysis of a network may begin by identifying which vertices are connected with others; that is, by construction of the connectivity matrix.

The connectivity matrix can then be used to analyze other levels of association among the data represented by the graph or digraph. For example, from the connectivity matrix in equation (8.2) on page 266, we have

\[
C^2 = \begin{bmatrix}
4 & 1 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 2 \\
1 & 2 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1
\end{bmatrix}.
\]

In terms of the application suggested on page 266 for airline connections, the matrix \(C^2\) represents the number of connections between the cities that consist of exactly two flights. From \(C^2\) we see that there are two ways to go from city \(y\) to city \(w\) in just two flights but only one way to go from \(w\) to \(y\) in two flights.

A power of a connectivity matrix for a nondirected graph is symmetric.

The Laplacian Matrix of a Graph

Spectral graph theory is concerned with the analysis of the eigenvalues of a graph. As mentioned above, there are two different definitions of the eigenvalues of a graph. The more useful definition, and the one we use here, takes the eigenvalues of a graph to be the eigenvalues of a matrix, called the Laplacian matrix, formed from the adjacency matrix and a diagonal matrix consisting of the degrees of the vertices.

Given the graph \(G\), let \(D(G)\) be a diagonal matrix consisting of the degrees of the vertices of \(G\) (that is, \(D(G) = \text{diag}(d(G))\)) and let \(C(G)\) be the adjacency matrix of \(G\). If there are no isolated vertices (that is if \(d(G) > 0\)), then the Laplacian matrix of the graph, \(L(G)\) is given by

\[
L(G) = I - D(G)^{-\frac{1}{2}}C(G)D(G)^{-\frac{1}{2}}.
\]  \hspace{1cm} (8.93)

Some authors define the Laplacian in other ways:

\[
L_a(G) = I - D(G)^{-1}C(G)
\]  \hspace{1cm} (8.94)

or

\[
L_b(G) = D(G) - C(G).
\]  \hspace{1cm} (8.95)

The eigenvalues of the Laplacian matrix are the eigenvalues of a graph. The definition of the Laplacian matrix given in equation (8.93) seems to be more useful in terms of bounds on the eigenvalues of the graph. The set of unique eigenvalues (the spectrum of the matrix \(L\)) is called the spectrum of the graph.
So long as \( d(G) > 0 \), \( L(G) = D(G)^{-\frac{1}{2}} L_a(G) D(G)^{-\frac{1}{2}} \). Unless the graph is regular, the matrix \( L_b(G) \) is not symmetric. Note that if \( G \) is \( k \)-regular, \( L(G) = I - C(G)/k \), and \( L_b(G) = L(G) \).

For a digraph, the degrees are replaced by either the indegrees or the outdegrees. (Some authors define it one way and others the other way. The essential properties hold either way.)

The Laplacian can be viewed as an operator on the space of functions \( f : V(G) \rightarrow \mathbb{R} \) such that for the vertex \( v \)

\[
L(f(v)) = \frac{1}{\sqrt{d_v}} \sum_{w \sim v} \left( \frac{f(v)}{\sqrt{d_v}} - \frac{f(w)}{\sqrt{d_w}} \right),
\]

where \( w \sim v \) means vertices \( w \) and \( v \) that are adjacent, and \( d_u \) is the degree of the vertex \( u \).

The Laplacian matrix is symmetric, so its eigenvalues are all real. We can see that the eigenvalues are all nonnegative by forming the Rayleigh quotient (equation (3.209)) using an arbitrary vector \( g \), which can be viewed as a real-valued function over the vertices,

\[
R_L(g) = \frac{\langle g, Lg \rangle}{\langle g, g \rangle} = \frac{\langle g, D^{-\frac{1}{2}} L_a D^{-\frac{1}{2}} g \rangle}{\langle g, g \rangle} = \frac{\langle f, L_a f \rangle}{\langle D^{\frac{1}{2}} f, D^{\frac{1}{2}} f \rangle} = \frac{\sum_{v \sim w} (f(v) - f(w))^2}{f^T D f},
\]  

where \( f = D^{-\frac{1}{2}} g \), and \( f(u) \) is the element of the vector corresponding to vertex \( u \). Because the Raleigh quotient is nonnegative, all eigenvalues are nonnegative, and because there is an \( f \neq 0 \) for which the Rayleigh quotient is 0, we see that 0 is an eigenvalue of a graph. Furthermore, using the Cauchy-Schwartz inequality, we see that the spectral radius is less than or equal to 2.

The eigenvalues of a matrix are the basic objects in spectral graph theory. They provide information about the properties of networks and other systems modeled by graphs. We will not explore them further here, and the interested reader is referred to Chung (1997) or other general texts on the subject.

