certain vectors (the eigenvectors). If \( Ax = \lambda x \) and \( |Ax| = 1 \), then
\[
|Ax| = |\lambda|x| = |\lambda|
\]
If \( \lambda \) is the eigenvalue with the greatest magnitude, then a corresponding unit eigenvector is an eigenvector in which the stretching effect of \( A \) is greatest. That is, the length of \( Ax \) is maximized when \( x = v_1 \), and \( |Av_1| = |\lambda||v_1| = |\lambda| \). This description of \( v_1 \) and \( |\lambda| \) has an analogue for rectangular matrices that will lead to the singular value decomposition.

**EXAMPLE 1** If \( A = \begin{bmatrix} \frac{4}{8} & \frac{11}{7} & \frac{14}{2} \\ \frac{8}{7} & \frac{7}{2} & \frac{9}{4} \end{bmatrix} \), then the linear transformation \( x \mapsto Ax \) maps the unit sphere \( \{x : |x| = 1\} \) in \( \mathbb{R}^2 \) onto an ellipse in \( \mathbb{R}^2 \), shown in Figure 1. Find a unit vector \( x \) at which the length \( |Ax| \) is maximized, and compute this maximum length.

**FIGURE 1** A transformation from \( \mathbb{R}^2 \) to \( \mathbb{R}^2 \).

**SOLUTION** The quantity \( |Ax|^2 \) is maximized at the same point that maximizes \( |Ax| \), and \( |Ax|^2 \) is easier to handle. Observe that
\[
|Ax| = |(Ax)^T Ax| = x(A^TA)x
\]
Also, \( A^TA \) is a symmetric matrix, since \( (A^TA)^T = A^TA \). So the problem now is to maximize the quadratic form \( x^TA^TAx \) subject to the constraint \( |x| = 1 \). By Theorem 6 in Section 7.3, the maximum value is the greatest eigenvalue \( \lambda_1 \) of \( A^TA \).

The maximum value is attained at a unit eigenvector of \( A^TA \) corresponding to \( \lambda_1 \), for the matrix \( A \) in this example, \( A^TA = \begin{bmatrix} 4 & 11 & 14 \\ 11 & 7 & 14 \\ 14 & 14 & 2 \end{bmatrix} \) is obtained. The eigenvalues of \( A^TA \) are \( \lambda_1 = 360, \lambda_2 = 90, \) and \( \lambda_3 = 40 \). Corresponding unit eigenvectors are, respectively,
\[
v_1 = \begin{bmatrix} 1/3 \\ 2/3 \\ -2/3 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -2/3 \\ 1/3 \\ 2/3 \end{bmatrix}, \quad v_3 = \begin{bmatrix} -2/3 \\ 1/3 \\ 2/3 \end{bmatrix}
\]
The maximum value of \( |Ax|^2 \), attained when \( x \) is the unit vector \( v_1 \), is the vector \( Av_1 \) is a point on the ellipse in Figure 1 farthest from the origin, namely,
\[
Av_1 = \begin{bmatrix} 4 & 11 & 14 \\ 11 & 7 & 14 \\ 14 & 14 & 2 \end{bmatrix} \begin{bmatrix} 1/3 \\ 2/3 \\ -2/3 \end{bmatrix} = \begin{bmatrix} 18 \\ 6 \\ 6 \end{bmatrix}
\]
For \( |x| = 1 \), the maximum value of \( |Ax| = |Av_1| = \sqrt{360} = 6\sqrt{10} \).
7.4 The Singular Value Decomposition

Thus \{A_1, \ldots, A_r\} is an orthogonal set. Furthermore, since the lengths of the vectors \(A_1, \ldots, A_r\) are the singular values of \(A\), and since there are \(r\) nonzero singular values, \(A_i \neq 0\) if and only if \(1 \leq i \leq r\). So \(A_1, \ldots, A_r\) are linearly independent vectors, and they are in \(\mathbb{C}^4\). Finally, for any \(y \in \text{Col } A\) and any \(x \in \text{Null } A\), we can write

\[
x = c_1 v_1 + \cdots + c_r v_r \quad \text{and} \quad y = Ax = c_1 A v_1 + \cdots + c_r A v_r = c_1 A v_1 + \cdots + c_r A v_r + 0 + \cdots + 0.
\]

Thus \(y\) is in \(\text{Span } \{A_1, \ldots, A_r\}\), which shows that \(\{A_1, \ldots, A_r\}\) is an (orthogonal) basis for \(\text{Col } A\). Hence rank \(A = \dim \text{Col } A = r\).

### NUMERICAL NOTE

In some cases, the rank of \(A\) may be very sensitive to small changes in the entries of \(A\). The obvious method of counting the number of pivot columns in \(A\) does not work well if \(A\) is now reduced by a computer. Roundoff error often creates an echelon form with full rank.

In practice, the most reliable way to estimate the rank of a large matrix \(A\) is to count the number of nonzero singular values. In this case, extremely small nonzero singular values are assumed to be zero for all practical purposes, and the effective rank of the matrix is the number obtained by counting the remaining nonzero singular values.

### The Singular Value Decomposition

The decomposition of \(A\) involves an \(m \times n\) “diagonal” matrix \(\Sigma\) of the form

\[
\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}
\]

where \(D\) is an \(r \times r\) diagonal matrix for some \(r\) not exceeding the smaller of \(m\) and \(n\). (If \(r\) equals \(m\) or \(n\) or both, some or all of the zero matrices do not appear.)

