Lecture Notes for the Linear Algebra for Statisticians (Math 530/488A)

Vladislav Kargin

December 26, 2022

Contents

Chapter 1

Matrix operations and row echelon reduction

Reading for this Lecture

1.1 Matrix products

Let x be an n-dimensional column vector with entries x_i, $i = 1, \ldots, n$ and A be an $m \times n$ matrix with entries a_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$. Then the *matrix-vector product* $b = Ax$ is the m-dimensional vector with entries

$$b_i = \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1, \ldots, m. \tag{1.1}$$

Sometimes it is convenient to write explicitly the dimensions of matrices in a subscript and then we write

$$b_{m\times 1} = A_{m\times n} x_{n\times 1}.$$

This formula can be interpreted in several ways:

1. One can think about x and b as elements of vector spaces $V = R^n$ and $W = \mathbb{R}^m$ written in specific coordinate bases. Then the formula $b = Ax$ shows how a *linear transformation A acts on vector x*. That is, the formula (1.1) explains how to calculate the coordinates of the image $b = (b_1, \ldots, b_m)$ in a basis of $W = \mathbb{R}^m$ from the coordinates of vector $x = (x_1, \ldots, x_n)$ in a basis of $V = \mathbb{R}^n$. In particular if $\boldsymbol{e}_j$ is the j-th basis vector of $\mathbb{R}^n$ so it has coordinates $(0, 0, \ldots, 1 \ldots, 0)$ with 1 in

the j-th place, than the calculation gives vector $A\boldsymbol{e}_j = (a_{1j}, \ldots, a_{mj})$. So, in this interpretation, the number a_{ij} is the i-th coordinate of the image $A\boldsymbol{e}_j$ in the given basis of W.

2. In a different interpretation of formula (1.1) matrix A is seen a collection of n column vectors a_i, each of dimension m. Then, the formula explains how to calculate the *linear combination of these vectors with coefficients provided by vector* x. It can be re-written as

$$\boldsymbol{b} = \sum_{i=1}^{n} x_i \boldsymbol{a}_i, \tag{1.2}$$

where we should remember that each x_i is a number and each a_i is an m-dimensional column vector with entries $(a_{1i}, \ldots, a_{mi})$. In short, in this interpretation $\boldsymbol{x}$ acts on columns of A to produce $\boldsymbol{b}$.

Example 1.1.1. Suppose our vector space V is the space of polynomials with real coefficients modulo x^{n+1}. The usual basis consists of polynomials $1, x, x^2, \ldots, x^n$.

1. What is the matrix of the differentiation operator $(D : P(x) \to P'(x))$in this basis?
2. What is the matrix of the integration operator $S : P(x) \to \int_0^x P(t)\,dt)$?
3. What is the matrix of the shift: $T : P(x) \to P(x+1)$?
4. Represent a sum of 3 polynomials, $P_i(x) \in V$, $i = 1, 2, 3$ by using matrix multiplication.

Similarly, if A and C are two matrices, A is $l \times m$ and C is $m \times n$, then we can define the *matrix-matrix product* $B = AC$,

$$B_{l\times n} = A_{l\times m} C_{m\times n},$$

with entries defined by

$$b_{ij} = \sum_{k=1}^{m} a_{ik} c_{kj}. \tag{1.3}$$

This product can also be interpreted in several ways:

1. If A and C are matrices representing linear transformations $\mathbb{R}^n \to \mathbb{R}^m$ and $\mathbb{R}^m \to \mathbb{R}^l$, respectively, then $B = AC$ is a matrix that represents a composition of these linear transformations $\mathbb{R}^n$ to $\mathbb{R}^l$. (As usual for compositions of maps, composition is from right to left: C acts first, and A second.)

2. Alternatively, we can think about B as matrix such that each column of B is a linear combination of columns of A. The first column of B is the linear combination of column vectors $\boldsymbol{a}_i$ of A with the coefficients equal to components of vector $\boldsymbol{c}_1$, which is the first column of C; the second column of B is the linear combination of $\boldsymbol{a}_i$ with the coefficients in $\boldsymbol{c}_2$ and so on. As a result we get n linear combinations and each of them is a column vector in $\mathbb{R}^l$.

Example 1.1.2. What is the matrix of the composition of the integration and differentiation operators from the previous example? That is, what is the matrix of $S \circ D$? What about $D \circ S$?

1.2 Transposition and conjugate transposition

A *transposition* of an $m \times n$ matrix A is the $n \times m$ matrix A^t for which the entry $(A^t)_{ij}$ equals the entry A_{ji} of the original matrix.[1]

In the situation when matrix A has complex-valued entries, it is typically more useful to define a *conjugate transpose* matrix A^*, with the entry $(A^t)_{ij}$ equal to $\overline{A_{ji}}$, where the line over the number denotes complex conjugation (that is, $\overline{x+iy} = x - iy$). [2] For real matrices $A^* = A^t$, so we will sometime use notation A^* for both real and complex matrices.

For various problem, an especially important class of real-valued matrices is that of symmetric matrices, which satisfy the condition $A = A^t$. For complex valued matrices, the equivalent concept is that of *Hermitian* matrices: $A = A^*$.

Trefethen and Bau book uses the name *adjoint* matrix for A^* and *self-adjoint* for Hermitian matrices by borrowing the terminology from the theory of linear operators.

A useful property of transposition is that $(AB)^* = B^*A^*$.

Indeed, $(AB)_{ij} = \sum_k A_{ik}B_{kj}$ for every i and j. So,

$$((AB)^*)_{ji} = \overline{(AB)_{ij}} = \sum_k \overline{A}_{ik}\overline{B}_{kj} = \sum_k (B^*)_{jk}(A^*)_{ki} = (B^*A^*)_{ji},$$

for every i and j, which implies the required identity. (We used a property of the complex conjugation that says that $\overline{zw} = \overline{z}\,\overline{w}$.)

[1] The notation A^T is also common.

[2] Sometimes, the congujate transpose matrix is denoted A^H.

1.3 Inner and outer products of vectors

The *inner product* (also called *dot product* or *scalar product*) of two m-column vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the matrix product of $\boldsymbol{x}^*$ and $\boldsymbol{y}$. It is a number (or 1×1 matrix)

$$\boldsymbol{x}^*\boldsymbol{y} = \sum_{i=1}^{m} \overline{x_i} y_i.$$

This product is often denoted $(\boldsymbol{x}, \boldsymbol{y})$ or $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$.

The *outer product* of column vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^m$ is the $m \times m$ matrix $\boldsymbol{x}\boldsymbol{y}^*$. The element in i-th row and j-th column is simply a product of i-th component of $\boldsymbol{x}$ and j-th component of $\overline{\boldsymbol{y}}$:

$$(\boldsymbol{x}\boldsymbol{y}^*)_{ij} = x_i \overline{y}_j.$$

(For real matrices, one can ignore complex conjugation in these formulas.) e

Remark: In this definition of the scalar product, the standard basis has the property $(\boldsymbol{e}_i, \boldsymbol{e}_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta: it is 1 if $i = j$ and 0 if $i \neq j$. Often, the scalar product is defined on a vector space V as a map $V \times V \to \mathbb{R}$ (in the real-valued case) that satisfy several axioms: $(c_1 v_1 + c_2 v_2, w) = c_1(v_1, w) + c_2(v_2, w)$, $(v, w) = (w, v)$. Then, it can be proved that one can find a basis in which $(\boldsymbol{e}_i, \boldsymbol{e}_j) = \delta_{ij}$ holds and then one can use the definition given above.

Example 1.3.1. At the space of all polynomials one can define the following scalar products:

1.
$$(P, Q) = \int_{-1}^{1} P(x)Q(x)\,dx,$$

2.
$$(P, Q) = \int_{-\infty}^{\infty} P(x)Q(x)e^{-x^2/2}\,dx.$$

These are valid scalar products, however, the standard basis does not have the required property $(x^k, x^l) \neq \delta_{k,l}$. We will see that one can find polynomials $P_k(x)$ that have the property $\big(P_k(x), P_l(x)\big) = \delta_{k,l}$. These polynomials are called orthogonal polynomials and they are quite important.

1.4 Elementary row transformations and LU decomposition

Recall that the elementary row operations are exchanges of rows, multiplication of a row by a non-zero constant, and subtraction of a row from another row.

From the second interpretation of the matrix product, which we introduced in Section 1.1, we know that we can manipulate columns of matrix A by multiplying A on the right by a matrix C. Similarly, we can manipulate rows of A by multiplying it on the *left* by a suitable matrix C.

In particular, elementary row transformations can be realized by multiplying matrices on the left by elementary matrices. For example, subtraction of the twice the row 1 from the row 2 can be realized by multiplying on the left by the following matrix

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -2 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & & & & \end{bmatrix}$$

These row manipulations are all that we need to perform the Gaussian elimination. If we apply the Gaussian elimination to a square matrix A, at the end we obtain an upper diagonal matrix U.

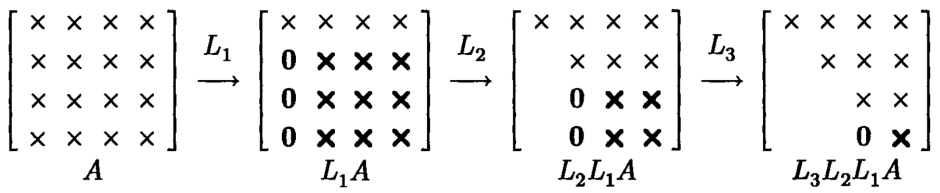


Figure 1.1: Schematics of Gaussian elimination

All the matrices we applied on the left were lower diagonal with 1 on the main diagonal, and the product of such matrices, $\widehat{L} = L_{n-1} \dots L_1$, is also lower diagonal with ones on the main diagonal. We get a formula $\widehat{L}A = U$, where U is upper diagonal.

It is easy to get $\widehat{L}$ by applying the row transformations to the extended matrix: $(A|I_n)$, where I_n is the $n \times n$ identity matrix. Then at the end of the row reduction process, we will get matrix $(U|\widehat{L})$.

The inverse of the matrix $\widehat{L}$ is also lower-diagonal. This can be checked by undoing the transformations one-by-one. For example the inverse of the matrix L_1 above is the matrix

$$L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & \dots \\ 2 & 1 & 0 & \dots \\ 0 & 0 & 1 & \dots \\ \dots & \dots & \dots & 0 \end{bmatrix}$$

And $(\widehat{L})^{-1} = L_1^{-1} \dots L_{n-1}^{-1}$.

So, finally, we get a decomposition

$$A = LU,$$

where $L = (\widehat{L})^{-1}$ is lower-diagonal with ones on the main diagonal and U is upper-diagonal.

This is called the LU decomposition of matrix A. It turns that if this factorization exists then it is unique.

Example 1.4.1 (Calculation of LU factorization).

The algorithm we have just described can fail if at the k-th step the entry $\widetilde{a}_{kk}$ of the transformed matrix $\widetilde{A}$ is zero. Also, from the numerical point of view it is not good if $\widetilde{a}_{kk}$ is very small even if it is not exactly 0. In these cases we can use an entry $\widetilde{a}_{ik}$ with $i > k$ to eliminate entries in column k. Typically, it is done by exchanging the rows i and k and then using the usual elimination operation.

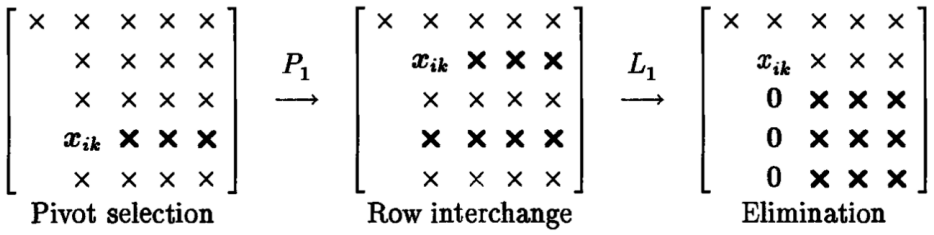


Figure 1.2: Elimination with pivoting

So, we get a formula

$$L_{n-1}P_{n-1}L_{n-2}P_{n-2}\dots L_1P_1A = U,$$

where P_i are permutation matrices, that is, matrices that have exactly one 1 in each row and column and 0s in all other places.

However, it seems that this breaks down the lower-diagonal structure of the matrix that premultiplies A. Somewhat surprising it turns out that this structure is not quite lost and we can write

$$L_{n-1}P_{n-1}L_{n-2}P_{n-2}\dots L_1P_1 = L'_{n-1}L'_{n-2}\dots L'_1P$$

where P is a permutation matrix and L'_i are lower-diagonal matrix. The proof of this a bit tricky and we refer the reader to the textbooks. However, it implies the following result.

Theorem 1.4.2. *Suppose A is an arbitrary square matrix. There is a permutation matrix P, a lower-diagonal matrix L with ones on the main diagonal and an upper-diagonal matrix U, so that*

$$PA = LU.$$

1.5 Exercises

Exercise 1.5.1. Let B be a 4 x 4 matrix to which we apply the following operations:

1. double column 1,
2. halve row 3,
3. add row 3 to row 1,
4. interchange columns 1 and 4,
5. subtract row 2 from each of the other rows,
6. replace column 4 by column 3,
7. delete column 1 (so that the column dimension is reduced by 1).

(a) Write the result as a product of eight matrices.

(b) Write it again as a product ABC (same B) of three matrices.

Exercise 1.5.2. Let X be a matrix

$$X = \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

where A and D are $n \times n$ and $m \times m$ matrices, respectively, and suppose that A is invertible.

The *Schur complement* matrix S is defined through the formula

$$\begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & S \end{bmatrix}$$

Express S in terms of matrices A, B, C and D. What are the dimensions of S?

Chapter 2

Dimension, Range, Nullspace, Rank, Inverse

2.1 Dimension of vector spaces

[This is a stub for a lecture note. For a good explanation see Section 2.5 in S. Treil "Linear Algebra done wrong"]

However, we would also like to mention the following useful fact. Let $V + W := \text{span}(V, W)$ denote the linear subspace formed by all linear combinations of vectors in V and W, and let $V \cap W$ denote the intersection of V and W. Then

$$\dim(V + W) = \dim(V) + \dim(W) - \dim(V \cap W). \tag{2.1}$$

(Intuitively, this is because we can form a basis of the $\text{span}(V, W)$ by taking a basis in $V \cap W$ and complementing it to bases in V and W respectively. Then, if we count elements of the basis of V and then all elements of basis of W, then all elements of the basis of $\text{span}(V, W)$ will be counted but the elements of the basis of $V \cap W$ will be counted twice, so we need to subtract their number from the final count.)

In case when $V \cap W = \{0\}$, the sum $V + W$ is called the direct sum and denoted $V \oplus W$. In this case, we have $\dim(V \oplus W) = \dim(V) + \dim(W)$.

2.2 Range, Nullspace, Rank, Nullity, Inverse Matrix

Definition 2.2.1. The *range* of an $m \times n$ matrix A, denoted Range(A), is the set of vectors in $\mathbb{R}^m$, that can be expressed as Ax for some x.

(It is also frequently denoted Im(A).)