If \( G \) is the graph represented in Figure 8.2 on page 264, with \( V(G) = \{a, b, c, d, e\} \), the degrees of the vertices of the graph are \( d(G) = (4, 2, 2, 3, 3) \). Using the adjacency matrix given in equation (8.1), we have
\[ L(G) = \begin{bmatrix}
  1 - \frac{\sqrt{2}}{4} & -\frac{\sqrt{2}}{4} & -\frac{\sqrt{3}}{6} & -\frac{\sqrt{3}}{6} \\
  \frac{\sqrt{2}}{4} & 1 & 0 & 0 - \frac{\sqrt{6}}{6} \\
  -\frac{\sqrt{3}}{6} & 0 & 1 & -\frac{1}{3} \\
  \frac{\sqrt{3}}{6} & -\frac{\sqrt{6}}{6} & 0 & 1 - \frac{1}{3}
\end{bmatrix}. \tag{8.97} \]

This matrix is singular, and the unnormalized eigenvector corresponding to the 0 eigenvalue is \((2\sqrt{14}, 2\sqrt{7}, 2\sqrt{7}, \sqrt{42}, \sqrt{42})\).

### 8.8.8 M-Matrices

In certain applications in physics and in the solution of systems of nonlinear differential equations, a class of matrices called \(M\)-matrices is important. The matrices in these applications have nonpositive off-diagonal elements. A square matrix all of whose off-diagonal elements are nonpositive is called a \(Z\)-matrix.

A \(Z\)-matrix that is positive stable (see page 125) is called an \(M\)-matrix. A real symmetric \(M\)-matrix is positive definite.

In addition to the properties that constitute the definition, \(M\)-matrices have a number of remarkable properties, which we state here without proof. If \(A\) is a real \(M\)-matrix, then:

- all principal minors of \(A\) are positive;
- all diagonal elements of \(A\) are positive;
- all diagonal elements of \(L\) and \(U\) in the \(LU\) decomposition of \(A\) are positive;
- for any \(i\), \(\sum_j a_{ij} \geq 0;\) and
- \(A\) is nonsingular and \(A^{-1} \geq 0.\)

Proofs of these facts can be found in Horn and Johnson (1991).

### Exercises

8.1. Ordering of nonnegative definite matrices.

a) A relation \(\succ\) on a set is a **partial ordering** if, for elements \(a, b,\) and \(c,\)

- it is reflexive: \(a \succ a;\)
- it is antisymmetric: \(a \succ b \succ a \implies a = b;\) and
- it is transitive: \(a \succ b \succ c \implies a \succ c.\)

Show that the relation \(\succeq\) (equation (8.19)) is a partial ordering.
b) Show that the relation \( \succ \) (equation (8.20)) is transitive.

8.2. Show that a diagonally dominant symmetric matrix with positive diagonals is positive definite.

8.3. Show that the number of positive eigenvalues of an idempotent matrix is the rank of the matrix.

8.4. Show that two idempotent matrices of the same rank are similar.

8.5. Under the given conditions, show that properties (a) and (b) on page 285 imply property (c).

8.6. Projections.

a) Show that the matrix given in equation (8.42) (page 287) is a projection matrix.

b) Write out the projection matrix for projecting a vector onto the plane formed by two vectors, \( x_1 \) and \( x_2 \), as indicated on page 287, and show that it is the same as the hat matrix of equation (8.55).

8.7. Show that the matrix \( X^T X \) is symmetric (for any matrix \( X \)).


A correlation matrix can be defined in terms of a Gramian matrix formed by a centered and scaled matrix, as in equation (8.72). Sometimes in the development of statistical theory, we are interested in the properties of correlation matrices with given eigenvalues or with given ratios of the largest eigenvalue to other eigenvalues.

Write a program to generate \( n \times n \) random correlation matrices \( R \) with specified eigenvalues, \( c_1, \ldots, c_n \). The only requirements on \( R \) are that its diagonals be 1, that it be symmetric, and that its eigenvalues all be positive and sum to \( n \). Use the following method due to Davies and Higham (2000) that uses random orthogonal matrices with the Haar uniform distribution generated using the method described in Exercise 4.7.

0. Generate a random orthogonal matrix \( Q \); set \( k = 0 \), and form

\[
R^{(0)} = Q \text{diag}(c_1, \ldots, c_n)Q^T.
\]

1. If \( r_{ii}^{(k)} = 1 \) for all \( i \) in \( \{1, \ldots, n\} \), go to step 3.

2. Otherwise, choose \( p \) and \( q \) with \( p < j \), such that \( r_{pp}^{(k)} < 1 < r_{qq}^{(k)} \) or \( r_{pp}^{(k)} > 1 > r_{qq}^{(k)} \), and form \( G^{(k)} \) as in equation (5.13), where \( c \) and \( s \) are as in equations (5.17) and (5.17), with \( a = 1 \).

Form \( R^{(k+1)} = (G^{(k)})^T R^{(k)} G^{(k)} \).

Set \( k = k + 1 \), and go to step 1.

3. Deliver \( R = R^{(k)} \).

8.9. Use the relationship (8.80) to prove properties 1 and 4 on page 304.

8.10. Leslie matrices.

a) Write the characteristic polynomial of the Leslie matrix, equation (8.88).

b) Show that the Leslie matrix has a single, unique positive eigenvalue.
8.11. Write out the determinant for an $n \times n$ Vandermonde matrix.
8.12. Write out the determinant for the $n \times n$ skew upper triangular Hankel matrix in (8.91).