### The Singular Value Decomposition

Let \(A\) be an \(m \times n\) matrix with rank \(r\). Then there exists an \(m \times n\) matrix \(\Sigma\) as in (3) for which the diagonal entries in \(D\) are the first \(r\) singular values of \(A\), \(c_1 \geq c_2 \geq \ldots \geq c_r > 0\), and there exist an \(m \times m\) orthogonal matrix \(U\) and an \(n \times n\) orthogonal matrix \(V\) such that

\[
A = U \Sigma V^T.
\]

Any factorization \(A = U \Sigma V^T\), with \(U\) and \(V\) orthogonal, \(\Sigma\) as in (3), and positive diagonal entries in \(D\), is called a singular value decomposition (or SVD) of \(A\). The matrices \(U\) and \(V\) are not uniquely determined by \(A\), but the diagonal entries of \(\Sigma\) are necessarily the singular values of \(A\). See Exercise 19. The columns of \(U\) in such a decomposition are called left singular vectors of \(A\), and the columns of \(V\) are called right singular vectors of \(A\).

### THEOREM 9

Suppose \(\{v_1, \ldots, v_n\}\) is an orthonormal basis of \(\mathbb{R}^n\) consisting of eigenvectors of \(A^* A\), arranged so that the corresponding eigenvalues of \(A^* A\) satisfy \(\lambda_1 \geq \cdots \geq \lambda_n > 0\), and suppose \(A\) has \(r\) nonzero singular values. Then \(\{A_1, \ldots, A_r\}\) is an orthogonal basis for \(\text{Col } A\), and rank \(A = r\).

**Proof:** Because \(v_i\) and \(A_1\) are orthogonal for \(i \neq j\),

\[
(A v_i)^T (A v_j) = v_i^T A^* A v_j = v_i^T (\lambda_i v_i) = 0.
\]
PROOF. Let \( \lambda_i \) and \( v_i \) be as in Theorem 9, so that \( \{ Av_1, \ldots, Av_r \} \) is an orthogonal basis for Col \( A \). Normalize each \( Av_i \) to obtain an orthonormal basis \( \{ u_1, \ldots, u_r \} \), where
\[
\begin{align*}
\frac{1}{\|Av_i\|} \quad Av_i &= \frac{1}{\|Av_i\|} \quad v_i = \frac{1}{\|Av_i\|} \quad u_i
\end{align*}
\]
and
\[
Av_i = \sigma_i u_i \quad (1 \leq i \leq r)
\] (4)

Now extend \( \{ u_1, \ldots, u_r \} \) to an orthonormal basis \( \{ u_1, \ldots, u_n \} \) of \( \mathbb{R}^n \), and let
\[
U = [u_1 \quad u_2 \quad \ldots \quad u_r \quad \ldots \quad u_n]
\]

By construction, \( U \) and \( V \) are orthogonal matrices. Also, from (4),
\[
AV = [Av_1 \quad \ldots \quad Av_r \quad 0 \quad \ldots \quad 0] = [\sigma_1 u_1 \quad \ldots \quad \sigma_r u_r \quad 0 \quad \ldots \quad 0]
\]

Let \( D \) be the diagonal matrix with diagonal entries \( \sigma_1, \ldots, \sigma_r \) and let \( \Sigma \) be as in (3) above. Then
\[
U \Sigma = [u_1 \quad u_2 \quad \ldots \quad u_n] \begin{bmatrix}
\sigma_1 & 0 & \cdots & 0 \\
0 & \sigma_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_r
\end{bmatrix}
= [\sigma_1 u_1 \quad \sigma_2 u_2 \quad \ldots \quad \sigma_r u_r \quad 0 \quad \ldots \quad 0]
= AV
\]

Since \( V \) is an orthogonal matrix, \( U \Sigma V^T = AV \).

The next two examples focus attention on the internal structure of a singular value decomposition. An efficient and numerically stable algorithm for this decomposition would use a different approach. See the Numerical Example at the end of the section.

EXAMPLE 3 Use the results of Examples 1 and 2 to construct a singular value decomposition of \( A = \begin{bmatrix} 4 & 11 \\ 8 & 7 \\ 4 \end{bmatrix} \).

SOLUTION A construction can be divided into three steps.

Step 1. Find an orthogonal diagonalization of \( A^T A \). That is, find the eigenvalues of \( A^T A \) and a corresponding orthonormal set of eigenvectors. If \( A \) had only two columns, the calculations could be done by hand. Larger matrices usually require a matrix calculator software. However, for the matrix \( A \) here, the eigenvalues for \( A^T A \) are provided in Example 1.

Step 2. Set up \( V \) and \( \Sigma \). Arrange the eigenvalues of \( A^T A \) in decreasing order. In Example 1, the eigenvalues are already listed in decreasing order: 360, 90, and 0. The corresponding unit eigenvectors, \( v_1, v_2, \) and \( v_3 \), are the right singular vectors of \( A \). Using Example 1, construct
\[
V = [v_1 \quad v_2 \quad v_3] = \begin{bmatrix}
\frac{1}{\sqrt{2}} \quad \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} \quad -\frac{1}{\sqrt{2}} \\
0 \quad 0
\end{bmatrix}
\]

The square roots of the eigenvalues are the singular values:
\[
\sigma_1 = \sqrt{360}, \quad \sigma_2 = \sqrt{90}, \quad \sigma_3 = 0
\]
The nonzero singular values are the diagonal entries of \( D \). The matrix \( \Sigma \) is the same size as \( A \), with \( D \) in its upper left corner and with 0's elsewhere.
\[
D = \begin{bmatrix} \sqrt{360} & 0 \\ 0 & \sqrt{90} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Sigma = [D \quad 0] = \begin{bmatrix} \sqrt{360} & 0 \\ 0 & \sqrt{90} \\ 0 & 0 \end{bmatrix}
\]