If we use Interpretation 2 for the product Ax, we can easily see that this is the set of all linear combinations of columns of matrix A. This set is a linear space because any linear combinations of vectors from Range(A) is still a linear combination of columns of A, and therefore belongs to Range(A)). Hence, Range(A) is also the linear space spanned by columns of A. (This is Theorem 1.1 in Trethefen-Bau.) For this reason it is also called the *column space* of A.

The *column rank* of a matrix A is the dimension of Range(A) (or its column space).

Definition 2.2.2. The *null-space* of $m \times n$ matrix A is the set of vectors $x \in \mathbb{R}^n$ such that $Ax = 0$. It is denoted Null(A) or ker(A).

The dimension of the nullspace is called *nullity* of A.

Intuitively, if we think about $Ax = 0$ as a system of m equation in n variables, then the nullity measures the size of the solution space. The rank measures the size of the space of those b which can be represented as $b = Ax$.

How we can we find out the rank and nullity of a matrix? The classical method is to reduce the matrix to its reduced row echelon form ("rref") by *elementary row transformations*.

Recall that the elementary row operations are exchanges of rows, multiplication of a row by a non-zero constant, and subtraction of a row from another row.

You can convince yourself that the elementary row operations do not change the dimension of the column space: if several columns are dependent/independent, then they remain dependent/independent after an elementary transformation used in the reduction process.

Formally, this is a consequence of the fact that the matrix multiplication has the distributive property. If columns $v_1, \ldots, v_k$ are linearly dependent: $c_1 v_1 + \ldots + c_k v_k = 0$, and an elementary row transformation is given by the left multiplication by matrix L, then the images of these rows are also linearly dependent:

$$c_1(Lv_1) + \ldots + c_k(Lv_k) = 0.$$

One can also go in the opposite direction because every elementary row operation can be undone by another elementary row operation.

Note, however, that the dimension of the null-space is also not changing under this process! *In fact, the null space itself is the same for the original and the transformed matrix.* (This is actually why the reduction to the row echelon form is used to solve systems of linear equations.) If $Ax = 0$ then $LAx = 0$ so x is in the null-space of the transformed matrix LA. Conversely, if $LAx = 0$, then we know that there is a matrix L^{-1} that undo the elementary transformation L so we can apply it on both sides of the equation and we get: $L^{-1}LAx = L^{-1}0$ so $Ax = 0$.

To summarize: the elementary row transformations do not change the rank and the nullity of a matrix A.

In particular we can reduce A to the reduced echelon row form and for the matrix in the reduced echelon form, it is easy to determine the rank and the nullity of the matrix. They are equal to the number of pivot and free variables, respectively.

Here is an example. Let the matrix be

$$A = \begin{bmatrix} 1 & 3 & 0 & -1 & 2 \\ 2 & 6 & 1 & -1 & 7 \\ 1 & 3 & 1 & 0 & 5 \end{bmatrix},$$

Then we can reduce it to the following matrix

$$\mathrm{rref}(A) = \begin{bmatrix} 1 & 3 & 0 & -1 & 2 \\ 0 & 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

then we have two pivot variables (corresponding to the entries shown in red) and three free variables (corresponding to blue entries), hence the rank and the nullity of the matrix are 2 and 3. (From the algorithmic point of view, the column space $Range(A)$ is generated by the first and the third columns of the original matrix A: $(1, 2, 1)$ and $(0, 1, 1)$, and the null space of A is the same as the null-space of $\mathrm{rref}(A)$ generated, for example, by vectors $(-3, 1, 0, 0, 0)$, $(1, 0, -1, 1, 0)$, and $(-2, 0, -3, 0, 1)$, – we simply set one free variable to 1 and all others to 0 and determine the value of all other variables. So to get the last vector we set $x_2 = 0$, $x_4 = 0$ and $x_5 = 1$; from the equation corresponding to the second row we get $x_3 = -3$, and from the equation corresponding to the first row we get $x_1 = -2$.)

This algorithm essentially proves one of the most fundamental theorems of linear algebra:

Theorem 2.2.3. *For an $m \times n$ matrix A,*

$$\textit{nullity}(A) + \textit{rank}(A) = n.$$

That is, the sum of dimensions of the range and the nullspace are equal to the number of columns. (If we think about A as a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$, then the dimension of the source space $\mathbb{R}^n$ equals the sum of the dimension of the null-space, that is all vectors that will go to 0, and the dimension of the range space.)

One can also define the *row space* and *row rank* similarly, as the linear space spanned by the rows of the matrix A and its dimension.

One can easily check that elementary row operations do not change the row space.[1] Indeed, let $x = c_1 r_1 + \dots c_k r_m$, where $r_1, \dots r_m$ are rows. The rows of a transformed matrix can be easily written as the linear combinations of the rows of the original matrix. For example, if we subtract twice the row 1 from the row 2 then the new rows will be $r'_1 = r_1$, $r'_2 = r_2 - 2r_1$, $r'_3 = r_3$, and so on. Then we can write x in terms of new rows. In our example this is $x = c_1 r'_1 + c_2(r'_2 + 2r'_1) + \dots + c_k r'_k$. So it is clear that x is in the row space of the transformed matrix LA.

It follows the elementary row operations preserve not only the column rank but also the row rank. So the row rank also can be computed as the row rank of the reduced row echelon. But for this form, the row rank equals the number of non-zero rows, so it equals the number of pivot variables! (Look at the previous example to convince yourself.) This proves another fundamental theorem of linear algebra:

Theorem 2.2.4. *For every matrix A,*

$$\textit{column rank}(A) = \textit{row rank}(A).$$

In particular one can simply talk about the rank of a matrix A, denoted $\text{rank}(A)$.[2]

It is clear that $\text{rank}(A) \leq \min\{n, m\}$.

Definition 2.2.5. For an $m \times n$ matrix A, if $\text{rank}(A) = \min\{n, m\}$, we say that matrix A is of *full rank*.

[1] They do not change the nullspace and they do not change the row space but they do change the columns space, although they preserve the dimension of the column space!

[2] For reference, more information about rank can be found in Chapter 2 of Strang's Linear Algebra book and Chapter 2, Section III of Hefferon's Linear Algebra book.

What does this mean intuitively that a matrix has full rank?

If $m \geq n$ then the matrix A is of full rank if A has $\text{rank}(A) = n$ so by the fundamental Theorem 2.2.3 $\text{nullity}(A) = 0$ and so A have the trivial null-space. In particular, if we consider m-by-n matrix A as a map from a linear space of all n-vectors $\mathbb{R}^n$ to the linear space of m-vectors $\mathbb{R}^m$ (our interpretation #1) then *this map is a bijection of* $\mathbb{R}^n$ *onto the column space* $\text{Range}(A)$. In particular, two different vectors of $\mathbb{R}^n$ must go to two different vectors of $\text{Range}(A)$. Otherwise we would have $A\boldsymbol{x}_1 = A\boldsymbol{x}_2$, so $A(\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0$ and if $\boldsymbol{x}_1 - \boldsymbol{x}_2 \neq 0$, we have a contradiction with the triviality of null-space.

If $m \leq n$ then the matrix A has full rank if $\text{rank}(A) = m$. Therefore, A is a *surjection of* R^n *on* $\mathbb{R}^m$. (That is, every vector b in $\mathbb{R}^m$ can be written as Ax for some $x \in \mathbb{R}^n$.)

If $m = n$ (the matrix A is square), and the matrix A has full rank, we see from the previous two facts that map A is a bijection of $\mathbb{R}^n$ on $\mathbb{R}^m \cong \mathbb{R}^n$ and so there exists an inverse transformation. One can check that this transformation is linear. The matrix of this inverse transformation is called the *inverse* of matrix A and denoted A^{-1}.

This is content of Theorem 1.2 in Trefethen-Bau.

Theorem 2.2.6. *A square $n \times n$ matrix A has full-rank if and only if there exists an $n \times n$ inverse matrix A^{-1} with the properties $AA^{-1} = A^{-1}A = I_n$, where I_n is the $n \times n$* identity matrix.

(The identity matrix has ones on the main diagonal and zeros everywhere else: $I_{kl} = 1$ if $k = l$ and $I_{kl} = 0$ if $k \neq l$.)

It is useful to note that it is enough to check only one of the conditions $AA^{-1} = A^{-1}A = I_n$ in the theorem. For example, if a square matrix A has a right inverse B such that $AB = I$, then it is also true that $BA = I$. (It is crucial here that the matrix A is square.)[3]

In particular, in this case for every vector $y = Ax$, we can recover x by using the inverse matrix, as $x = A^{-1}y$.

For numerical applications, it is important to remember that one does not need to calculate the inverse matrix A^{-1} in order to solve the equation $y = Ax$ for one single vector y. The Gaussian elimination (that is, the reduction to the echelon form) which you studied in the first linear algebra

[3]The proof is as follows. Since $AB = I$, hence B has trivial nullspace. (Otherwise we could find a vector $x \neq 0$ such that $Bx = 0$, hence $ABx = 0$. But $(AB)x = Ix = x$ so we have $x = 0$, contradiction.) By the rank-nullity theorem the range of B is $\mathbb{R}^n$ and so it is a bijective transformation $\mathbb{R}^n \mapsto \mathbb{R}^n$. Hence B has a right inverse. Call it X. For this matrix X, we have $BX = I$. Then $(AB)X = X$ and $A(BX) = A$. By associativity of matrix multiplication $X = A$ and so we proved that $BA = I$.

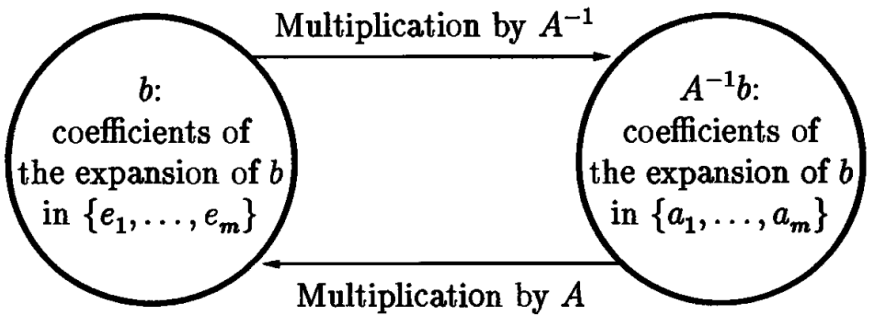


Figure 2.1: Interpretation of multiplication by A, $b = Ax$, as "the change of basis" operation.

course[4] is significantly more efficient and simple method to do it. The only reason for inverting matrix A is if you plan to solve many equations $y = Ax$ for various y.

We have two interpretations for multiplication by matrix A. Correspondingly, there are two interpretations for the multiplication by the inverse matrix A^{-1}.

1. If multiplication by a square matrix A is interpreted as a linear transformation $x \to y = Ax$ from $V = \mathbb{R}^n$ to $W = \mathbb{R}^n$, then multiplication by A^{-1} is simply the inverse transformation $y \to x$ from $W \to V$.

2. If the multiplication is understood as taking a linear combination of columns of A with coefficients from x, then $A^{-1}y$ is the vector of coefficients of the expansion of y in *the basis of columns of* A. In other words, multiplication by A^{-1} can be understood as the change of basis operation. (We are given y in a standard basis, and we have a new basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. We write the coordinates of each $\boldsymbol{v}_i$ in the standard basis as i-th column of matrix A. Then the coordinates of y in the new basis are given by the entries of the vector $A^{-1}y$.)

Note, however, that in this calculation of coefficients of a vector in the basis of columns of A we assumed that A is $n \times n$ and the range of A equals the target space R^n. We will talk more about this later when A is $m \times n$, $m > n$ and so the range can be smaller than $\mathbb{R}^m$.

[4]Chapter 1 in Strang, or Chapter 1, Section I of Hefferon

It is also worthwhile to mention two useful properties of the inverse operation:

1. $(AB)^{-1} = B^{-1}A^{-1}$.
2. $(A^t)^{-1} = (A^{-1})^t$.

To see that the first property holds, note that AB is the composition of the linear maps A and B in which B acts first, and A second. We can invert this composition by doing A^{-1} first and B^{-1} second. This corresponds to the product $B^{-1}A^{-1}$.

For the second property, let $B := A^{-1}$. Then, we have $AB = I$. By taking the transposition on both sides and using one of the properties of tthe transposition operation, we get $B^tA^t = I$. This means that $B^t = (A^t)^{-1}$.

Example 2.2.7 (Calculation of the inverse matrix: Gauss-Jordan method). Example from Strang p.53. Calculate the inverse of the following matrix

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix}$$

Answer:

$$A^{-1} = \begin{bmatrix} \frac{12}{16} & -\frac{5}{16} & -\frac{6}{16} \\ \frac{4}{8} & -\frac{3}{8} & -\frac{2}{8} \\ -1 & 1 & 1 \end{bmatrix}$$

2.3 Rank-1 matrices

We looked at the matrices of full rank. What about the matrices of small rank? One important case occurs when we have a matrix of rank one. In this case, the dimension of the range is 1, and by the fundamental Theorem 2.2.3 the dimension of the null space is $n - 1$.

Since the dimension of column space is 1, it means that all columns are proportional to a single column $(b_1, \ldots, b_m)^t$. So, the matrix can be written as follows:

$$A = \begin{bmatrix} a_1b_1 & a_2b_1 & \ldots & a_nb_1 \\ a_1b_2 & a_2b_2 & \ldots & a_nb_2 \\ \ldots & & & \\ a_1b_m & a_2b_m & \ldots & a_nb_m \end{bmatrix} = \boldsymbol{b}\boldsymbol{a}^t,$$

where $\boldsymbol{b} = (b_1, \ldots, b_m)^t$ and $\boldsymbol{a} = (a_1, \ldots, a_n)^t$ are two column vectors. Hence, we can conclude that *every matrix of rank 1 is an outer product of two (non-zero) vectors.*

2.4 Exercises

Exercise 2.4.1. Let $v_1 = [1, 2, 3]^t$, $v_2 = [0, 1, 3]^t$, $v_3 = [1, 0, 1]^t$. Find the coordinates of the vector $x = e_1 = [1, 0, 0]^t$ in the basis $\{v_1, v_2, v_3\}$.

Exercise 2.4.2. Write the matrix $\left(\left((AB)^t\right)^{-1}\right)^t$ in terms of A^{-1} and B^{-1}.

Exercise 2.4.3. By using the reduction to the rref form, find the bases for the column space and nullspace of A and the solution to $Ax = b$:

$$A = \begin{bmatrix} 2 & 4 & 6 & 4 \\ 2 & 5 & 7 & 6 \\ 2 & 3 & 5 & 2 \end{bmatrix} \quad b = \begin{bmatrix} 4 \\ 3 \\ 5 \end{bmatrix}$$

Exercise 2.4.4. Let $f_1, \dots f_8$ be a set of functions defined on the interval $[1, 8]$ with the property that for any numbers $d_1, \dots, d_8$, there exists a set of coefficients $c_l, \dots, c_8$ such that

$$\sum_{j=1}^{8} c_j f_j(i) = d_i, \quad i = 1, \dots, 8.$$

(a) Show by appealing to the theorems of lecture 1 in Trefethen, Bau that $d_1, \dots, d_8$ determine $c_1, \dots, c_8$ uniquely.

