6.2 Zero and identity matrices

Zero matrix. A zero matrix is a matrix with all elements equal to zero. The zero matrix of size $m \times n$ is sometimes written as $0_{m\times n}$, but usually a zero matrix is denoted just 0, the same symbol used to denote the number 0 or zero vectors. In this case the size of the zero matrix must be determined from the context.

Identity matrix. An identity matrix is another common matrix. It is always square. Its *diagonal* elements, *i.e.*, those with equal row and column indices, are all equal to one, and its off-diagonal elements (those with unequal row and column indices) are zero. Identity matrices are denoted by the letter I. Formally, the identity matrix of size n is defined by

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j, \end{cases}$$

for $i, j = 1, \ldots, n$. For example,

$$\left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right], \qquad \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

are the 2×2 and 4×4 identity matrices.

The column vectors of the $n \times n$ identity matrix are the unit vectors of size n. Using block matrix notation, we can write

$$I = \left[\begin{array}{cccc} e_1 & e_2 & \cdots & e_n \end{array}\right],$$

where e_k is the kth unit vector of size n.

Sometimes a subscript is used to denote the size of an identity matrix, as in I_4 or $I_{2\times 2}$. But more often the size is omitted and follows from the context. For example, if

$$A = \left[\begin{array}{ccc} 1 & 2 & 3 \\ 4 & 5 & 6 \end{array}\right],$$

then

$$\left[\begin{array}{cc} I & A \\ 0 & I \end{array}\right] = \left[\begin{array}{ccccc} 1 & 0 & 1 & 2 & 3 \\ 0 & 1 & 4 & 5 & 6 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

The dimensions of the two identity matrices follow from the size of A. The identity matrix in the 1,1 position must be 2×2, and the identity matrix in the 2,2 position must be 3×3. This also determines the size of the zero matrix in the 2,1 position.

The importance of the identity matrix will become clear later, in §10.1.

Sparse matrices. A matrix A is said to be *sparse* if many of its entries are zero, or (put another way) just a few of its entries are nonzero. Its *sparsity pattern* is the set of indices (i, j) for which $A_{ij} \neq 0$. The *number of nonzeros* of a sparse matrix A is the number of entries in its sparsity pattern, and denoted $\mathbf{nnz}(A)$. If A is $m \times n$ we have $\mathbf{nnz}(A) \leq mn$. Its *density* is $\mathbf{nnz}(A)/(mn)$, which is no more than one. Densities of sparse matrices that arise in applications are typically small or very small, as in 10^{-2} or 10^{-4}. There is no precise definition of how small the density must be for a matrix to qualify as sparse. A famous definition of sparse matrix due to the mathematician James H. Wilkinson is: A matrix is sparse if it has enough zero entries that it pays to take advantage of them. Sparse matrices can be stored and manipulated efficiently on a computer.

Many common matrices are sparse. An $n \times n$ identity matrix is sparse, since it has only n nonzeros, so its density is $1/n$. The zero matrix is the sparsest possible matrix, since it has no nonzero entries. Several special sparsity patterns have names; we describe some important ones below.

Like sparse vectors, sparse matrices arise in many applications. A typical customer purchase history matrix (see page 111) is sparse, since each customer has likely only purchased a small fraction of all the products available.

Diagonal matrices. A square $n \times n$ matrix A is *diagonal* if $A_{ij} = 0$ for $i \neq j$. (The entries of a matrix with $i = j$ are called the *diagonal entries*; those with $i \neq j$ are its *off-diagonal* entries.) A diagonal matrix is one for which all off-diagonal entries are zero. Examples of diagonal matrices we have already seen are square zero matrices and identity matrices. Other examples are

$$\begin{bmatrix} -3 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 1.2 \end{bmatrix}.$$

(Note that in the first example, one of the diagonal elements is also zero.)

The notation $\mathbf{diag}(a_1, \ldots, a_n)$ is used to compactly describe the $n \times n$ diagonal matrix A with diagonal entries $A_{11} = a_1, \ldots, A_{nn} = a_n$. This notation is not yet standard, but is coming into more prevalent use. As examples, the matrices above would be expressed as

$$\mathbf{diag}(-3, 0), \qquad \mathbf{diag}(0.2, -3, 1.2),$$

respectively. We also allow $\mathbf{diag}$ to take one n-vector argument, as in $I = \mathbf{diag}(\mathbf{1})$.

Triangular matrices. A square $n \times n$ matrix A is *upper triangular* if $A_{ij} = 0$ for $i > j$, and it is *lower triangular* if $A_{ij} = 0$ for $i < j$. (So a diagonal matrix is one that is both lower and upper triangular.) If a matrix is either lower or upper triangular, it is called *triangular*. For example, the matrices

$$\begin{bmatrix} 1 & -1 & 0.7 \\ 0 & 1.2 & -1.1 \\ 0 & 0 & 3.2 \end{bmatrix}, \qquad \begin{bmatrix} -0.6 & 0 \\ -0.3 & 3.5 \end{bmatrix},$$

are upper and lower triangular, respectively.

A triangular $n \times n$ matrix A has up to $n(n+1)/2$ nonzero entries, *i.e.*, around half its entries are zero. Triangular matrices are generally not considered sparse matrices, since their density is around 50%, but their special sparsity pattern will be important in the sequel.

6.3 Transpose, addition, and norm

6.3.1 Matrix transpose

If A is an $m \times n$ matrix, its *transpose*, denoted A^T (or sometimes A' or A^*), is the $n \times m$ matrix given by $(A^T)_{ij} = A_{ji}$. In words, the rows and columns of A are transposed in A^T. For example,

$$\begin{bmatrix} 0 & 4 \\ 7 & 0 \\ 3 & 1 \end{bmatrix}^T = \begin{bmatrix} 0 & 7 & 3 \\ 4 & 0 & 1 \end{bmatrix}.$$

If we transpose a matrix twice, we get back the original matrix: $(A^T)^T = A$. (The superscript T in the transpose is the same one used to denote the inner product of two n-vectors; we will soon see how they are related.)

Row and column vectors. Transposition converts row vectors into column vectors and vice versa. It is sometimes convenient to express a row vector as a^T, where a is a column vector. For example, we might refer to the m rows of an $m \times n$ matrix A as $\tilde{a}_i^T, \ldots, \tilde{a}_m^T$, where $\tilde{a}_1, \ldots, \tilde{a}_m$ are (column) n-vectors. As an example, the second row of the matrix

$$\begin{bmatrix} 0 & 7 & 3 \\ 4 & 0 & 1 \end{bmatrix}$$

can be written as (the row vector) $(4, 0, 1)^T$.

It is common to extend concepts from (column) vectors to row vectors, by applying the concept to the transposed row vectors. We say that a collection of row vectors is linearly dependent (or independent) if their transposes (which are column vectors) are linearly dependent (or independent). For example, 'the rows of a matrix A are linearly independent' means that the columns of A^T are linearly independent. As another example, 'the rows of a matrix A are orthonormal' means that their transposes, the columns of A^T, are orthonormal. 'Clustering the rows of a matrix X' means clustering the columns of X^T.

Transpose of block matrix. The transpose of a block matrix has the simple form (shown here for a 2×2 block matrix)

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^T = \begin{bmatrix} A^T & C^T \\ B^T & D^T \end{bmatrix},$$

where A, B, C, and D are matrices with compatible sizes. The transpose of a block matrix is the transposed block matrix, with each element transposed.