Step 3. Construct \( U \). When \( A \) has rank \( r \), the first \( r \) columns of \( U \) are the normalized vectors obtained from \( Av_1, \ldots, Av_r \). In this example, \( A \) has two nonzero singular values, so rank \( A = 2 \). Recall from equation (2) and the paragraph before Example 2 that \( \|Av_1\| = \sigma_1 / \sigma_1 = 1 \) and \( \|Av_2\| = \sigma_2 / \sigma_2 = 1 \). Thus
\[
\begin{align*}
\frac{1}{\sigma_1} Av_1 &= \frac{1}{\sqrt{360}} \begin{bmatrix} 18 \\ 6 \\ 0 \end{bmatrix} \\
\frac{1}{\sigma_2} Av_2 &= \frac{1}{\sqrt{90}} \begin{bmatrix} 3 \\ -9 \\ 0 \end{bmatrix}
\end{align*}
\]

Note that \( [v_1, v_2] \) is already a basis for \( \mathbb{R}^2 \). Thus no additional vectors are needed for \( U \), and \( U = [v_1 \quad v_2] \). The singular value decomposition of \( A \) is
\[
A = \begin{bmatrix} 3/\sqrt{10} & 1/\sqrt{10} \\ 1/\sqrt{10} & -3/\sqrt{10} \end{bmatrix} \begin{bmatrix} \sqrt{360} & 0 \\ 0 & \sqrt{90} \end{bmatrix} \begin{bmatrix} 1/3 & 2/3 \\ 2/3 & -1/3 \end{bmatrix}
\]

EXAMPLE 4 Find a singular value decomposition of \( A = \begin{bmatrix} 1 & 0 \\ -2 & 2 \end{bmatrix} \).

SOLUTION First, compute \( A^T A = \begin{bmatrix} 9 & -9 \\ -9 & 9 \end{bmatrix} \). The eigenvalues of \( A^T A \) are 18 and 0, with corresponding unit eigenvectors
\[
\begin{align*}
v_1 &= \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}, \quad v_2 &= \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}
\end{align*}
\]

These unit vectors form the columns of \( V \).
\[
V = [v_1 \quad v_2] = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}
\]

The singular values are \( \sigma_1 = \sqrt{18} = 3\sqrt{2} \) and \( \sigma_2 = 0 \). Since there is only one nonzero singular value, the "matrix" \( D \) may be written as a single number. That is, \( D = 3\sqrt{2} \).

The matrix \( \Sigma \) is the same size as \( A \), with \( D \) in its upper left corner.
\[
\Sigma = \begin{bmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}
\]

To construct \( U \), first construct \( Av_1 \) and \( Av_2 \):
\[
Av_1 = \begin{bmatrix} 2/\sqrt{2} \\ -4/\sqrt{2} \end{bmatrix}, \quad Av_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\]
As a check on the calculations, verify that \( \| A v_1 \| = \sigma_1 = 3 \sqrt{2} \). Of course, \( A v_2 = 0 \) because \( \langle A v_1 \rangle = \sigma_2 = 0 \). The only column found for \( U \) so far is
\[
\begin{bmatrix}
\frac{1}{\sqrt{2}}
\\frac{1}{\sqrt{2}}
\end{bmatrix}
\]

The other columns of \( U \) are found by extending the set \( \{ u_1 \} \) to an orthonormal basis for \( \mathbb{R}^2 \). In this case, we need two orthogonal unit vectors \( u_1 \) and \( u_2 \) that are orthogonal to \( u_1 \). (See Figure 3.) Each vector must satisfy \( u_1^T x = 0 \), which is equivalent to the equations
\[ x_1 - 2 x_2 + 2 x_3 = 0. \]
A basis for the solution set of this equation is
\[
\begin{bmatrix}
2
1
0
\end{bmatrix}, \quad \begin{bmatrix}
2
0
1
\end{bmatrix}
\]

(Check that \( u_1 \) and \( u_2 \) are each orthogonal to \( u_1 \).) Apply the Gram-Schmidt process (with normalizations) to \( \{ w_1, w_2 \} \), and obtain
\[
\begin{bmatrix}
\frac{2}{\sqrt{5}}
0
\end{bmatrix}, \quad \begin{bmatrix}
\frac{1}{\sqrt{5}}
\frac{2}{\sqrt{5}}
\end{bmatrix}
\]

Finally, set \( U = [ u_1 \; u_2 \; u_3 ] \), take \( \Sigma \) and \( V^T \) from above, and write
\[
A = \begin{bmatrix}
1 & -1 \\
-2 & 2 \\
2 & -2
\end{bmatrix} = \begin{bmatrix}
1/2 & \sqrt{2}/2 & 0 \\
-2/3 & 1/2 & 1/2 \\
0 & 4/\sqrt{5} & \sqrt{5}/2
\end{bmatrix} = \begin{bmatrix}
1/\sqrt{2} & -1/\sqrt{2} \\
0 & 1/\sqrt{2} \\
0 & 1/\sqrt{2}
\end{bmatrix}
\]

Applications of the Singular Value Decomposition

The SVD is often used to estimate the rank of a matrix, as noted above. Several other numerical applications are described briefly below, and an application to image processing is presented in Section 7.5.

**EXAMPLE 5** (The Condition Number) Most numerical calculations involving the equation \( Ax = b \) are as reliable as possible when the SVD of \( A \) is used. The orthogonal matrices \( U \) and \( V \) do not affect lengths of vectors or angles between vectors (Theorem 7 in Section 6.2). Any possible instabilities in numerical calculations are identified in \( \Sigma \). If the singular values of \( A \) are extremely large or very small, roundoff errors are almost inevitable, but an error analysis is aided by knowing the entries in \( \Sigma \) and \( V \).

If \( A \) is an invertible \( n \times n \) matrix, then the ratio \( \sigma_0/\sigma_1 \) is the largest and smallest singular values gives the condition number of \( A \). Exercises 41–43 in Section 6.3 showed how the condition number affects the sensitivity of a solution of \( Ax = b \) to changes for errors in \( x \) and \( A \). (Actually, a "condition number" of \( A \) can be computed in several ways, but the definition given here is widely used for studying \( Ax = b \).)