(b) Let A be the 8×8 matrix representing the linear mapping from data $d_1, \dots, d_8$ to coefficients $c_1, \dots, c_8$. What is the i, j entry of A^{-1}?

Exercise 2.4.5. Let u and v are two vectors in $\mathbb{R}^n$. The matrix $A = I + uv^*$ is known as a *rank-one perturbation of the identity*. Show that if A is nonsingular (that is, if it has an inverse), then its inverse has the form $A^{-1} = I + \alpha uv^*$ for some scalar α and give an expression for α. For what u and v is A singular? If it is singular, what is Null(A)?

Chapter 3

Norms and orthogonality

Reading for this Chapter

- Trefethen, Bau: Lecture 2, 3
- Strang: Chapter 3.

3.1 Vector norms

Intuitively, a norm is a way to measure how long is a given vector.

Mathematically, a *norm* is a non-negative function on a linear space, which has the property $\|cv\| = |c|\|v\|$, and satisfy the triangle inequality: $\|u + v\| \leq \|u\| + \|v\|$. It is also required that $\|u\| = 0$ implies that $u = 0$.

These axioms are satisfied by the Euclidean length of vector u,

$$\|u\| = \sqrt{u_1^2 + \ldots + u_n^2} = \sqrt{u^* u}.$$

(We will sometimes write $|u|$ instead of $\|u\|$ if no confusion can arise.) This norm is called the Euclidean norm, and it is the usual norm that will use here.

It is useful to know that there are other norms besides the Euclidean norm. For example, a p-norm is defined for every $p \geq 1$ as follows. If $x \in \mathbb{R}^n$, then

$$\|x\|_p = \Big(\sum_{i=1}^{n} |x_i|^p \Big)^{1/p}.$$

$$\|x\|_1 = \sum_{i=1}^{m} |x_i|,$$

$$\|x\|_2 = \left(\sum_{i=1}^{m} |x_i|^2 \right)^{1/2} = \sqrt{x^* x},$$

$$\|x\|_\infty = \max_{1 \le i \le m} |x_i|,$$

$$\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p \right)^{1/p} \quad (1 \le p < \infty).$$

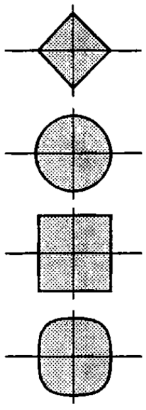


Figure 3.1: Unit balls for different vector norms

This is an exercise that this function is indeed a norm.[1]

If we look at $p \to \infty$ then we get a so-called supremum norm:

$$\|x\|_\infty = \sup_i |x_i|.$$

In this notation, the Euclidean norm can be called 2-norm since it corresponds to the case $p = 2$. So, more proper notation for this norm would be $\|v\|_2$. However, we will usually use this norm rather than any other p-norm and so we will skip this subscript.

The great advantage of the 2-norm (i.e., the Euclidean norm) is that it equals the square root of the inner product of the vector with itself. Because of this, it enjoys some properties which are not true for other norms. For example, if we want to find out what is the point in a linear subspace with the smallest distance from a given point, where the distance is measured using the 2-norm, then we can use the orthogonal projection operator (which we discuss later). In contrast, if we measure distance not in the usual 2-norm but in a different norm, then this would not be true anymore and it would be more difficult to find this point.

On the other hand, the p-norms for $p \neq 2$ are sometimes used in modern statistics, so you should know about them. For example, the *lasso regression* uses the 1-norm of vectors.

[1] In contrast, one can check that if $p < 1$, then $\|x\|_p$ is not a norm. This is a couple of additional exercises. First is to check that if $\|\cdot\|$ is a norm, then this implies that the unit ball $B = \{x : \|x\| \le 1\}$ must be convex. And the second is to check that if $p < 1$, then the unit ball is not convex.

For the Euclidean vector norm, besides the triangle inequality, we have the Cauchy-Schwarz inequality:

$$|x^t y| \leq \|x\|\|y\|,$$

and for more general p-norms, we have the Holder inequality:

$$|x^t y| \leq \|x\|_p \|y\|_q,$$

where $p^{-1} + q^{-1} = 1$.

3.2 Orthogonality

Orthogonality is a very useful concept. If we have an orthogonal system of vectors, then it becomes much simpler to calculate the length of a linear combination of these vectors provided that the length is measured by using the usual Euclidean norm (that is the 2-norm). In addition the orthogonal matrices have the property that corresponding transformations conserve the lengths.

3.2.1 Orthogonal vectors and matrices

Definition 3.2.1. Two vectors u, v are called *orthogonal* (or *perpendicular*) if their inner product is zero, $u^* v = 0$. In this case we will write $u \perp v$.

A set of vectors $u_1, \ldots, u_n$ is called an *orthogonal system* if they are all non-zero and they are pairwise orthogonal: $u_i \perp u_j$ for all $i \neq j$. It is called an *orthonormal system* if it is an orthogonal system and each of these vectors have length 1.

One useful thing about systems of orthogonal vectors is that we can use them to decompose an arbitrary vector in orthogonal components. First of all, we have the following result.

Theorem 3.2.2. *The vectors of an orthogonal system are linearly independent.*

Proof. Suppose they are dependent. The we can write, after reordering these vectors,

$$v_1 = \sum_{i=2}^{n} \lambda_i v_i,$$

where at least one of λ_i is not zero. Say, $\lambda_i \neq 0$. Then $(v_1, v_i) = \lambda_i |v_i|^2 \neq 0$, and vectors v_1 and v_i are not orthogonal. □

And here is how to decompose an arbitrary vector v as a linear combination of the vectors in the orthonormal system and a "residual".

Theorem 3.2.3. *Let $\{u_1, \dots u_n\}$ is an orthonormal set of vectors in $\mathbb{R}^m$, where $m \geq n$. Then for every vector $v \in \mathbb{R}^m$, there exists a unique decomposition:*

$$v = r + \sum_{i=1}^{n} c_i u_i,$$

in which vector r is orthogonal to each of vectors u_i. The coefficients can be computed as $c_i = (u_i, v) = (u_i^* v)$.

Note: the theorem remains valid for complex vectors.

Proof. The existence will be proved if we show that

$$r = v - \sum_{i=1}^{n} (u_i^* v) u_i$$

is orthogonal to each of vectors u_i. By multiplying with u_i, we get

$$(r, u_i) = (v, u_i) - (u_i, v)(u_i, u_i) = 0,$$

which is the required property.

For uniqueness, we note that if we have two decompositions like that, then we can subtract them. As a result we would have that either $r = r'$, and u_i are linearly dependent, or $r \neq r'$ and the orthogonal set $r - r', u_1, \dots, u_n$ is linearly dependent. Both are not possible by Theorem 3.2.2.

□

A real matrix is called *orthogonal* if:
(i) it is a square matrix, and
(ii) The set of its column vectors is orthonormal.

(If a matrix has complex entries and satisfies conditions (i) and (ii), it is called a *unitary* matrix.) The usual notation for orthogonal and unitary matrices is Q and U.

The definition essentially says that Q is orthogonal (unitary) if and only if it is square and

$$Q^* Q = I,$$

where I is the identity matrix, that is, $I_{ij} = \delta_{ij}$.

For *square* matrices A and B, the identity $AB = I$ implies that $BA = I$. (This is a good exercise.) If Q is orthogonal then we also have $QQ^* = I$. (Note that if Q is not square then it can happen that $Q^*Q = I$, but $QQ^* \neq I$.)

Theorem 3.2.3 implies that the columns of the $n \times n$ orthogonal matrix Q form a basis in Q (since in this case the maximal number of linearly independent vectors is n), and the coefficients of a vector v in this basis can be computed very conveniently as $c = Q^*v$.

Finally, an important property of linear transformations with orthogonal (or unitary) matrix Q is that they preserves lengths of vectors.

$$\|Qv\|^2 = (Qv)^*Qv = v^*Q^*Qv = v^*v = \|v\|^2.$$

3.2.2 Orthogonal subspaces and complements

Two linear subspaces V and W are orthogonal to each other ($V \perp W$) if *every (non-zero) vector in V is orthogonal to every (non-zero) vector in W*.

Note that the *intersection of two orthogonal subspaces is always zero* (i.e., the trivial subspace). Indeed, if v belongs to two subspaces simultaneously, then $v \perp v$ and $\|v\|^2 = v^*v = 0$, which implies that $v = 0$.

Let $V \subset \mathbb{R}^m$ be a linear subspace. Then its *orthogonal complement* of V in $\mathbb{R}^m$, denoted $V^\perp$, is the largest linear subspace in $\mathbb{R}^m$ orthogonal to V. Alternatively, it is the set of all vectors u that are orthogonal to V. Formally:

$$V^\perp = \{u \in \mathbb{R}^m : u^*v = 0 \text{ for all } v \in V\}.$$

Theorem 3.2.4. *For any $V \subset \mathbb{R}^m$,*

$$\operatorname{span}(V, V^\perp) = \mathbb{R}^m.$$

Proof. Let $W = \operatorname{span}(V, V^\perp)$. Suppose, by seeking contradiction, that $W \neq \mathbb{R}^m$. Then, we can find a vector $x \notin W$. Take an orthonormal basis $w_1, \ldots, w_n$ of W and apply Theorem 3.2.3 to write $x = r + \sum_{i=1}^n c_i w_i$. This gives a vector $r \neq 0$ such that $r \notin W$ and therefore $r \notin V^\perp$. In addition, r is orthogonal to all w_i. In particular, $r \perp W$. Since $V \subset W$ so $r \perp V$ and therefore $r \in V^\perp$. Contradiction.

□

(This proof have one non-clear step since we assumed that we are always able to build an orthonormal basis of W. Later we will see how to find this basis by the Gram-Schmidt orthogonalization process.)

Theorem 3.2.5. *If $V \in \mathbb{R}^m$ and $\dim(V) = k$ then $\dim(V^\perp) = m - k$.*

Proof. Let $W = \text{span}(V, V^\perp) \subset \mathbb{R}^m$. Then we can use the fact that $V \cap V^\perp = 0$ and therefore by (2.1), $\dim(W) = \dim(V) + \dim(V^\perp)$, and therefore $\dim(V^\perp) = \dim(W) - \dim(V)$.

So it is enough to show that $W = \mathbb{R}^m$. But this holds by Theorem 3.2.4. □

Corollary 3.2.6. *If $V^\perp$ is orthogonal complement to V in $\mathbb{R}^m$, then V is an orthogonal complement to $V^\perp$ in $\mathbb{R}^m$.*

Proof. It is clear that all vectors in V are orthogonal to all vectors in $V^\perp$ (simply by definition of $V^\perp$). So $V \subset (V^\perp)^\perp$. In addition, we can calculate that $\dim(V) = \dim((V^\perp)^\perp)$ and we use the fact that if one linear space is a subspace of another one and they have the same dimension then they must coincide. □

Since V and $V^\perp$ have zero intersection and $\dim(V) + \dim(V^\perp) = m$, therefore we can construct the basis of R^m by taking the union of the bases of V and $V^\perp$. In particular, every vector u in $\mathbb{R}^n$ can be represented in a unique fashion as $v + w$ where $v \in V$ and $w \in W$.

How can we calculate this decomposition? We will see it later.

Now, here is an important example of orthogonal complements.

Theorem 3.2.7. *For an $m \times n$ matrix A*

1. *The nullspace of A is the orthogonal complement of the row space of A (i.e., the range of A^t):*

$$\text{Null}(A)^\perp = \text{Range}(A^t) \text{ and } \text{Range}(A^t)^\perp = \text{Null}(A).$$

2. *The range of A is the orthogonal complement of the left nullspace of A:*

$$\text{Range}(A)^\perp = \text{Null}(A^t) \text{ and } \text{Null}(A^t)^\perp = \text{Range}(A).$$

Proof. It is enough to prove one of these claims, since the proof of the other follows by considering the transpose of matrix A. Let us prove the first one.

If $u \in \mathbb{R}^n$ is in the row space of A then $u = A^*z$ for some vector $z \in R^m$ (since u is a linear combination of rows of A, or the columns of A^*, and we can form the linear combination by multiplying the matrix A^* by a vector z). Let $v \in \text{Null}(A)$, then we can calculate the scalar product

$$u^*v = u^*v = (A^*z)^*v = z^*Av = 0,$$

where the last step holds because $v \in \mathbb{R}^n$ is in the null-space of A.

Hence the row space and the null-space are orthogonal. There is one small detail left, namely, that we cannot find a larger orthogonal space to the null-space.

The dimension of the row space of A is $rank(A)$ and the dimension of the null-space of A is $n -$ rank so the sum of dimensions is n and we can conclude that they are not only orthogonal but are actually complements of each other. □

In particular, it gives a method to calculate the orthogonal complement to a subspace spanned by vectors $\boldsymbol{c}_1, \ldots \boldsymbol{c}_n$. Write the matrix C^t with rows given by $\boldsymbol{c}_1^t, \ldots \boldsymbol{c}_n^t$, and calculate its nullspace (that is, the basis of the nullspace).

Example 3.2.8 (Example of calculation). Find a vector in the orthogonal complement to the column space of matrix

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 2 & 4 \\ 2 & 2 \end{bmatrix}.$$

3.3 Exercises

Exercise 3.3.1. Let $V = \mathbb{R}^5$ and let U be the subspace of V spanned by the vectors

$$\begin{bmatrix} 1 \\ 2 \\ -1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 2 \\ 2 \\ 1 \\ -2 \end{bmatrix},$$

and W the subspace of V spanned by the vectors

$$\begin{bmatrix} 3 \\ 2 \\ -3 \\ 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -4 \\ -1 \\ -2 \\ 1 \end{bmatrix}.$$

Determine the dimension of $U \cap W$.

Exercise 3.3.2. Find all vectors that are perpendicular to (1, 4, 4, 1) and (2, 9, 8, 2).

Exercise 3.3.3. In the vector space $V = \mathbb{R}^5$, consider the subspace U spanned by the vectors

$$\begin{bmatrix} 2 \\ 2 \\ 1 \\ 7 \\ -3 \end{bmatrix}, \begin{bmatrix} -4 \\ 1 \\ -12 \\ 6 \\ -4 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 3 \\ 4 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 3 \\ 1 \\ 2 \end{bmatrix}, \text{ and } \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

(a) Compute $\dim U$.

(b) Which of the vectors

$$\begin{bmatrix} 4 \\ 0 \\ 5 \\ -3 \\ -1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 8 \\ 4 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 \\ 2 \\ 4 \\ 0 \\ 0 \end{bmatrix}, \text{ and } \begin{bmatrix} 1 \\ 0 \\ 5 \\ 0 \\ 2 \end{bmatrix}$$

belong to U?

Exercise 3.3.4. Find a matrix A, which is in reduced echelon form, and satisfies $\dim(\text{Range}(A^t)^\perp) = 4$, $\dim(\text{Range}(A)^\perp) = 1$.

Exercise 3.3.5. Let A be a real symmetric matrix. An *eigenvector* of matrix A is a non-zero vector x such that $Ax = \lambda x$ for some number λ which is called the *eigenvalue* corresponding to the eigenvector x.

Prove that if x and y are eigenvectors corresponding to distinct real eigenvalues λ_1 and λ_2, then x and y are orthogonal.