**EXAMPLE 6** (Bases for Fundamental Subspaces) Given an SVD for an \( n \times n \) matrix \( A \), let \( u_1, \ldots, u_n \) be the left singular vectors, \( v_1, \ldots, v_n \), the right singular vectors, and \( \sigma_1, \ldots, \sigma_n \) the singular values, and let \( r \) be the rank of \( A \). By Theorem 9,
\[
[u_1, u_2, \ldots, u_r]
\]
is an orthonormal basis for \( \text{Col} A \).

7.4 The Singular Value Decomposition

Recall from Theorem 3 in Section 6.1 that \((\text{Col} \ A)^T = \text{Nul} \ A^T\). Hence
\[
\{ u_{r+1}, \ldots, u_n \}
\]
is an orthonormal basis for \( \text{Nul} \ A^T \).

Since \( \| A v_i \| = \sigma_i \) for \( 1 \leq i \leq n \), and \( \sigma_i = 0 \) if and only if \( i > r \), the vectors \( v_1, \ldots, v_r \) span a subspace of \( \text{Nul} \ A \) of dimension \( n - r \). By the Rank Theorem, \( \dim \text{Nul} \ A = n - \text{rank} \ A \). It follows that
\[
\{ v_1, \ldots, v_r \}
\]
is an orthonormal basis for \( \text{Row} \ A \).

From (5) and (6), the orthogonal complement of \( \text{Nul} \ A^T \) is \( \text{Col} \ A \). Interchanging \( A \) and \( A^T \), note that \((\text{Nul} \ A)^T = \text{Col} \ A^T = \text{Row} \ A \). Hence, from (7),
\[
\{ v_1, \ldots, v_r \}
\]
is an orthonormal basis for \( \text{Row} \ A \).

Figure 4 summarizes (5)–(8), but shows the orthogonal basis \( \{ v_1, \ldots, v_r, u_1, \ldots, u_n \} \) for \( \text{Col} \ A \) instead of the normalized basis, to remind you that \( A v_i = \sigma_i u_i \) for \( 1 \leq i \leq r \).

Explicit orthonormal bases for the four fundamental subspaces determined by \( A \) are useful in some calculations, particularly in constrained optimization problems.

**THEOREM**

The Invertible Matrix Theorem (concluded)

Let \( A \) be an \( n \times n \) matrix. Then the following statements are each equivalent to the statement that \( A \) is an invertible matrix.
1. \((\text{Col} \ A)^T = \{ 0 \}\).
2. \((\text{Nul} \ A)^* = \mathbb{R}^n\).
3. \(\text{Row} \ A = \mathbb{R}^n\).
4. \(A \) has \( n \) nonzero singular values.
EXAMPLE 7 (Reduced SVD and the Pseudoinverse of A). When $\Sigma$ contains rows or columns of zeros, a more compact decomposition of $A$ is possible. Using the notation established above, let $r = \text{rank } A$, and partition $U$ and $V$ into submatrices whose first blocks contain $r$ columns:

\[
U = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix}, \quad V = \begin{bmatrix} V_r & V_{n-r} \end{bmatrix},
\]

where $U_r = \{u_1, \ldots, u_r\}$ and $V_r = \{v_1, \ldots, v_r\}$.

Then $U_r$ is $m \times r$ and $V_r$ is $n \times r$. (To simplify notation, we consider $U_{m-r}$ or $V_{n-r}$, even though one of them may have no columns.) Then partitioned matrix multiplication shows that

\[
A = U_r U_r^T + U_{m-r} U_{m-r}^T = U_r V_r^T + U_{m-r} V_{m-r}^T.
\]

This factorization of $A$ is called a reduced singular value decomposition of $A$. Since the diagonal entries in $\Sigma$ are nonzero, $D$ is invertible. The following matrix is called the pseudoinverse (also, the Moore-Penrose inverse) of $A$:

\[
A^+ = V_r D_r^{-1} U_r^T
\]

Supplementary Exercises 12–14 at the end of the chapter explore some of the properties of the reduced singular value decomposition and the pseudoinverse.

EXAMPLE 8 (Least-Squares Solution). Given the equation $Ax = b$, use the pseudoinverse of $A$ in (10) to define

\[
s = A^+ b = V_r D_r^{-1} U_r^T b
\]

Then, from the SVD in (9),

\[
\begin{align*}
A & = (U_r V_r^T) (D_r 0) (V_r^T U_r b) \\
& = U_r D_r^{-1} U_r^T b \\
& = U_r U_r^T b
\end{align*}
\]

Because $V_r^T V_r = I_r$.

It follows from (5) that $U_r U_r^T b$ is the orthogonal projection of $b$ onto Col $A$. (See Theorem 10 in Section 6.3.) Thus $s$ is a least-squares solution of $Ax = b$. In fact, $s$ has the smallest length among all least-squares solutions of $Ax = b$. See Supplementary Exercise 14.

**NUMERICAL NOTE**

Examples 1–4 and the exercises illustrate the concept of singular values and suggest how to perform calculations by hand. In practice, the computation of $A^+$ should be avoided since any errors in the entries of $A$ are squared in the entries of $A^+$. There exist fast iterative methods that produce the singular values and singular vectors of $A$ accurately to many decimal places.

Further Reading


**PRACTICE PROBLEMS**

1. Given a singular value decomposition, $A = U \Sigma V^T$, find an SVD of $A^T$. How are the singular values of $A$ and $A^T$ related?

2. For any $n \times n$ matrix $A$, use the SVD to show that there is an $n \times n$ orthogonal matrix $Q$ such that $A^T A = Q^T (A^T A) Q$.

Remark: Practice Problem 2 establishes that for any $n \times n$ matrix $A$, the matrices $AA^T$ and $A^T A$ are orthogonally similar.

**EXERCISES**

Find the singular values of the matrices in Exercises 1–4.

1. \[
\begin{bmatrix}
1 & 0 \\
0 & -3
\end{bmatrix}
\]

2. \[
\begin{bmatrix}
-3 & 0 \\
0 & 8
\end{bmatrix}
\]

3. \[
\begin{bmatrix}
1 & -1 \\
2 & 1
\end{bmatrix}
\]

4. \[
\begin{bmatrix}
1 & 2 \\
3 & 4
\end{bmatrix}
\]

Find an SVD of each matrix in Exercises 5–12. (Hint: In Exercise 11, one choice for $U$ is $\begin{bmatrix} 2/3 & -1/3 \end{bmatrix}$, and in Exercise 12, one column of $U$ can be $\begin{bmatrix} 1/\sqrt{2} \\
1/\sqrt{2}\end{bmatrix}$.)

5. \[
\begin{bmatrix}
-2 & 0 \\
0 & 0
\end{bmatrix}
\]

6. \[
\begin{bmatrix}
-3 & 0 \\
0 & -2
\end{bmatrix}
\]

7. \[
\begin{bmatrix}
-2 & 1 \\
2 & 2
\end{bmatrix}
\]

8. \[
\begin{bmatrix}
4 & 0 \\
0 & 4
\end{bmatrix}
\]

9. \[
\begin{bmatrix}
0 & -1 \\
1 & 0
\end{bmatrix}
\]

10. \[
\begin{bmatrix}
7 & 1 \\
0 & 2
\end{bmatrix}
\]

11. \[
\begin{bmatrix}
0 & -1 \\
4 & 2
\end{bmatrix}
\]

12. \[
\begin{bmatrix}
1 & 1 \\
0 & -1
\end{bmatrix}
\]

13. Find the SVD of $A = \begin{bmatrix} 3 & 2 \\
2 & 3 \end{bmatrix}$. (Hint: Work with $A^T$.)

14. In Exercises 7, find a unit vector $x$ at which $Ax$ has maximum length.

15. Suppose $A$ is square and invertible. Find a singular value decomposition of $A^{-1}$.

16. Show that if $A$ is square, then $|det A|$ is the product of the singular values of $A$.

17. Suppose $A$ is square and invertible. Find a singular value decomposition of $A^{-1}$.

18. Show that the columns of $V$ are eigenvectors of $A^T A$, the columns of $U$ are eigenvectors of $A A^T$, and the diagonal
SOLUTIONS TO PRACTICE PROBLEMS

1. If $A = U \Sigma V^T$, where $\Sigma$ is $m \times n$, then $A^T = (V^T \Sigma^T)^T = V \Sigma^T U^T$. This is the SVD of $A^T$ because $V$ and $U$ are orthogonal matrices and $\Sigma^T$ is an $n \times m$ "diagonal" matrix. Since $\Sigma$ and $\Sigma^T$ have the same nonzero diagonal entries, $A^T$ has the same nonzero singular values. [Note: If $A$ is $2 \times n$, then $A^T A$ is only $2 \times 2$ and its eigenvalues may be easier to compute (by hand) than the eigenvalues of $A A^T$.]

2. Use the SVD to write $A = U \Sigma V^T$, where $U$ and $V$ are $n \times n$ orthogonal matrices and $\Sigma$ is an $n \times n$ diagonal matrix. Notice that $U^T U = I = V^T V$ and $\Sigma^2 = \Sigma$, since $U$ and $V$ are orthogonal matrices and $\Sigma$ is a diagonal matrix. Substituting the SVD for $A$ into $A A^T = C^2$ and $A^T A$ results in

$$AA^T = U \Sigma V^T U = \Sigma V^T U = \Sigma V^T U = U \Sigma V^T,$$

and

$$A^T A = (U V^T)^T U V^T = V \Sigma^T U = V \Sigma^T U = V \Sigma V^T.$$

Let $Q = V U^T$. Then

$$Q^T (A A^T) Q = (V U^T)^T (U \Sigma V^T) (U V^T) = U \Sigma V^T.$$

7.5 APPLICATIONS TO IMAGE PROCESSING AND STATISTICS

The satellite photographs in this chapter’s introduction provide an example of multivariate, or multivariate, data—information organized so that each datum in the data set is identified with a point (vector) in $R^4$. The main goal of this section is to explain a technique, called principal component analysis, used to analyze such multivariate data. The calculations will illustrate the use of orthogonal diagonalization and the singular value decomposition.

Principal component analysis can be applied to any data that consist of lists of measurements made on a collection of objects or individuals. For instance, consider a chemical process that produces a plastic material. To monitor the process, 200 samples are taken of the material produced, and each sample is subjected to a battery of eight tests, such as melting point, density, ductility, tensile strength, and so on. The laboratory report for each sample is a vector in $R^8$, and the set of such vectors forms an $8 \times 300$ matrix, called the matrix of observations.

Loosely speaking, we can say that the process control data are eight-dimensional. The next two examples describe data that can be visualized graphically.

EXAMPLE 1

An example of two-dimensional data is given by the weights and heights of $N$ college students. Let $X_i$ denote the observation vector in $R^2$ that lists the weight and height of the $i$th student. If $W$ denotes weight and $H$ height, then the matrix of observations has the form

$$\begin{bmatrix}
w_1 & w_2 & \cdots & w_N \\
H_1 & H_2 & \cdots & H_N \\
\end{bmatrix}$$

The set of observation vectors can be visualized as a two-dimensional scatter plot. See Figure 1.