Chapter 4

Gram-Schmidt process and QR decomposition

Reading for this Chapter

- Trefethen, Bau: Lecture 7
- Strang: Section 3.4

In some cases we are given a basis $(a_1, a_2, \ldots)$ of a linear space V and we want to construct a orthonormal basis $(q_1, q_2, \ldots, q_n)$. More generally, we are given an increasing sequence of spaces (a *flag*)

$$V_1 \subset V_2 \subset \ldots \subset V_n,$$

where $V_k = span(a_1, \ldots, a_k)$, and we want to construct an orthonormal system of vectors $q_1, \ldots, q_n$ so that $V_k = span(q_1, \ldots q_n)$. This can be easily done by the process that is called the *Gram-Schmidt orthogonalization*.

The process is recursive. At step 1, we take vector a_1 and normalize it to have the unit length:

$$q_1 = \frac{1}{r_{11}} a_1,$$

where $r_{11} = \|a_1\|$.

At step k we take vector a_k and subtract its projection on the subspace V_{k-1}. This is easy to do because we already know $(q_1, \ldots, q_{k-1})$, which form an orthonormal basis of V_{k-1}. After this, we normalize the resulting vector so that it has the unit length.

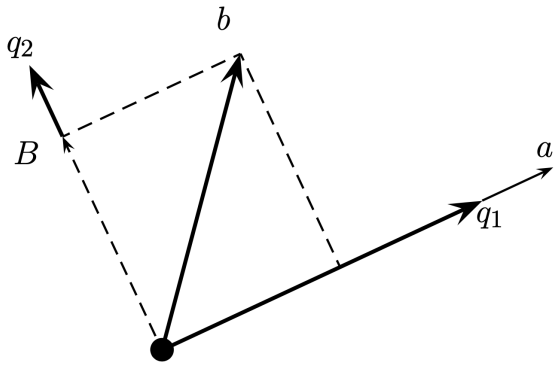


Figure 4.1: A step of the Gram-Schmidt orthogonalization. $\boldsymbol{q}_1 = \boldsymbol{a}/\|\boldsymbol{a}\|$, $B = \boldsymbol{b} - (\boldsymbol{q}_1^*\boldsymbol{b})\boldsymbol{q}_1$, $\boldsymbol{q}_2 = B/\|B\|$.

So,

$$\begin{aligned} v_k &= a_k - (q_1^* a_k)q_1 - \ldots - (q_{k-1}^* a_k)q_{k-1}, \\ q_k &= \frac{1}{r_{kk}} v_k, \end{aligned}$$

where $r_{kk} = \|v_k\|$. (Note also that $r_{kk} = q_k^* v_k = q_k^* a_k$.)

The process will continue without interruption, provided that the inclusions $V_{k-1} \subset V_k$ are strict, which is the same as that the matrix A with columns $a_1, \ldots, a_n$ has full rank.

The formulas above can also be written differently, as

$$\begin{aligned} a_1 &= r_{11}q_1, \\ a_2 &= r_{12}q_1 + r_{22}q_2, \\ a_3 &= r_{13}q_1 + r_{23}q_2 + r_{33}q_3, \\ &\ldots \\ a_n &= r_{1n}q_1 + r_{2n}q_2 + \ldots + r_{nn}q_n, \end{aligned}$$

where $r_{ij} = q_i^* a_j$ when $i \le j$.

In a matrix form it can be written as

$$A = \widehat{Q}\widehat{R},$$

where A is an $m \times n$ matrix, $\widehat{Q} = [q_1, \ldots, q_n]$ is an $m \times n$ matrix with orthonormal columns and R is an upper-diagonal $n \times n$ matrix with positive diagonal elements.

$$\widehat{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1n} \\ 0 & r_{22} & r_{23} & \dots & r_{2n} \\ 0 & 0 & r_{33} & \dots & r_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & r_{nn} \end{bmatrix}$$

This factorization is called the *reduced QR factorization* and the above argument shows that if matrix A has full rank, then this factorization exists and is unique. By extending matrix $\widehat{Q}$ to an orthogonal $m \times m$ matrix Q, and $\widehat{R}$ to an upper-diagonal $m \times n$ matrix R one can obtain the full QR factorization, although this factorization is not unique.

Above, we showed how to calculate the QR factorization by using the Gram-Schmidt orthogonalization. There exists a faster method to calculate this factorization based on so-called Householder reflections. For details, see the textbook by Trefethen and Bau.

Here is an example of a QR factorization from Strang's textbook.

$$\begin{aligned} A &= \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 1/\sqrt{2} & \sqrt{2} \\ & 1/\sqrt{2} & \sqrt{2} \\ 0 & 0 & 1 \end{bmatrix} = QR \end{aligned}$$

The Gram-Schmidt orthogonalization is a general process, which can be applied not only to vectors in $\mathbb{R}^m$ but also to functions in a linear space of functions. One only needs to define the scalar product of two functions. For example if $f(x)$ and $g(x)$ are two real-valued functions, then we can define the scalar product as an integral, provided that the integral is convergent.

There are different ways to define the scalar products,

$$(f, g) := \int_{-\infty}^{\infty} f(x)g(x)\,dx, \text{ or} \tag{4.1}$$

$$(f, g) := \int_{-1}^{1} f(x)g(x)\,dx, \text{ or} \tag{4.2}$$

$$(f, g) := \int_{-\infty}^{\infty} e^{-x^2} f(x)g(x)\,dx, \text{ or} \tag{4.3}$$

$$\dots$$

Two functions are called orthogonal if their scalar product is zero, and the norm of a function is defined naturally as $\|f\| = \sqrt{(f,f)}$.

Many properties of the vector norms can be extended to functions. In particular the triangle inequality and the Cauchy-Schwarz inequalities hold.

The Gram-Schmidt orthogonalization can also be applied to a system of functions $f_1(x), \ldots, f_n(x)$ and results in a system of orthonormal functions $q_1(x), \ldots, q_n(x)$.

For example, many famous families of polynomials can be obtained in this way by applying orthogonalization procedure to polynomials $1, x, x^2, x^3, \ldots$ with respect to various scalar products.

The Legendre polynomials are orthonormal with respect to scalar product 4.2, Hermite's polynomials are orthonormal with respect to scalar product 4.3, etc.

This is important for the problems when one approximates functions by other functions.

4.1 Exercises

Exercise 4.1.1. From the nonorthogonal $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, $\boldsymbol{a}_3$, find orthonormal vectors $\boldsymbol{q}_1$, $\boldsymbol{q}_2$, $\boldsymbol{q}_3$.

$$\boldsymbol{a}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \boldsymbol{a}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \boldsymbol{a}_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix},$$

Write matrix A as QR decomposition $A = QR$.

Chapter 5

Projectors

5.1 Definition and properties

Reading:

- Section 3.3 in Strang
- Lecture 6 in Trefethen-Bau

In the previous chapter we have seen that if we have a system of n orthonormal vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ in $\mathbb{R}^m$, then we can project every vector $\boldsymbol{x} \in \mathbb{R}^m$ on their span by calculating the projection vector

$$P(\boldsymbol{x}) = (\boldsymbol{q}_1^* \boldsymbol{x})\boldsymbol{q}_1 + \ldots + (\boldsymbol{q}_n^* \boldsymbol{x})\boldsymbol{q}_n, \tag{5.1}$$

and the residual vector

$$\boldsymbol{r} = \boldsymbol{x} - P(\boldsymbol{x}),$$

and we know that $\boldsymbol{r}$ is orthogonal to every vector $\boldsymbol{q}_i$.

We can write the projection operator in formula (5.1) by using the matrix $Q = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n]$. It easy to see that

$$P(\boldsymbol{x}) = QQ^* \boldsymbol{x},$$

The matrix QQ^* is a symmetric square $m \times m$ matrix with the following property.

$$(QQ^*)^2 = Q(Q^*Q)Q^* = QQ^*,$$

because Q^*Q is the identity matrix by orthonormality of vectors $\boldsymbol{q}_i$. Intuitively it says that projecting the same vector twice on the same subspace does not change the results.

The residual can be written as

$$\boldsymbol{r} = (I - QQ^*)\boldsymbol{x},$$

and it is easy to see that $(I - QQ^*)$ is also a symmetric matrix that has the property that

$$(I - QQ^*)^2 = I - QQ^*$$

Now, we want to look at things more generally. What does it mean to project a vector on a subspace from an algebraic point of view?

Definition 5.1.1. A *projector* is a square matrix P that satisfies the equation $P^2 = P$.

We do not require P to be symmetric at this moment.

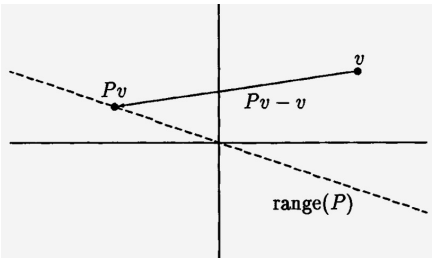


Figure 5.1: Oblique Projector

The *complementary projector* for a projector P is defined as $I - P$. It is indeed a projector, since $(I-P)^2 = I - 2P + P^2 = I - P$.

It is called complementary because if we apply $I-P$ and then apply P then we get zero. So while P maps $\mathbb{R}^m$ on Range(P), $I-P$ maps $\mathbb{R}^m$ on nullspace of P. Formally, we have the following proposition.

Proposition 5.1.2. Range$(I - P)$ = Null(P).

Proof. 1. $P(I - P)v = 0$, so Range$(I - P) \subset$ Null(P);

2. if some vector $u \in$ Null(P) then $Pu = 0$ and we can write $u = (I-P)u$ so Null$(P) \subset$ Range$(I - P)$.

□

Now, we have a decomposition of an arbitrary vector $v = Pv + (I - P)v$ into a sum of two vectors. One of them is from the range of P and another one is from the nullspace of P. So we have a decomposition of vector v as a sum of projections on the range and nullspace of P.

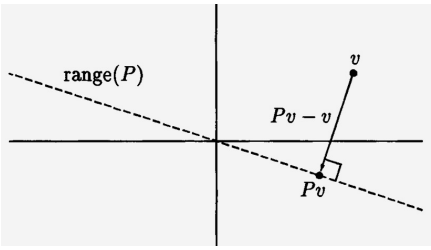


Figure 5.2: Ortogonal Projector

The most useful projectors are orthogonal projectors, for which the *vectors in this decomposition are orthogonal to each other*. However, we need an additional condition to ensure that this property holds.

Example 5.1.3. Consider the outer product of two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. Then,

$$(\boldsymbol{u}\boldsymbol{v}^*)^2 = (\boldsymbol{v}^*\boldsymbol{u})\boldsymbol{u}\boldsymbol{v}^*.$$

So, according to Definition 5.1.1, this matrix is a projector if and only if $(\boldsymbol{v}, \boldsymbol{u}) = \boldsymbol{v}^*\boldsymbol{u} = 1$. However, the range of this matrix is the set of all vectors that are proportional to u and the nullspace is all vertices orthogonal to v. So the matrix is an orthogonal projector only if $\boldsymbol{v}$ is proportional to $\boldsymbol{u}$.

We can define the orthogonal projector as a projector P for which the range and nullspace of P are orthogonal. However, it turns out that there is a simpler definition, which is equivalent to this requirement.

Definition 5.1.4. An *orthogonal projector* P is a projector which is symmetric (or hermitian in the complex case): $P^* = P$.

Let us check that the range and nullspace of P are indeed orthogonal to each other. Indeed if $\boldsymbol{u}_1 \in \text{Range}(P)$ and $\boldsymbol{u}_2 \in \text{Null}(P)$, then we can write $\boldsymbol{u}_1 = P\boldsymbol{v}$ and so

$$u_2^* u_1 = u_2^* P v = (P^* u_2)^* v = 0^* v = 0.$$

Note that this argument would not work if $P^* \neq P$.

Note that the complementary projector is also orthogonal since $(I - P)^* = I - P^* = I - P$.

Example 5.1.5. Suppose v is a column vector that has unit length. Then matrix $P = vv^t$ is an orthogonal projector. Indeed,

$$P^2 = (vv^*)(vv^*) = v(v^*v)v^* = vv^* = P,$$

where in the second equality we used the fact that the matrix product is associative and in the third equality that the vector has unit length. It is also clear that $P^* = P$.

The projector in Example 5.1.5 is called a rank-one projector because its range is one-dimensional: it is spanned by the vector v. Geometrically,

$Px = v(v^*x)$ is the projection of the vector x on the line L that has the direction vector v.

One particular case is when $v = \frac{1}{\sqrt{n}}[1, 1, \ldots, 1]^*$. In this case $vv^* = \frac{1}{n}J$, where J is the $n \times n$ matrix consisting of all 1s.

Example 5.1.6. We have seen at the beginning of this section that if a matrix Q has column vectors $q_1, q_2, \ldots, q_n$ which form an orthonormal set, then matrix $P = QQ^*$ is an orthogonal projector on the linear space spanned by these vectors.

It is sometimes useful to write $P = QQ^*$ as a sum of rank-one projectors from the previous example.

$$P = QQ^* = \sum_{i=1}^{n} q_i q_i^*$$

Example 5.1.7. Now consider even more general case, when we want to project on a vector space V spanned by vectors $a_1, a_2, \ldots, a_n$ which are not necessarily orthogonal. Let matrix A have columns a_i. Then we claim that the orthogonal projection on V is

$$P = A(A^*A)^{-1}A^*$$

Here we assume that A^*A is full rank and therefore invertible. (This requirement is equivalent to the requirement that columns $a_1, \ldots, a_n$ of matrix A are linearly independent, or equivalently that the nullspace of A is trivial. See the Strang's book p.184 for the proof of the fact that $\text{Null}(A^*A) = \text{Null}(A)$.)

First, by direct checking, $P^2 = P$ and $P^* = P$, so P is an orthogonal projection and we only need to check that it has the correct range which should be V and the correct null-space, which should be the orthogonal complement to V, denoted $V^\perp$.

Indeed, if a vector y is in V, then this means that it is in the range space of A, that is, there is a vector x such that $y = Ax$. In this case it is obvious that

$$Py = PAx = A(A^*A)^{-1}A^*Ax = Ax = y,$$

so P preserves vectors in V. It remains to show that the vectors in the orthogonal complement to V are sent to 0 by P. Since every vector $v \in V^\perp$ is orthogonal to every column in A, so we can write that $A^*v = 0$. Then it is obvious that

$$Pv = A(A^*A)^{-1}A^*v = 0.$$

5.2 Relation to Least Squares Regression

- Section 3.3 in Strang
- Lecture 11 in Trefethen-Bau

In statistics we often need to solve the following problem:

$$y_i = \beta_1 x_i^{(1)} + \ldots \beta_n x_i^{(n)} + \varepsilon_i, \tag{5.2}$$

where $i = 1, \ldots, m$ labels observations, y_i is the value of the variable that we want to explain in observation i, and $x_i^{(1)}, \ldots, x_i^{(n)}$ are the values of n "explanatory" variables in observation i. (They often called "features" in machine learning.) In statistics, a linear regression is usually has a constant term. Here we treat the constant term on the equal basis with other coefficients. For example, we can think about the vector $\boldsymbol{x}^{(1)} = (x_1^{(1)}, \ldots, x_n^{(1)})$ as the vector in which all components are equal to 1.

The numbers ε_i are "error terms". In statistics, y_i and sometimes also x_i are treated as random variables and ε_i are usually assumed to be taken from a random process, often from a process of i.i.d. random variables and sometimes from the process of i.i.d. Gaussian random variables. In this example we are not interested in the random nature of y_i, x_i and ε_i. We simply assume that we observed vectors $\boldsymbol{y} = (y_i)$, and $\boldsymbol{x}^{(k)} = (x_i^{(k)})$ where $i = 1, \ldots, n$ and $k = 1, \ldots, p$ but that we do not know β_k and ε_i.

One simple statistical method is Ordinary Linear Regression. It prescribes to choose those coefficients β_j, $j = 1, \ldots, n$ that the sum of the squares of ε_j is at its minimum. (There is also a generalized least squares method that weights different error terms differently.)

Another view on this problem is that we simply trying to solve an overdetermined system of equations, where the number of equations m exceeds the number of variables n. In this case, there is no exact solution and we trying to minimize the norm of the vector of the residual terms ε_i.

We want to develop a simple formula for these values of β_j.

Let us introduce $m \times 1$ vector $y = [y_1, \ldots, y_m]$, an $m \times n$ matrix X with entries $X_{ij} = x_i^{(j)}$, the $n \times 1$ vector of coefficients $\beta = [\beta_1, \ldots, \beta_n]$, and $m \times 1$ vector of errors $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]$.

Then we can re-write equation (5.2) as

$$y = X\beta + \varepsilon,$$

Our task is to minimize the norm of vector ε, which we can write as

$$(y - X\beta)^*(y - X\beta) \to \min$$

We can write the first order conditions as

$$\frac{\partial}{\partial\beta}(y - X\beta)^*(y - X\beta) = 0.$$

Here, $\frac{\partial}{\partial\beta}f(\beta)$ is the vector of partial derivatives $\frac{\partial}{\partial\beta_j}f(\beta)$. One can check directly that this leads to equations:

$$X^*(y - X\beta) = 0,$$

or

$$X^*X\beta = X^*y. \tag{5.3}$$

(Indeed

$$\frac{\partial}{\partial\beta_j}\sum_i(y_i - \sum_k X_{ik}\beta_k)^2 = -2\sum_i X_{ij}(y_i - \sum_k X_{ik}\beta_k),$$

and this is equivalent to equation (5.3).)

In the traditional statistics, $m > n$, the number of observations exceeds the number of explanatory variables. For this reason the rank of a typical X equals n, so it is a full rank. It follows that X^*X is invertible and we can solve equation (5.3) as

$$\beta = (X^*X)^{-1}X^*y \tag{5.4}$$

The equations in (5.3) are called *normal equations* and the matrix

$$X^+ = (X^*X)^{-1}X^*$$

is sometimes called the *pseudoinverse* of matrix X.

In statistical applications we are also interested in estimated true values of y_i, when the noise ε_i is filtered out. So we define the fitted values of y as $\widehat{y} = X\beta$. Then

$$\widehat{y} = X(X^*X)^{-1}X^*y.$$

This is the linear combination of explanatory random variables which minimizes the norm of the error term $e = y - X\beta$.

From the point of view of linear algebra, $\widehat{y}$ is the orthogonal projection of vector y on the linear space spanned by the vectors of the explanatory variables $x^{(1)}$, ..., $x^{(n)}$. The matrix of the projection is

$$P = X(X^*X)^{-1}X^*$$

Note the matrix $(X^*X)^{-1}$ is n-by-n, so the normal equations 5.3 are n equations in n variables. They can be solved in various ways, for example by Gaussian elimination, which is has the work of around n^3 floating point operations (flops). Since the matrix is symmetric and positive definite, this can be solved also by Cholesky factorization in $n^3/3$ flops. We will discuss the Cholesky factorization later. In addition, one has to compute X^*X which requires mn^2 flops.

One other method of solving the normal equations is through the QR factorization. Essentially its idea is to find the orthogonal basis in the space spanned by the columns of X and calculate the projection by using this basis.

Technically, we compute the QR factorization $X = QR$. Then the normal equations become $R^*R\beta = R^*Q^*y$, which leads to the equation $R\beta = Q^*y$. So the algorithm is

1. compute the QR factorization for X.

2. Calculate $b = Q^*y$ [coefficients of the projected vector in the basis given by columns of Q.]

3. Solve the equation

$$R\beta = b.$$

[This gives the coefficients of the projected vector in the old basis.]

The last equation is easy to solve recursively because the matrix R is upper-diagonal. The work is dominated by the first step and requires approximately $2mn^2 - \frac{2}{3}n^3$ flops which is worse than the Cholesky factorization method if m is large. Apparently, however, this method behaves better with respect to accumulation of numerical errors. Its another advantage is that there is a variant of QR factorization algorithm adapted for sparse matrices (that is, the matrices that have large number of zero entries). Since the sparseness is lost in computation of A^*A, the QR algorithm can have an advantage in speed in these cases.

Example 5.2.1. Let

$$A = \begin{bmatrix} 1, & 0 \\ 0, & 1 \\ 1, & 0 \end{bmatrix}$$

What is the orthogonal projector P onto Range(A) and what is the image under P of the vector $v = [1, 2, 3]^*$?

Solution:

$$A^*A = \begin{bmatrix} 2, & 0 \\ 0, & 1 \end{bmatrix}, (A^*A)^{-1} = \begin{bmatrix} 1/2, & 0 \\ 0, & 1 \end{bmatrix}, (A^*A)^{-1}A^* = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix},$$

$$P = A(A^*A)^{-1}A^* = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix}.$$

So,

$$P[1, 2, 3]^* = [2, 2, 2]^*.$$

Note that if one wants only to calculate Pv with $v = [1, 2, 3]^*$, then one does not need to calculate $(A^*A)^{-1}$. One can simply calculate sequentially $u = A^*v$, then solve the system $A^*Ax = u$ and finally calculate Ax.

Example 5.2.2. Projection on a plane.

Example 5.2.3. Curve fitting.

There are some generalizations of the linear regression when one puts different weights on different observations. In this case we can introduce a weighted 2-norm:

$$\|x\|_{2,w} = \sqrt{\sum_{i=1}^{n} w_i x_i^2},$$

where w_i are some positive weights. In this case one wants to minimize

$$\|y - X\beta\|_{2,w}^2,$$

and this leads to the formulas

$$\widehat{\beta} = (X^*WX)^{-1}X^*Wy,$$
$$\widehat{y} = X(X^*WX)^{-1}X^*Wy,$$

where W is the diagonal matrix with weights w_i on the main diagonal.

The prediction matrix $X(X^*WX)^{-1}X^*W$ is a projector, but it is not symmetric so this projection is not orthogonal. However, it turns out that it can be thought as orthogonal if we define the orthogonality differently, namely if we say that two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal if $\sum_{i=1}^{n} w_i u_i v_i = 0$.

Other generalizations of the linear regression are related to the choice of other vector norms. However, they not always result in prediction being given by a linear projection.

Recently, there was a lot of interest in the statistical community in the case when the number of explanatory variables n exceeds the number of observations m. In this case, the $n \times n$ matrix X^*X is not invertible and we cannot solve the normal equations.

A popular approach is to change a minimization target. Instead of minimizing the norm of the error term $\|y - X\beta\|$, one minimizes a "regularized loss function", which is either

$$\|y - X\beta\|^2 + \lambda\|\beta\|_2^2, \text{ or } \|y - X\beta\|^2 + \lambda\|\beta\|_1,$$

where λ is a regularization parameter, $\|\beta\|_2$ and $\|\beta\|_1$are the 2-norm and 1-norm of the parameter vector $\|\beta\|$, respectively. These two methods are called the *ridge regression* and the *lasso regression*, respectively.

The ridge regression leads to the solution

$$\widehat{\beta} = (X^*X + \lambda I_n)^{-1}X^*Y.$$

Note that this is not a projection matrix.

The solution of the lasso regression formula cannot be given by a simple formula but there are efficient algorithms for its calculation.

The lasso regression became recently especially popular since it often results in a vector β which has a lot of zeros as its components.

5.3 Exercises

Exercise 5.3.1. If P is an orthogonal projector, then the matrix $I - 2P$ is orthogonal. Prove this algebraically, and try to give a geometric interpretation for the transformation represented by matrix $I - 2P$.

Exercise 5.3.2. Let E be the $m \times m$ matrix that extracts the even part of an m-vector: $Ex = (x + Fx)/2$, where F is the $m \times m$ matrix that flips $(x_1, \ldots, x_m)^*$ to $(x_m, \ldots, x_1)^*$. Is E an orthogonal projector, an oblique projector, or not a projector at all? What are its entries?

Exercise 5.3.3. Given an $m \times n$ matrix A with $m \geq n$, show that A^*A is non-singular if and only if A has full rank.

Chapter 6

Determinants

Reading for this Chapter

- Strang: Chapter 4.

6.1 Definitions

Consider an $n \times n$ matrix A with columns $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$. We might be interested in the volume of the parallelepiped (or, in simpler terms, the box) spanned by vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$.

However, one difficulty is that this function is somewhat complicated. if we call this function $\mathrm{vol}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$, then the equality

$$\mathrm{vol}(\boldsymbol{a}_1 + \boldsymbol{a}_1', \ldots, \boldsymbol{a}_n) = \mathrm{vol}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) + \mathrm{vol}(\boldsymbol{a}_1', \ldots, \boldsymbol{a}_n) \qquad (6.1)$$

sometimes holds and sometimes not: for example, volume on the left is zero if $\boldsymbol{a}_1' = -\boldsymbol{a}_1$ and the volumes on the right are (almost always) positive.

For this and other reasons, it is useful to define the signed volume of the parallelepiped. The absolute value of the signed volume equals the regular volume and its sign is determined by the *orientation* of the system of vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$. We will not define the orientation rigorously but only note that it can be either positive or negative, and the orientation preserved by rotations but changes sign after a reflection.

We denote this signed volume by $\mathrm{Vol}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$. In particular $\mathrm{Vol}(v_1, v_2) = -\mathrm{Vol}(v_2, v_1)$ and $\mathrm{Vol}(-v_1, v_2) = -\mathrm{Vol}(v_1, v_2)$.

The geometric definition of the determinant of matrix A is that

$$\det(A) = \mathrm{Vol}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n).$$

This definition is a bit unsatisfactory since it depends on the definitions of volume and orientation and so it is not purely algebraic. It also does not explain how to compute the determinant.

The second definition is axiomatic. The determinant is a function that maps matrix $A = [a_1, \ldots, a_n]$ to real numbers and satisfies the following axioms:

1. For all real numbers $c \in \mathbb{R}$,

$$\det[c\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] = c \det[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n], \tag{6.2}$$

   (More generally, this identity should hold for all c from the field over which we define the matrices.)

2.

$$\det[\boldsymbol{a}_1 + \boldsymbol{a}'_1, \ldots, \boldsymbol{a}_n] = \det[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] + \det[\boldsymbol{a}'_1, \ldots, \boldsymbol{a}_n], \tag{6.3}$$

3. For every $1 \leq i < j \leq n$, we have

$$\det[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_i, \ldots, \boldsymbol{a}_j, \ldots, \boldsymbol{a}_n] = -\det[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_j, \ldots, \boldsymbol{a}_i, \ldots, \boldsymbol{a}_n]. \tag{6.4}$$

4. For the basis vectors $\boldsymbol{e}_1 = (1, 0, 0 \ldots, 0)$, $\boldsymbol{e}_2 = (0, 1, 0, \ldots, 0)$, ..., $\boldsymbol{e}_n = (0, 0, 0, \ldots 1)$, we have

$$\det[\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n] = 1 \tag{6.5}$$

This is a nice definition but it is not clear why this function exists. (The signed volume satisfies these axioms but our goal is to avoid using the signed volume in the definition.)

We are going to show the existence of the determinant by writing the function $\det(A)$ explicitly in terms of the entries of A and checking that it satisfies the axioms. Unfortunately, this definition is somewhat cumbersome.

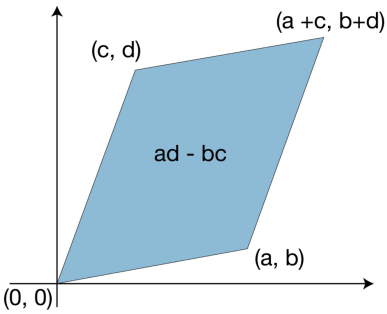


Figure 6.1: Area of a parallelogram in $\mathbb{R}^2$

It can perhaps be guessed by considering examples in 2-dimensional and 3-dimensional space. For example, if we have two vectors $v_1 = (a, b)$ and $v_2 = (c, d)$ then one can show that the area of the corresponding parallelogram is $|ad - bc|$. One can also develop a formula for the volume of 3-dimensional

parallelepiped. From this, one can guess the general definition.

Here is the constructive definition of the determinant:

$$\det(A) = \sum_{\pi \in S_n} \varepsilon(\pi) a_{1\pi(1)} a_{2\pi(2)} \dots a_{n\pi(n)}. \tag{6.6}$$

Here the sum is over all permutations of the set $\{1, 2, \dots, n\}$. A permutation is the bijective mapping of this set to itself. For example, we can define a permutation of the set $\{1, 2, 3, 4\}$ by setting $\pi(1) = 3, \pi(2) = 4, \pi(3) = 2, \pi(4) = 1$. This permutation can also be written in two-line notation:

$$\pi = \begin{matrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{matrix}$$

or simply in one-line notation 3421 (since the first line is always the same). For each permutation, we can define its length $l(\pi)$ as the minimal number of switches of two elements which is needed to bring it to the identical transformation. For example for our transformation $\pi = 3421$, we can undo it as follows:

$$3421 \xrightarrow{31} 1423 \xrightarrow{42} 1243 \xrightarrow{43} 1234,$$

so the length of this permutation is three.

Then we define the function $\varepsilon(\pi) := (-1)^{l(\pi)}$, and now our formula (6.6) is well-defined.

For example, for a 2×2 matrix A we have only two permutations 12 and 21 with lengths 0 and 1 respectively, and the formula for the determinant is

$$\det(A) = a_{11}a_{22} - a_{12}a_{21},$$

with the first term corresponding to the identity permutation 12 and the second to the permutation 21.

If $n = 3$, we have one permutation of length 0: 123, three permutations of length 1: 213, 132, and 321, and two permutations of length 2: 231 and 312. So the formula for the determinant in this case is

$$\begin{aligned} \det(A) = {} & a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ & - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31}. \end{aligned}$$

The number of terms in these formulas grows very fast so they are not useful for the actual computation of the determinants except for small matrices.

The terms in the definition of the determinant can be re-organized to give the recursive formula for the determinant in terms of the determinants of sub-matrices.

This is called the *cofactor expansion* of the determinant.