EXAMPLE 2

The first three photographs of Railroad Valley, Nevada, shown in the chapter introduction can be viewed as one image of the region, with three wavebands, because simultaneous measurements of the region were made at three separate wavelengths. Each photograph gives different information about the same physical region. For instance, the pixel corresponding to the upper-left corner of each photograph corresponds to the same place on the ground (about 30 meters by 30 meters). To each pixel there corresponds an observation vector in $R^3$ that lists the signal intensities for that pixel in the three spectral bands.

Typically, the image is $2000 \times 2000$ pixels, so there are 4 million pixels in the image. The data for the image form a matrix with 3 rows and 4 million columns (with columns arranged in any convenient order). In this case, the "multidimensional" character of the data refers to the three spectral dimensions rather than the two spatial dimensions that naturally belong to any photograph. The data can be visualized as a cluster of 4 million points in $R^3$, perhaps as in Figure 2.

Mean and Covariance

To prepare for principal component analysis, let $[X_1 \ldots X_N]$ be a $p \times N$ matrix of observations, such as described above. The sample mean, $\bar{M}$, of the observation vectors...
### Example 3

Three measurements are made on each of four individuals in a random sample from a population. The observation vectors are

\[
X_i = \begin{bmatrix} 2 \\ 1 \\ 4 \\ 7 \\ 8 \\ 6 \\ 7 \\ 2 \end{bmatrix}
\]

Compute the sample mean and the covariance matrix.

**Solution**

The sample mean is

\[
M = \frac{1}{4} \begin{bmatrix} 1 \\ 2 \\ 4 \\ 7 \\ 8 \\ 5 \\ 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 20 \\ 16 \\ 4 \\ 5 \\ 5 \end{bmatrix}
\]

Subtract the sample mean from \(X_1, \ldots, X_4\) to obtain

\[
\hat{X}_i = \begin{bmatrix} -2 \\ -1 \\ 2 \\ 4 \\ 3 \\ 0 \\ 3 \\ 4 \end{bmatrix}
\]

and

\[
B = \begin{bmatrix} -2 & -1 & 2 & 3 \\ -4 & 2 & 4 & 0 \\ -4 & 8 & -4 & 0 \end{bmatrix}
\]

The sample covariance matrix is

\[
S = \frac{1}{3} \begin{bmatrix} -4 & -2 & -2 & -4 \\ -2 & -2 & 4 & 0 \\ -4 & 8 & -4 & 0 \\ 10 & 6 & 6 & 30 \\ 2 & 4 & 4 & 18 \\ -2 & 8 & 8 & 24 \\ -2 & 4 & 4 & 24 \\ 0 & -8 & 8 & 96 \end{bmatrix}
\]

To discuss the entries in \(S = [s_{ij}]\), let \(X\) represent a vector that varies over the set of observation vectors and denote the coordinates of \(X\) by \(x_1, \ldots, x_4\). Then \(x_1\), for example, is a scalar that varies over the set of first coordinates of \(X_1, \ldots, X_4\). Thus, if \(P \in \mathbb{R}^{p \times p}\) is any diagonal matrix with \(s_{ij}\) on the diagonal, \(S\) is called the **variance** of \(x_j\).

The variance of \(x_j\) measures the spread of the values of \(x_j\). (See Exercise 13.) In Example 3, the variance of \(x_1\) is 10 and the variance of \(x_3\) is 32. The fact that 32 is more than 10 indicates that the set of third entries in the response vectors contains a wider spread of values than the set of first entries.

The **total variance** of the data is the sum of the variances on the diagonal of \(S\). In general, the sum of the diagonal entries of a square matrix \(S\) is called the trace of the matrix, written \(tr(S)\). Thus

\[
\text{total variance} = tr(S)
\]

The entry \(s_{ij}\) in \(S\) for \(i \neq j\) is called the **covariance** of \(x_i\) and \(x_j\). Observe that in Example 3, the covariance between \(x_1\) and \(x_4\) is 0 because the (1, 4)-entry in \(S\) is 0. Statisticians say that \(x_1\) and \(x_4\) are **uncorrelated**. Analysis of the multivariate data in \(X_1, \ldots, X_4\) is greatly simplified when most or all of the variables \(x_1, \ldots, x_4\) are uncorrelated, that is, when the covariance matrix of \(X_1, \ldots, X_4\) is diagonal or nearly diagonal.

### Principal Component Analysis

For simplicity, assume that the matrix \([X_1 \quad \cdots \quad X_4]\) is already in mean-deviation form. The goal of principal component analysis is to find an orthogonal \(p \times p\) matrix \(P = [u_1 \quad \cdots \quad u_p]\) that determines a change of variable, \(X = PX\), or

\[
\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}
\]

with the property that the new variables \(y_1, \ldots, y_p\) are uncorrelated and are arranged in order of decreasing variance.

The orthogonal change of variable \(X = PX\) means that each observation vector \(X_k\), for \(k = 1, \ldots, N\), is orthogonal to all other vectors \(X_i\), for \(i \neq k\).

It is then possible to verify that for any orthogonal \(P\), the covariance matrix of \(Y_1, \ldots, Y_p\) is \(P^TSP\) (Exercise 11). So the desired orthogonal matrix \(P\) is one that makes \(P^TSP\) diagonal. Let \(D\) be a diagonal matrix with the eigenvalues \(\lambda_1, \lambda_2, \ldots, \lambda_p\) of \(S\) on the diagonal, arranged so that \(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0\), and let \(P\) be an orthogonal matrix whose columns are the corresponding unit eigenvectors \(u_1, \ldots, u_p\). Then \(S = PDP^T\) and \(P^TSP = D\).

The unit eigenvectors \(u_1, \ldots, u_p\) of the covariance matrix \(S\) are called the **principal components** of the data (in the matrix of observations). The first principal component is the eigenvector corresponding to the largest eigenvalue of \(S\), the second principal component is the eigenvector corresponding to the second largest eigenvalue, and so on.

The first principal component \(u_1\) determines the new variable \(y_1\) in the following way: Let \(c_1, \ldots, c_p\) be the entries in \(u_1\). Since \(u_1^T u_1 = 1\) is the first row of \(P^T\), the equation

\[
y_1 = u_1^T X = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p
\]

is a linear combination of the original variables \(x_1, \ldots, x_p\), using the entries in the eigenvector \(u_1\) as weights. In a similar fashion, \(u_i\) determines the variable \(y_i\), and so on.
EXAMPLE 4 The initial data for the multispectral image of Railroad Valley (Example 2) consisted of 4 million vectors in \( R^7 \). The associated covariance matrix is:

\[
S = \begin{bmatrix}
2382.78 & 2611.84 & 2136.20 \\
2611.84 & 3106.47 & 2553.90 \\
2136.20 & 2553.90 & 2650.71
\end{bmatrix}
\]

Find the principal components of the data, and list the new variable determined by the first principal component.

**SOLUTION** The eigenvalues of \( S \) and the associated principal components (the unit eigenvectors) are:

\[
\lambda_1 = 7614.23 \quad \lambda_2 = 427.63 \quad \lambda_3 = 98.10 \\
\mathbf{u}_1 = \begin{bmatrix} 0.5417 \\ -0.4994 \\ 0.6834 \end{bmatrix} \quad \mathbf{u}_2 = \begin{bmatrix} -0.3026 \\ 0.8179 \\ -0.1441 \end{bmatrix}
\]

Using two decimal places for simplicity, the variable for the first principal component is:

\[
y_1 = 0.54x_1 + 0.53x_2 + 0.56x_3
\]

This equation was used to create photograph (d) in the chapter introduction. The variables \( x_1, x_2, \) and \( x_3 \) are the signal intensities in the three spectral bands. The values of \( x_3 \), converted to a gray scale between black and white, produced photograph (a).

Similarly, the values of \( x_2 \) and \( x_3 \) produced photographs (b) and (c), respectively. At each pixel in photograph (d), the gray scale value is computed from \( y_1 \), a weighted linear combination of \( x_1, x_2, \) and \( x_3 \). In this sense, photograph (d) "displays" the first principal component of the data.

In Example 4, the covariance matrix for the transformed data, using variables \( y_1, y_2, \) and \( y_3 \), is:

\[
D = \begin{bmatrix}
7614.23 & 0 & 0 \\
0 & 427.63 & 0 \\
0 & 0 & 98.10
\end{bmatrix}
\]

Although \( D \) is obviously simpler than the original covariance matrix \( S \), the merit of constructing the new variables is not yet apparent. However, the variances of the variables \( y_1, y_2, \) and \( y_3 \) appear on the diagonal of \( D \), and obviously the first variance in \( D \) is much larger than the other two. As we shall see, this fact will permit us to view the data as essentially one-dimensional rather than three-dimensional.

Reducing the Dimension of Multivariate Data

Principal component analysis is potentially valuable for applications in which most of the variation, or dynamic range, in the data is due to variations in only a few of the new variables, \( y_1, \ldots, y_p \).

It can be shown that an orthogonal change of variables, \( X = PY \), does not change the total variance of the data. (Roughly speaking, this is true because left-multiplication by \( P \) does not change the lengths of vectors or the angles between them. See Exercise 12.) This means that if \( S = PP^T \), then

\[
\begin{bmatrix}
\text{total variance} \\
\text{of } y_1, \ldots, y_p
\end{bmatrix} = \begin{bmatrix}
\text{total variance} \\
\text{of } x_1, \ldots, x_n
\end{bmatrix} = \text{tr}(S) = \lambda_1 + \cdots + \lambda_p
\]

The variance of \( y_j \) is \( \lambda_j \), and the quotient \( \lambda_j / \text{tr}(S) \) measures the fraction of the total variance that is "explained" or "captured" by \( y_j \).

---

EXAMPLE 5 Compute the various percentages of variance of the Railroad Valley multispectral data that are displayed in the principal component photographs, (d)-(f), shown in the chapter introduction.

**SOLUTION** The total variance of the data is:

\[
\text{tr}(D) = 7614.23 + 427.63 + 98.10 = 8139.96
\]

[Verify that this number also equals tr(S).] The percentages of the total variance explained by the principal components are:

\[
\begin{array}{ccc}
\text{First component} & \text{Second component} & \text{Third component} \\
7614.23 & 8139.96 & 3 \times 427.63 & 5 \times 98.10 & 1 \times 98.10 & 1.2\% \\
93.5\% & 5.3\% & 0.3\% & 0.3\% & 0.3\% & 0.3\%
\end{array}
\]

In a sense, 93.5% of the information collected by Landsat for the Railroad Valley region is displayed in photograph (d), with 5.3% in (e) and only 1.2% remaining for (f).

The calculations in Example 5 show that the data have practically no variance in the third (new) coordinate. The values of \( y_3 \) are all close to zero. Geometrically, the data points lie nearly in the plane \( y_3 = 0 \), and their locations can be determined fairly accurately by knowing only the values of \( y_1 \) and \( y_2 \). In fact, \( y_2 \) also has relatively small variance, which means that the points lie approximately along a line, and the data are essentially one-dimensional. See Figure 2, in which the data resemble a popsicle stick.