Let A be an $n \times n$ matrix with entries a_{ij}. Let $A^{(ij)}$ be a matrix which is obtained by removing row i and column j. Then the *cofactor*

$$C_{ij} := (-1)^{i+j} \det\left(A^{(ij)}\right).$$

The *cofactor expansion* along the row i is the formula

$$\det(A) = \sum_{j=1}^{n} a_{ij} C_{ij}.$$

For example, for the first row, we have the expansion:

$$\begin{aligned} \det(A) &= a_{11}C_{11} + a_{12}C_{12} + \ldots + a_{1n}C_{1n} \\ &= a_{11} \det\left(A^{(11)}\right) - a_{12} \det\left(A^{(12)}\right) + \ldots + (-1)^{n+1} a_{1n} \det\left(A^{(1n)}\right) \end{aligned}$$

We have also an analogous expansion along column j:

$$\det(A) = \sum_{i=1}^{n} a_{ij} C_{ij}.$$

(So altogether, there are $2n$ different expansions, n along the rows and n along the columns.)

This result can be proved from the basic definition (6.6). In a sense, this is simply a way to organize formula (6.6) as a recursive calculation. We omit the proof.

The cofactor expansion can be used to calculate the determinant recursively but for large matrices this is usually much slower than by reducing the matrix to the upper-diagonal form, the method which we describe a bit later.

6.2 Properties of the determinant

One case in which it is easy to calculate the determinant from the definition (6.6) or from a cofactor expansion is the case in which the matrix is either lower diagonal or upper-diagonal. In this case, it is easy to see that the only non-zero term in the sum in formula (6.6) is the term corresponding to the identity permutation $\pi = 12 \ldots n$. This leads to the following theorem.

Theorem 6.2.1. *If a square matrix A is upper-diagonal, or lower-diagonal, then*

$$\det A = a_{11}a_{22}\dots a_{nn}.$$

Formula (6.6) is also very useful from the theoretical point of view. First of all, it is possible to show that the determinant defined in this way actually satisfies the axioms listed above. So, the determinant which we defined axiomatically actually exists.

Second, this definition allows us to prove an important theorem.

Theorem 6.2.2. *For every square matrix A, we have:*

$$\det(A^t) = \det(A).$$

(Note that we use here the transposition sign t instead of $*$. Even if the matrix is complex, we should use the transposition, not the conjugate transposition so that the theorem holds true.)

Proof. For the transposed matrix A^t, we have

$$\begin{aligned}\det(A^t) &= \sum_{\pi\in S_n} \varepsilon(\pi) a_{\pi(1)1}a_{\pi(2)2}\dots a_{\pi(n)n} \\ &= \sum_{\pi\in S_n} \varepsilon(\pi) a_{1\pi^{-1}(1)}a_{2\pi^{-1}(2)}\dots a_{n\pi^{-1}(n)},\end{aligned}$$

where π^{-1} is the inverse permutation to the permutation π. For example, if $\pi = \begin{bmatrix}1 & 2 & 3\\ 3 & 1 & 2\end{bmatrix}$, then $\pi^{-1} = \begin{bmatrix}3 & 1 & 2\\ 1 & 2 & 3\end{bmatrix} = \begin{bmatrix}1 & 2 & 3\\ 2 & 3 & 1\end{bmatrix}$.

It turns out that the length of the inverse permutation π^{-1} equals the length of permutation π. (If we can undo the permutation π by some sequence of transpositions then we can undo π^{-1} by a related sequence of transpositions that has the same length.) Hence $\varepsilon(\pi^{-1}) = \varepsilon(\pi)$ and we can continue the formula above as:

$$\det(A^t) = \sum_{\pi\in S_n} \varepsilon(\pi^{-1}) a_{1\pi^{-1}(1)}a_{2\pi^{-1}(2)}\dots a_{n\pi^{-1}(n)}.$$

However if π in this sum are all possible permutations of the set $\{1,\dots,n\}$, then π^{-1} also run over all possible permutations of this set. So it is actually the same sum as in definition of $\det(A)$ and we conclude that $\det(A^t) = \det(A)$. □

In particular, this implies that properties (6.2) - (6.4) hold not only for matrix A with *columns* $a_1, \ldots, a_n$ but also for matrix with *rows* $a_1, \ldots, a_n$.

Here are useful consequences of these basic properties:

Theorem 6.2.3. *1. If one row in the matrix is a multiple of another row then the determinant equals 0.*

2. If we add a multiple of row a_i to any other row a_j then the determinant will not change.

Proof. For the proof of the first property, without loss of generality let the second row be a multiple of the first row, then

$$\det([a_1; ca_1; \ldots]) = c \det([a_1; a_1; \ldots]) = -c \det([a_1; a_1; \ldots]),$$

where in the last equality we exchanged rows 1 and 2 and used the property (6.4). This implies that the determinant is zero.

For the proof of the second property, again without loss of generality, assume that we added a multiple of the first row to the second row. Then we have:

$$\begin{aligned} \det([a_1; a_2 + ca_1; \ldots]) &= \det([a_1; a_2; \ldots]) + \det([a_1; ca_1; \ldots]) \\ &= \det([a_1; a_2; \ldots]), \end{aligned}$$

which is what we wanted to prove. (The first equality uses property (6.3) and the second one uses the property that we just proved.) □

In particular, this theorem means that if we do Gaussian elimination on matrix A (by adding the multiples of rows above to the rows below but without multiplying the rows by a constant, and without exchanging the rows), then the determinant of the matrix will not change and eventually we will be left with an upper-diagonal matrix U that have the same determinant as the original matrix. Hence, by Theorem 6.2.1 the determinant of A equals the product of the diagonal elements of U.

If we had to exchange the rows, then a more general formula applies:

$$\det(A) = (-1)^r u_{11} u_{22} \ldots u_{nn}, \tag{6.7}$$

where r is the number of times we exchanged the rows, and u_{11}, ..., u_{nn} are the pivots, that is the diagonal elements of U.

This property gives an effective method to calculate the determinants.

Example 6.2.4. Calculate the determinant of

$$A = \begin{bmatrix} 1 & -4 & 2 \\ -2 & 8 & -9 \\ -1 & 7 & 0 \end{bmatrix}$$

in two different ways, by using the cofactor expansion and the Gaussian elimination.

By using the cofactor expansion over the first row, we get:

$$\begin{aligned} \det(A) &= 1 \times \det \begin{bmatrix} 8 & -9 \\ 7 & 0 \end{bmatrix} - (-4) \times \det \begin{bmatrix} -2 & -9 \\ -1 & 0 \end{bmatrix} + 2 \times \det \begin{bmatrix} -2 & 8 \\ -1 & 7 \end{bmatrix} \\ &= 63 - 36 - 12 = 15. \end{aligned}$$

By raw reduction, we find

$$\begin{bmatrix} 1 & -4 & 2 \\ -2 & 8 & -9 \\ -1 & 7 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & -4 & 2 \\ 0 & 0 & -5 \\ 0 & 3 & 2 \end{bmatrix} \sim \begin{bmatrix} 1 & -4 & 2 \\ 0 & 3 & 2 \\ 0 & 0 & -5 \end{bmatrix},$$

where we used one exchange of rows. Consequently,

$$\det(A) = (-1)^1 \times 1 \times 3 \times (-5) = 15.$$

For large matrices, the method that uses Gaussian elimination is much more effective than the cofactor expansion method.

Example 6.2.5 (The Vandermonde determinant). One important determinant which pop-ups in many parts of mathematics is the Vandermonde determinant:

$$\det(W) = \det \begin{bmatrix} x_1^{n-1} & x_2^{n-1} & \dots & x_n^{n-1} \\ x_1^{n-2} & x_2^{n-2} & \dots & x_n^{n-2} \\ \cdots & \cdots & \cdots & \cdots \\ x_1 & x_2 & \dots & x_n \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

By the definition, this should be a polynomial of variables $x_1, x_2, \dots, x_n$ and that every term in this polynomial has the same total degree. It is clear that the determinant equal to zero if $x_i = x_j$ for some $i \neq j$. So, the polynomial should be divisible by all of the differences $x_i - x_j$. By checking the total degree, it follows that the polynomial is equal to the product of

these differences up to the constant term and the by looking on a specific term like $x_1^{n-1}x_2^{n-2}\dots x_{n-1}$ one can find this constant. This results in the following formula:

$$\det(W) = \prod_{i<j}(x_i - x_j).$$

The formula 6.7 for determinant in terms of pivots implies the following important property of the determinants.

Theorem 6.2.6. *A square matrix A is invertible if and only if* $\det(A) \neq 0$.

Proof. The matrix A is invertible if and only if it has full rank, hence, if and only if after raw reduction all the variables u_{ii} are valid pivots, that is, $u_{ii} \neq 0$, hence, by formula (6.7) if and only if $\det(A) \neq 0$. □

Another important result is that determinant multiplicative.

Theorem 6.2.7. *Let A and B be two $n \times n$ matrices, then*

$$\det AB = \det A \det B.$$

Sketch of the proof. The argument has two cases. The first case is when one of the matrices A and B is non-invertible (has a non-zero null-space). Then it is easy to check that AB is also non-invertible and we are done by Theorem 6.2.6.

The second case is when A and B are both invertible. Say, A is invertible. Then we can reduce the matrix A by elementary row operations to the identity matrix. (This is the process by which we obtain the reduced row echelon form. In this case this form is the identity matrix because A is invertible). We can reverse the process and write A as a product of matrices corresponding to elementary transformations:

$$A = E_s E_{s-1} \dots E_1.$$

Then we check that the claim of the Theorem holds for the case when we multiply B by an elementary matrix. That is, if E is an elementary matrix and X is an arbitrary $n \times n$ matrix, then $\det(EX) = \det(E)\det(X)$. This holds because of the properties of the determinant in (6.2)–(6.4) and in Theorem 6.2.3. Then, by induction, we have:

$$\det(A) = \det(E_s)\det(E_{s-1})\dots\det(E_1),$$

and

$$\det(AB) = \det(E_s)\det(E_{s-1})\dots\det(E_1)\det(B),$$

which implies that $\det(AB) = \det(A)\det(B)$. □

The unpleasant part about this proof is that it depends very much on the fact that A is a finite matrix. The determinants can be generalized to some linear transformations in infinite-dimensional spaces by requiring that the axioms are satisfied and checking the existence and uniqueness properties. This proof will not work for these generalizations. Alternatively, one can define a certain function of matrix A as $\det(AB)/\det(B)$ (where B is considered a fixed non-singular matrix) and check that it satisfies all axioms. Then by uniqueness, one can conclude that this function equal the determinant of A.

Corollary 6.2.8. *For a non-singular square matrix A, we have:*

$$\det(A^{-1}) = \frac{1}{\det(A)}$$

Corollary 6.2.9. *Suppose $n \times n$ matrix V has columns $v_1, \dots, v_n$, and let A be another $n \times n$ matrix. Then*

$$\mathrm{Vol}(Av_1, Av_2, \dots, Av_n) = \det(A)\,\mathrm{Vol}(v_1, v_2, \dots v_n).$$

Proof. Since $Av_1, Av_2, \dots, Av_n$ are columns of the matrix AV, we have

$$\begin{aligned}\mathrm{Vol}(Av_1, Av_2, \dots, Av_n) &= \det(AV) = \det(A)\det(V)\\ &= \det(A)\,\mathrm{Vol}(v_1, v_2, \dots v_n).\end{aligned}$$

□

In other words, suppose we have a box with the sides given by vectors $v_1, v_2, \dots, v_n$ and suppose this box has the oriented volume V, and we apply a linear transformation A to this box. Then this box will be mapped to a box with the oriented volume $\det(A)V$. This gives another interpretation of the determinant: it is a scale factor by which a linear transformation A extends the volume elements.

6.3 Inverse matrix and Cramer formula

A formula for inverse matrix.

Recall that cofactors are defined as

$$C_{ij} := (-1)^{i+j} \det\left(A^{(ij)}\right).$$

We can think about them as entries of a matrix C. Recall also that the (row) *cofactor expansion* along the row i is the formula

$$\det(A) = \sum_{j=1}^{n} a_{ij} C_{ij}.$$

Another use of cofactors is that they provide us with a formula for the inverse matrix.

Theorem 6.3.1. *Let A be a non-singular $n \times n$ matrix. Then*

$$A^{-1} = \frac{1}{\det(A)} C^t,$$

meaning that

$$(A^{-1})_{ij} = \frac{1}{\det(A)} C_{ji}.$$

Sketch of the proof: We need to prove that

$$AC^t = \det(A) I,$$

that is, that

$$\sum_{j=1}^{n} a_{ij} C_{kj} = \det(A) \delta_{ik}.$$

For $k = i$ this is simply the cofactor expansion, while for $k \neq i$, the left-hand side is the cofactor expansion for the determinant of the matrix which is obtained from matrix A by replacing the row k with the row i (and keeping all other rows intact). However, this new matrix has two identical rows and therefore its determinant is 0. □

The formula is important theoretically. However, in numerical computations, the inverse is easier to find by using the Gaussian elimination method.

Example 6.3.2. The inversion formula for 2×2 matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Example 6.3.3. If a matrix A has integer entries and $\det A = 1$, then its inverse is also an integer matrix.

A related result is Cramer's formula for the solution of linear equations.

Theorem 6.3.4. *For an invertible matrix A, the k-th entry of the solution of the equation $Ax = b$ is given by the formula*

$$x_k = \frac{\det B_k}{\det A},$$

where the matrix B_k is obtained from A by replacing column number k of A by the vector b.

Sketch of the proof: The solution is

$$x = A^{-1}b = \frac{1}{\det A} C^t b,$$

so

$$x_k = \frac{1}{\det A} \sum_{i=1}^{n} C_{ik} b_i,$$

and one can identify the sum as the cofactor expansion for the determinant of the matrix B_k along the column k. □

Again, from the practical point of view, this formula is not very useful for calculations. However, from the theoretical viewpoint, it means that the solution of *any* system of equations can be written in terms of determinants.

6.4 Advanced properties of determinant

The theory of determinants have some beautiful identities. Here are several of them, which we give without proof.

Determinant of a block-diagonal matrix. Suppose A and D are square $k \times k$ and $l \times l$ matrices respectively, and let B is a $k \times l$ matrix. Then we

can form a block matrix $\begin{bmatrix} A & B \\ 0 & D \end{bmatrix}$ which is a square $(k+l) \times (k+l)$ matrix. (Here 0 denotes a $l \times k$ matrix of zeros.) Then

$$\det \begin{bmatrix} A & B \\ 0 & D \end{bmatrix} = \det(A) \det(D).$$

Schur's identity is a generalization of this formula. Suppose A is square and invertible and B, C and D are such that the matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ is square, then

$$\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det(A) \det(D - CA^{-1}B)$$

Silvester's determinantal identity. Let A and B be $m \times n$ and $n \times m$ matrices, respectively. Then

$$\det(I_m + AB) = \det(I_n + BA)$$

The Cauchy - Binet formula. The Cauchy-Binet formula allows one to calculate the determinant of AB if A and B are not square. So, it is a generalization of the product formula for the determinant. Suppose A is $m \times n$ and B is $n \times m$ and assume that $m \leq n$ (otherwise, it is easy to show that $\det(AB) = 0$. Then one has the formula:

$$\det(AB) = \sum_S \det(A(:, S)) \det(B(S, :))$$

where the sum is over all m-element subsets S of the set $\{1, \ldots, n\}$, $A(:, S)$ is an $m \times m$ matrix whose columns are the columns of A at indices from S, and $B(S, :)$ is an $m \times m$ matrix whose rows are the rows of A at indices from S.

6.5 Exercises

Exercise 6.5.1. Use a determinant to identify all values of t and k such that the following matrix is singular. Assume that h and k must be real numbers.

$$A = \begin{bmatrix} 0 & 1 & t \\ -3 & 10 & 0 \\ 0 & 5 & k \end{bmatrix}$$

Exercise 6.5.2. Let $A = [\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}]$ be a 4×4 matrix whose determinant is equal to 2. What is the determinant of $B = [\boldsymbol{d}, \boldsymbol{b}, 3\boldsymbol{c}, \boldsymbol{a} + \boldsymbol{b}]$? Explain.

Exercise 6.5.3. By applying row operations to produce an upper triangular U, compute the following determinants:

1.
$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

2.
$$A = \begin{bmatrix} 1 & t & t^2 & t^3 \\ t & 1 & t & t^2 \\ t^2 & t & 1 & t \\ t^3 & t^2 & t & 1 \end{bmatrix}$$

Exercise 6.5.4. True or false, with reason if true and counterexample if false:

1. If A and B are identical except that $b_{11} = 2a_{11}$, then $\det(B) = 2\det(A)$.
2. The determinant is the product of the pivots.
3. If A is invertible and B is singular, then $A + B$ is invertible.
4. If A is invertible and B is singular, then AB is singular.
5. The determinant of $AB - BA$ is zero.

Exercise 6.5.5. Find the determinant of an $n \times n$ matrix $A = I + J$ where I is the identity matrix and J is a matrix with all entries equal to 1.

Chapter 7

Eigenvalues and eigenvectors

Reading for this Chapter

- Strang: Chapter 5.
- Trefethen-Bau: Lecture 24

7.1 Definition and relation to characteristic polynomial

The main goal of the theory of eigenvalues and eigenvectors is to determine a basis in which the matrix of a transformation has the simplest possible form.

For this theory, some knowledge of complex numbers is unavoidable. Some basics are summarized in appendix.

7.1.1 Eigenvalue diagonalization

A non-zero vector $x \in \mathbb{C}^m$ is an *eigenvector* of an $m \times m$ matrix A if $Ax = \lambda x$ and then λ is its corresponding *eigenvalue*. The set of all eigenvalues of a matrix A is called the spectrum of the matrix A. It is a subset of the plane of complex numbers.

Can the spectrum be empty? The answer is "no" as we will see a bit later.

Now, if λ is an eigenvalue, then the corresponding eigenvectors form a linear space, which is called an *eigenspace*. We denote it by E_λ. The dimension of this eigenspace is called the *geometric multiplicity* of λ.

Given that we know an eigenvalue λ of A, it is easy to calculate the corresponding eigenspace. It is simply the null-space of matrix $A - \lambda I$. The calculation of eigenvalues is more involved. We will discuss methods for doing that later.

The importance of eigenvectors and eigenvalues stems from the following observation. Suppose that we have a basis $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of $\mathbb{C}^m$ that consists from eigenvectors of matrix A. Then we have

$$A \begin{bmatrix} | & | & \ldots & | \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_n \\ | & | & \ldots & | \end{bmatrix} = \begin{bmatrix} | & | & \ldots & | \\ \lambda_1 \boldsymbol{x}_1 & \lambda_2 \boldsymbol{x}_2 & \ldots & \lambda_n \boldsymbol{x}_n \\ | & | & \ldots & | \end{bmatrix}$$
$$= \begin{bmatrix} | & | & \ldots & | \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_n \\ | & | & \ldots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix},$$

or

$$AX = X\Lambda,$$

where X is the matrix with columns x_1, ..., x_n, and Λ is the diagonal matrix with diagonal entries equal to eigenvalues $\lambda_1, \ldots, \lambda_n$. Since $x_1, \ldots, x_n$ is a basis, the matrix X is full rank and therefore, it is invertible. So,

$$A = X\Lambda X^{-1},$$

This factorization of matrix A is called the *eigenvalue diagonalization* of matrix A.

Intuitively, in the basis of eigenvectors, the linear transformation A looks as a stretch in the directions given by eigenvectors by factors given by the eigenvalues.

Of course, a rotation is difficult to imagine like a stretch transformation. Indeed, the eigenvalue diagonalization is impossible over the real numbers, since there are no real eigenvectors. However, it is still possible over the complex numbers.

Still, even over complex numbers, the diagonalization can be impossible if there are not enough eigenvectors to form a basis. One case, when it is always possible is when there are n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. In this case, it is possible to show that the corresponding n eigenvectors are linearly independent and so form a basis.

More generally, the basis exists if the sum of geometric multiplicities of eigenvalues of A is n.

Now, how can we calculate the eigenvalues and their geometric multiplicities?

7.1.2 Characteristic polynomial

Definition 7.1.1. The *characteristic polynomial* of A is the polynomial

$$p_A(z) = \det(zI - A).$$

(Remark: sometimes the characteristic polynomial is defined as $\det(A - zI)$, which is different from our definition by the sign if the size of them matrix n is odd.)

A very important theorem connects eigenvalues and the characteristic polynomial.

Theorem 7.1.2. *A number λ is an eigenvalue of A if and only if $p_A(\lambda) = 0$.*

Proof. Indeed, λ is an eigenvalue if and only if there is a non-zero vector x (its corresponding eigenvector), such that $(\lambda I - A)x = 0$. This happens if and only if the square matrix $\lambda I - A$ is singular (that is, if it is invertible). And here we can use a property of the determinant that the singularity of matrix $\lambda I - A$ is equivalent to $\det(\lambda I - A) = 0$ (Theorem 6.2.6). □

Example 7.1.3. Find the characteristic polynomial, eigenvalues and eigenvectors of the following matrices:

$$\begin{bmatrix} 1 & 2 \\ 8 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}.$$

Answers: For the first example: $p(z) = (z-1)^2 - 16$, $\lambda_{1,2} = -3, 5$, $x_1 = [1, 2]^t$, $x_2 = [1, -2]^t$.

For the second example : $p(z) = (z-1)^2 + 4$, $\lambda_{1,2} = 1 \pm 2i$, $x_1 = [1, i]^t$, $x_2 = [1, -i]^t$.

Example 7.1.4 (Rotation matrix). What are eigenvalues of matrix

$$R_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}?$$

Example 7.1.5 (Orthogonal matrices). What can be said about eigenvalues of orthogonal (or unitary) matrices?

Orthogonal matrices preserve the norm of vectors, so it is easy to see that all their eigenvalues must have absolute value 1.

Example 7.1.6 (Projections). What can be said about eigenvalues and eigenspaces of the projection matrices?

For projection matrices, all eigenvalues are either 0 or 1 and the corresponding eigenspaces are the nullspace and the range of the matrix.

Theorem 7.1.2 implies that every matrix has at least one eigenvalue and so its spectrum is not empty. This is a consequence of the fundamental theorem of algebra that says that every polynomial which is not identically constant has at least one root, which might be a complex number.

Moreover, the characteristic polynomial $p_A(z)$ of an $m \times m$ matrix A has degree m and the fundamental theorem of algebra gives us some additional information. Namely, we can write p_A in the form

$$p_A(z) = (z - \lambda_1)(z - \lambda_2) \dots (z - \lambda_m),$$

where λ_i are eigenvalues of A. The number of times a given eigenvalue λ_j appears in this product is called the *algebraic multiplicity* of λ_j. An eigenvalue is called *simple* if its algebraic multiplicity is 1.

In particular, we see that the number of distinct eigenvalues is between 1 and m. If all roots of $p_A(z)$ are simple, then A has m distinct eigenvalues and, as we prove later, the matrix A is diagonalizable. (This is the generic situation. If all entries of A are real numbers chosen at random from a continuous distribution, then with probability 1 the roots of $p_A(z)$ are simple. If the entries are not real but say integer, and the matrix A is large then the probability that a root is not simple is not zero but very small.)

Now how do geometric and algebraic multiplicities are related?

Example 7.1.7. Here is an example that the geometric and algebraic multiplicities of an eigenvalue can be different. Consider matrices

$$A = \begin{bmatrix} 2 & & \\ & 2 & \\ & & 2 \end{bmatrix} \text{ and } B = \begin{bmatrix} 2 & 1 & \\ & 2 & 1 \\ & & 2 \end{bmatrix}$$

The characteristic polynomial for both matrices is $p(z) = (z - 2)^3$, so the only eigenvalue is $\lambda = 2$ and it has the algebraic multiplicity 3 for both matrices. However it is easy to check that the eigenspace of $\lambda = 2$ is the whole space $\mathbb{R}^3$ in case of matrix A, and the line spanned by the vector $e_1 = (1, 0, 0)$ in case of matrix B.

It is possible to show that the algebraic multiplicity is never smaller than the geometric multiplicity. (See Theorem 7.2.2 below.) We say that an eigenvalue is *defective* if its algebraic multiplicity is greater than its geometric multiplicity. A matrix is *defective* if it has one or more defective eigenvalues. It turns out that a matrix is diagonalizable if and only if it is not defective.

7.2 Change of basis and similarity of matrices

Suppose that a linear transformation L has matrix A in the standard basis. Now we consider another basis $x_1, \ldots, x_n$ and want to know what is the matrix of L in the new basis. To understand this, let $X = [x_1, \ldots, x_n]$ be a matrix whose columns are $x_1, \ldots, x_n$ and let a vector v has coordinates $b_1, b_2 \ldots b_n$ in the new basis. That means that $v = b_1 x_1 + \ldots b_n x_n$, so its coordinates in the *standard* basis are the entries of the vector Xb. This vector will go to AXb under transformation L. The entries of this vector are still coordinates in the standard basis, so in order to re-write it in the new basis, we apply the inverse transformation and get $X^{-1}AXb$.

The result: the matrix of the linear transformation L in the basis $x_1, \ldots, x_n$ equals $X^{-1}AX$.

For a non-singular matrix X, matrices A and $X^{-1}AX$ are called *similar*. Intuitively, they can be thought as representations of the same linear transformation in two different bases, with the basis transformation given by X.

Theorem 7.2.1. *If X is non-singular, then A and $X^{-1}AX$ have the same characteristic polynomial, eigenvalues, and algebraic and geometric multiplicities.*

In particular we see that all of these quantities are properties of the linear transformation represented by A rather than of the matrix itself. They remain the same in every basis.

Proof. First we show that the characteristic polynomials are the same, by using properties of the determinant:

$$\begin{aligned} p_{X^{-1}AX}(z) &= \det\left(zI - X^{-1}AX\right) = \det\left(X^{-1}(zI - A)X\right) \\ &= \det\left(X^{-1}\right)\det(zI - A)\det\left(X\right) \\ &= \det(zI - A) = p_A(z). \end{aligned}$$

The agreement of the characteristic polynomials implies that the eigenvalues and its algebraic multiplicities are the same for A and $X^{-1}AX$.

In order to show that the geometric multiplicities agree, it is easy to check that if E_λ is an eigenspace for A, then $X^{-1}E_\lambda$ is an eigenspace for $X^{-1}AX$ corresponding to eigenvalue λ, and conversely. In addition, these subspaces are bijectively mapped on each other so they have the same dimension. □

Theorem 7.2.2. *The algebraic multiplicity of an eigenvalue λ is at least as great as its geometric multiplicity.*

Proof. Let n be the geometric multiplicity of λ for matrix A, and let $\widehat{V}$ be an $m \times n$ matrix with the columns that form an orthonormal basis for E_λ. Then $A\widehat{V} = \lambda\widehat{V}$. (This simply says that every column of $\widehat{V}$ is in E_λ so the multiplication by A acts as a multiplication by the eigenvalue λ on this column.)

Let us extend $\widehat{V}$ to a square unitary matrix V. Then it is easy to check that

$$B = V^*AV = \begin{bmatrix} \lambda I_{n\times n} & C \\ 0 & D \end{bmatrix},$$

where C is $n \times (m-n)$ and D is $(m-n) \times (m-n)$. Note that B is similar to A. We calculate by using the definition of the determinant:

$$\begin{aligned} \det(zI - B) &= \det(zI - \lambda I)\det(zI - D) \\ &= (z-\lambda)^n \det(zI - D). \end{aligned}$$

Therefore the algebraic multiplicity of λ as eigenvalue of B is at least n. Since A is similar to B, it has the same algebraic multiplicity for λ, and so the algebraic multiplicity of λ in A is no less than n, its geometric multiplicity. □

7.3 More on diagonalizability

Recall that the matrix A is diagonalizable if and only if it can be written as $X\Lambda X^{-1}$ where Λ is a diagonal matrix. When a matrix is diagonalizable? In principle, the answer is given by the following theorem.

Theorem 7.3.1. *An $m \times m$ matrix A is diagonalizable if and only if it is non-defective.*

That is, when we say that a matrix is diagonalizable or that a matrix is non-defective, we describe the same property of matrices.

Sketch of the proof of Theorem 7.3.1. If matrix A is diagonalizable then $A = X\Lambda X^{-1}$, so it is similar to a diagonal matrix Λ and hence has same eigenvalues with same multiplicities. It is easy to check that a diagonal matrix is non-defective, so Λ is non-defective and the same holds for A.

In the converse direction, assume that the matrix A is non-defective. Then the dimension of each eigenspace equals to the algebraic multiplicity of the corresponding eigenvalue. Hence the sum of the dimensions of these

eigenspaces equal to m. If we choose a basis in each of these eigenspaces, and combine the bases together then we obtain the set of m linearly independent eigenvectors [This is the place where a more accurate argument is needed.] If these m independent eigenvectors are formed into the columns of a matrix X, then X is nonsingular and we have $A = X\Lambda X^{-1}$. □

However, it might be not easy to check if an eigenvalue is defective. A simpler sufficient condition for diagonalizability is that all eigenvalues are distinct.

Theorem 7.3.2. *If all eigenvalues of matrix A are distinct then A is diagonalizable.*

Proof. It is enough to show that the set of eigenvectors $x_1, \dots x_n$ corresponding to the eigenvalues $\lambda_1, \dots \lambda_n$ is linearly independent. Suppose not. Moreover, let us choose a set of eigenvectors such that they are linearly dependent but any proper subset of them is not linearly dependent. Without loss of generality, we can assume that this is the set $\{x_1, \dots, x_k\}$ where $k \leq n$. Note that $k \geq 2$ because $x_1 \neq 0$ so the set that consists of one eigenvector is not linearly dependent. Without loss of generality, we can rename the eigenvectors so that

$$x_1 = \sum_{i=2}^{n} c_i x_i,$$

and not all $c_i = 0$ because an eigenvector $x_1 \neq 0$. Then $Ax_1 = \sum_{i=2}^{n} c_i A x_i$, which means that

$$\lambda_1 x_1 = \sum_{i=2}^{n} c_i \lambda_i x_i$$

If we multiply the first equation by λ_1 and subtract it from the second equation, then we get

$$0 = \sum_{i=2}^{n} c_i (\lambda_i - \lambda_1) x_i$$

Since $\lambda_i \neq \lambda_1$ we see that not all coefficients in this sum are zero. Therefore, the set of vectors $\{x_2, \dots, x_k\}$ is linearly dependent, contrary to our assumption about the minimality of the set $\{x_1, x_2, \dots, x_k\}$. This contradiction shows that $\{x_1, \dots, x_n\}$ is linearly independent and therefore we can diagonalize matrix A. □

What happens if a matrix is non-diagonalizable? In this case it is possible to show that there is a matrix X such that $X^{-1}AX$ has a Jordan form. In this form the matrix is block-diagonal and every block has the form

$$B = \begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ \dots & \dots & \dots & \dots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{bmatrix}$$

(The block can be 1×1 with only λ inside it.)