Characterizations of Principal Component Variables

If \( y_1, \ldots, y_p \) arise from a principal component analysis of a \( p \times N \) matrix of observations, then the variance of \( y_j \) is as large as possible in the following sense:

If \( u \) is any unit vector and if \( y = uX \), then the variance of the values of \( y \) as \( X \) varies over the original data \( X_1, \ldots, X_N \) turns out to be \( u^T S u \). By Theorem 8 in Section 7.3, the maximum value of \( u^T S u \) over all unit vectors \( u \) is the largest eigenvalue \( \lambda_1 \) of \( S \), and this variance is attained when \( u \) is the corresponding eigenvector \( u_1 \). In the same way, Theorem 8 shows that \( y_2 \) has maximum possible variance among all variables \( y = u^T X \) that are uncorrelated with \( y_1 \). Likewise, \( y_3 \) has maximum possible variance among all variables uncorrelated with both \( y_1 \) and \( y_2 \), and so on.

**NUMERICAL NOTE**

The singular value decomposition is the main tool for performing principal component analysis in practical applications. If \( B \) is a \( p \times N \) matrix of observations in mean-deviation form, and if \( A = (1 / \sqrt{N} - 1)B^T \), then \( A^T A \) is the covariance matrix, \( S \). The squares of the singular values of \( B \) are the eigenvalues of \( S \), and the right singular vectors of \( A \) are the principal components of the data.

As mentioned in Section 7.4, iterative calculation of the SVD of \( A \) is faster and more accurate than an eigenvalue decomposition of \( S \). This is particularly true, for instance, in the hyperspectral image processing (with \( p = 224 \)) mentioned in the chapter introduction. Principal component analysis is completed in seconds on specialized workstations.

Further Reading

### Excerpts from the Document

#### 7.5 EXERCISES

1. Find the covariance matrix for the data.
2. Make a principal component analysis of the data to find a single size index that explains most of the variation in the data.

9. Suppose three tests are administered to a random sample of college students. Let $X_1, X_2, X_3$ be observation vectors in $\mathbb{R}^3$ that list the three scores of each student, and for $j = 1, 2, 3$, let $c_j$ denote student $j$'s score on the $j$th exam. Suppose the covariance matrix of the data is

$$ S = \begin{bmatrix} 5 & 2 & 0 \\ 2 & 6 & 2 \\ 0 & 2 & 7 \end{bmatrix} $$

Let $y$ be an "index" of student performance, with $y = c_1x_1 + c_2x_2 + c_3x_3$ and $c_1^2 + c_2^2 + c_3^2 = 1$. Choose $c_1, c_2, c_3$ so that the variance of $y$ over the data set is as large as possible. ([hint: The eigenvalues of the sample covariance matrix are $4, 3.6, 9.1$].)

10. (M) Repeat Exercise 9 with $S = \begin{bmatrix} 5 & 4 & 2 \\ 4 & 11 & 2 \\ 2 & 4 & 5 \end{bmatrix}$

11. Given multivariate data $X_1, \ldots, X_n$ in $\mathbb{R}^p$ as mean-deviation form, let $P$ be a $p \times p$ matrix, and define $Y_k = P^T X_k$ for $k = 1, \ldots, N$. a. Show that $Y_1, \ldots, Y_n$ are in mean-deviation form. ([hint: Let $w$ be the vector in $\mathbb{R}^p$ with $1$ in each entry. Then $[X_1 \ldots X_n] w = 0$ (the zero vector in $\mathbb{R}^n$).])

b. Show that if the covariance matrix of $X_1, \ldots, X_n$ is $S$, then the covariance matrix of $Y_1, \ldots, Y_n$ is $P^T SP$.

12. Let $X$ denote a vector that varies over the columns of a $p \times N$ matrix of observations, and let $P$ be a $p \times p$ orthogonal matrix. Show that the change of the total variance of the data. ([hint: By Exercise 11, it suffices to show that $(P^T SP)^T = (S)$.) Use a property of the trace mentioned in Exercise 25 in Section 5.4.)

13. The sample covariance matrix is a generalization of $\Sigma$ (from Chapter 5) for the variance of a sample of $N$ scalar measurements $x_1, \ldots, x_N$, if $M$ is the average of $\ldots, x_N$, then the sample variance is given by

$$ \frac{1}{N-1} \sum_{k=1}^{N} (x_k - M)^2. $$

---

#### Solutions to Practice Problems

1. First arrange the data in mean-deviation form. The sample mean vector is easily seen to be $\mathbf{M} = \begin{bmatrix} 130 \\ 65 \end{bmatrix}$ Subtracted $\mathbf{M}$ from the observation vectors (the columns in the table) and obtain $\mathbf{B} = \begin{bmatrix} -10 & -5 & -5 & 5 & 15 \\ -4 & -5 & -1 & 3 & 7 \end{bmatrix}$ Then the sample covariance matrix is $S = \begin{bmatrix} 1 & -10 & -5 & 5 & 15 \\ -5 & -4 & -5 & -1 & 3 \\ 5 & -5 & -1 & 3 & 7 \\ 15 & 15 & 15 & 15 & 15 \end{bmatrix} = \begin{bmatrix} 100 & 47.5 \\ 47.5 & 25.0 \end{bmatrix}$

2. The eigenvalues of $S$ are (to two decimal places) $\lambda_1 = 123.02$ and $\lambda_2 = 1.98$.

The unit eigenvector corresponding to $\lambda_1$ is $u = \begin{bmatrix} 0.90 \\ 0.43 \end{bmatrix}$ (Since $S$ is $2 \times 2$, the computations can be done by hand if a matrix program is not available.) For the size index, let $y = 900u^T + 436\hat{w}$, where $\hat{w}$ and $\hat{u}$ are weight and height, respectively, in mean-deviation form. The variance of this index over the data set is 123.02. Because the total variance is $tr(S) = 100 + 25 = 125$, the size index accounts for practically all (98.4%) of the variance of the data.

The original data for Practice Problem 1 and the line determined by the first principal component $u$ are shown in Figure 4. (In parametric vector form, the line is $x = \mathbf{M} + tu$.) It can be shown that the line is the best approximation to the data,