7.4 The determinant and trace of A and eigenvalues

Theorem 7.4.1. *The determinant and the trace of a matrix A are equal to the product and the sum of the eigenvalues of A, respectively, counted with their algebraic multiplicities.*

Proof. We have already proved the statement about the determinant for diagonalizable matrices in a previous lecture. In general, we set $z = 0$ in the definition of the characteristic polynomial and obtain the required formula.

For the trace recall that the trace equals to the sum of diagonal elements of the matrix. From the definition of the determinant we see that in the expansion of $\det(zI - A)$ in powers of z the coefficient before z^{m-1} is $-\operatorname{tr}(A)$. (Indeed, in order to ensure that we have $m-1$ variables z in one of the determinant products, we need to take z from every diagonal element of the matrix $zI - A$ except one. This forces the last choice to be a $-a_{ii}$ from the remaining diagonal element. After summing over i, we obtain $-\operatorname{tr} A$.) On the other hand expanding $(z - \lambda_1) \dots (z - \lambda_m)$, we find that this coefficient is $-\sum_{i=1}^{m} \lambda_i$. This completes the proof.

□

7.5 Functions of matrices

If a matrix is not defective, then we have enough linearly independent eigenvectors $x_1, \dots, x_n$ to build a full-rank square matrix $X = [x_1, x_2, \dots, x_n]$. Then we have the diagonalization formula:

$$A = X \Lambda X^{-1},$$

where Λ is the diagonal matrix with the eigenvalues on the main diagonal.

This formula can be useful to compute functions of matrix A. For example if we want to calculate the k-th power of matrix A, A^k, then this formula gives us:

$$A^k = X\Lambda^k X^{-1} = X \begin{bmatrix} \lambda_1^k & 0 & \dots & 0 \\ 0 & \lambda_2^k & \dots & 0 \\ 0 & 0 & \dots & \lambda_n^k \end{bmatrix} X^{-1}$$

Similarly, if we have a polynomial $p(z) = \sum_{k=0}^{K} c_k z^k$, then

$$p(A) := \sum_{k=0}^{K} c_k A^k = X p(\Lambda) X^{-1} = X \begin{bmatrix} p(\lambda_1) & 0 & \dots & 0 \\ 0 & p(\lambda_2) & \dots & 0 \\ 0 & 0 & \dots & p(\lambda_n) \end{bmatrix} X^{-1}$$

More generally, this formula is valid for convergent power series and for all functions that can be written as convergent power series. For example,

$$e^A := \sum_{k=0}^{K} \frac{1}{k!} A^k = X e^{\Lambda} X^{-1} = X \begin{bmatrix} e_1^{\lambda} & 0 & \dots & 0 \\ 0 & e_2^{\lambda} & \dots & 0 \\ 0 & 0 & \dots & e_n^{\lambda} \end{bmatrix} X^{-1}$$

Example 7.5.1. Let

$$A = \begin{bmatrix} 4 & 3 \\ 1 & 2 \end{bmatrix}$$

Find A^{2022} by diagonalizing A.

The characteristic polynomial is $p(z) = (z-4)(z-2) - 3 = z^2 - 6z + 5 = (z-1)(z-5)$. So, the eigenvalues are 1 and 5 and the corresponding eigenvectors are $[1, -1]^t$ and $[3, 1]^t$. So the diagonalization is

$$A = X \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix} X^{-1},$$

where

$$X = \begin{bmatrix} 3 & 1 \\ 1 & -1 \end{bmatrix} \text{ and } X^{-1} = -\frac{1}{4} \begin{bmatrix} -1 & -1 \\ -1 & 3 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 & 1 \\ 1 & -3 \end{bmatrix}$$

It follows that

$$\begin{aligned} A^{2022} &= \frac{1}{4}\begin{bmatrix}3 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}5^{2022} & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & -3\end{bmatrix} \\ &= \frac{1}{4}\begin{bmatrix}3 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}5^{2022} & 5^{2022}\\ 1 & -3\end{bmatrix} = \frac{1}{4}\begin{bmatrix}3\times 5^{2022}+1 & 3\times 5^{2022}-3\\ 5^{2022}-1 & 5^{2022}+3\end{bmatrix} \\ &\approx 5^{2022}\frac{1}{4}\begin{bmatrix}3 & 3\\ 1 & 1\end{bmatrix}. \end{aligned}$$

Note that we got a matrix with the columns which are very close to a multiple of the eigenvector of the largest eigenvalue 5. This observations generalizes to other matrices (under some condition) and can be used to find the largest eigenvalue of a matrix. (Or rather, the eigenvalue with the largest absolute value.)

7.6 Unitary diagonalization; Schur decomposition

So far, we looked at the diagonalization matrices X which are simply invertible. One of the most important properties of Hermitian matrices is that they are diagonalizable and moreover, they admit a unitary diagonalization:

$$A = Q\Lambda Q^*,$$

where Λ is diagonal matrix with real entries and Q is a unitary matrix. (If A is a real symmetric matrix, then Q is an orthogonal matrix.)

In the case when all eigenvalues are different this is easy to prove. First, let us prove that all eigenvalues are real.

(a) Let λ be an eigenvalue of A with eigenvector x. Then,

$$(x, Ax) = \lambda x^* x = \lambda \|x\|^2.$$

At the same time, by using $A = A^*$,

$$(x, Ax) = (A^* x, x) = (Ax, x) = (\lambda x)^* x = \overline{\lambda}\|x\|^2.$$

Since x is non-zero vector, $\|x\| \neq 0$, so we find that $\lambda = \overline{\lambda}$ and that means that λ is real.

Then, let us prove that all eigenvectors are orthogonal.

(b) Let $\lambda_1 \neq \lambda_2$ are eigenvalues of A corresponding to eigenvectors x_1 and x_2, respectively. Then,

$$(x_2, Ax_1) = \lambda_1 (x_2, x_1).$$

On the other hand,

$$(x_2, Ax_1) = (A^*x_2, x_1) = (Ax_2, x_1) = \overline{\lambda}_2(x_2, x_1).$$

By using the fact that we proved in (a), λ_2 is real: $\overline{\lambda}_2 = \lambda_2$, so we can write:

$$\begin{aligned} \lambda_1(x_2, x_1) &= \lambda_2(x_2, x_1), \\ (\lambda_1 - \lambda_2)(x_2, x_1) &= 0. \end{aligned}$$

Since we assumed $\lambda_1 \neq \lambda_2$, hence $(x_2, x_1) = 0$ and so the eigenvectors are orthogonal to each other.

So, if all eigenvalues are different, then we can find a basis that consist of orthonormal system of eigenvalues. What if they are not? In general, we can prove the unitary diagonalization of Hermitian matrices by proving the existence of a so-called Schur factorization of an arbitrary square matrix.

A *Schur* factorization of a matrix A is a factorization $A = QTQ^*$, where Q is unitary and T is upper-triangular.

Theorem 7.6.1. *Every matrix A has a Schur factorization.*

Remark: Moreover, if matrix A is real and all its eigenvalues are real then it is possible to choose Q and T to be real in this factorization.

Proof. The proof is by induction on the dimension m of A. Suppose $m \geq 2$. Every matrix A has at least one eigenvalue λ by one of our previous results. Let x be a unit eigenvector belonging to λ and set it as a first column of a unitary matrix U. Then, we can check that

$$U^*AU = \begin{bmatrix} \lambda & w^* \\ 0 & B \end{bmatrix}.$$

By inductive hypothesis, there exists a Schur factorization VTV^* of B. Then, we can set

$$Q = U \begin{bmatrix} 1 & 0 \\ 0 & V \end{bmatrix},$$

and check that

$$Q^*AQ = \begin{bmatrix} \lambda & w^*V \\ 0 & T \end{bmatrix},$$

which is the desired Schur factorization. □

Corollary 7.6.2. *If $A^* = A$, then A admits unitary diagonalization:*

$$A = Q\Lambda Q^*, \tag{7.1}$$

where Q is unitary and Λ is diagonal with real entries.

Remark: if A is real then by using the remark after the theorem about the Schur diagonalization, we can show that Q can be chosen real.

The formula (7.1) is often written in the following form:

$$A = \sum_{i=1}^{n} \lambda_i q_i q_i^*,$$

where λ_i are eigenvalues of A and $\{q_i\}$ is an orthonormal basis of eigenvectors.

Simultaneous diagonalization Two Hermitian matrices A and B are called *simultaneously diagonalizable* if we can find a unitary matrix U such that

$$A = U\Lambda_A U^*,$$
$$B = U\Lambda_B U^*,$$

where Λ_A and Λ_B are the diagonal matrices with eigenvalues of A and B, respectively, on the main diagonal.

Theorem 7.6.3. *Hermitian matrices A and B are simultaneously diagonalizable if and only they commute, that is, if $AB = BA$.*

Proof. Let us look at the simple case when all eigenvalues of matrices A and B are distinct. Suppose x is an eigenvector of A with eigenvalue λ. Then $A(Bx) = BAx = \lambda Bx$, so Bx is also an eigenvector of A with the same eigenvalue (or zero). We assumed that all eigenvalues of A are simple, so Bx must be proportional to x and so x is also an eigenvector of B. This implies that we can take the matrix of (normalized) eigenvectors of A as U.

The other case, in which eigenvalues can have multiplicity greater than 1, is more complicated and we omit the proof. □

Some other classes of matrices also admit unitary diagonalization. The general criteria is that a square matrix A admits unitary diagonalization if and only if $A^*A = AA^*$. Such matrices are called *normal*. The proof of this fact follows from the Schur decomposition theorem by checking that for an upper-triangular matrix T, $T^*T = TT^*$ can hold only if T is diagonal.

One example of normal matrices which are not symmetric are unitary matrices.

7.7 Applications

7.7.1 Difference equations

Reading: Section 5.3 in Strang's book

A one-dimensional difference equation has the form

$$x_n = c_1 x_{n-1} + c_2 x_{n-2} + \ldots + c_k x_{n-k}$$

Here x_n is a sequence of numbers. We are given some initial conditions $x_{k-1}, x_{k-2} \ldots x_0$ or $x_0, x_{-1}, \ldots, x_{-(k-1)}$ and look to find what is the behavior of x_n for large n.

This equation can be written as the matrix equation if we introduce k-vectors $x^{(n)} = [x_n, x_{n-1}, \ldots, x_{n-k+1}]^*$ and matrix

$$A = \begin{bmatrix} c_1 & c_2 & \ldots & c_k \\ 1 & 0 & 0 \ldots 0 & 0 \\ 0 & 1 & 0 \ldots 0 & 0 \\ 0 & 0 & 1 \ldots 0 & 0 \\ 0 & 0 & 0 \ldots 1 & 0 \end{bmatrix}$$

Then we can write the difference equation in the form

$$x^{(n)} = A x^{(n-1)}. \tag{7.2}$$

The solution of this equation is $x^{(s)} = A^s x^{(0)}$. Hence if we want to know the behavior of the sequence x_n for large n we need to know the behavior of powers of the matrix A^s.

If we can diagonalize the matrix A then we have

$$\begin{aligned} A &= X \Lambda X^{-1}, \\ A^s &= X \Lambda^s X^{-1} \end{aligned} \tag{7.3}$$

If we know both Λ and the matrix of eigenvectors X we can write an explicit formula for $x^{(n)}$. In fact, we often don't need to calculate the matrix X because formula (7.3) implies that we can write the solution as

$$x_n = \sum_{i=1}^{k} a_i \lambda_i^n, \tag{7.4}$$

where λ_i are eigenvalues of matrix A and a_i are some coefficients which can be calculated from the initial conditions.

In addition, this formula often allows us to find the asymptotic behavior of x_n. Suppose λ_1 is an eigenvalue of A that has the largest absolute value: $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_k|$. If in addition, we assume that $a_1 \neq 0$, then we have $x_n \sim a_1\lambda_1^n$. In particular, if $|\lambda_1| < 1$ then the sequence declines to zero, and if $|\lambda_1| > 1$ then the sequence grows unboundedly.

Remark: It can be proved that the characteristic polynomial of A is

$$p_A(z) = z^k - c_1 z^{k-1} - \ldots - c_{k-1}z - c_k, \tag{7.5}$$

so the eigenvalues of A are roots of this polynomial, and our method is found to be equivalent to a popular method of solving difference equations. Namely, solve the characteristic equation (7.5) and then find the coefficients in (7.4) from the initial conditions.

Many other dynamic problems in biology, engineering and physics can be cast in the form (7.2) with $x^{(k)}$ that describe the state of a system at time k, and A that describe the evolution of the state. In this case, the stability of the system depends on the size of the eigenvalue with the largest absolute value.

Example 7.7.1 (Fibonacci numbers). A classic example for this concept is the Fibonacci numbers, which are defined by the relation:

$$F_n = F_{n-1} + F_{n-2}.$$

and the initial condition $F_1 = F_2 = 1$. Then we can define vector $\boldsymbol{f}_n = (F_{n+1}, F_n)^t$, with $\boldsymbol{f}_0 = (1, 0)^t$, and

$$\begin{aligned} A &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{bmatrix}^{-1}, \end{aligned}$$

where $\lambda_1 = (1 + \sqrt{5})/2$ and $\lambda_2 = (1 - \sqrt{5})/2$ are eigenvalues of matrix A. Then,

$$A^n = \begin{bmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{bmatrix} \begin{bmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{bmatrix}^{-1},$$

One can use this formula to calculate F_n. Alternatively, we can note that this formula implies that $F_n = a\lambda_1^n + b\lambda_2^n$, where a and b are some coefficients that do not depend on n. We can find a and b from equations $F_0 = a + b$

and $F_1 = a\lambda_1 + b\lambda_2$. The advantage of this method is that we do not need to calculate the matrix of eigenvectors X and the inverse matrix X^{-1}.

After some calculation, we can get:

$$F_n = \frac{1}{\sqrt{5}}\Big(\lambda_1^n - \lambda_2^n\Big).$$

Since $|\lambda_1| > |\lambda_2|$ we find that

$$F_n \sim \frac{1}{\sqrt{5}}\Big(\frac{1+\sqrt{5}}{2}\Big)^n$$

For example, $F_{30} = 832,040$ and the right hand side is $832,040 + 2.4063 \times 10^{-7}$.

7.7.2 Linear Differential Equations

See Strang 5.4.

If we have a system of linear differential equations $x'(t) = Ax(t)$, where x is a vector with k components, when diagonalization of A decouples this system of equation. We get the formula: $x'(t) = S\Lambda S^{-1}x(t)$, where Λ is the diagonal matrix with eigenvalues of A on the main diagonal.

Let $u(t) = S^{-1}x(t)$. Each equation in the resulting system has the form $u_i'(t) = \lambda_i u_i(t)$ so its solution is $u_i(t) = e^{\lambda_i t}u_i(0)$. In vector form: $u(t) = e^{\Lambda t}u(0)$.

Therefore, the solution of the original system is

$$x(t) = Se^{\Lambda t}S^{-1}x(0) = e^{At}x(0).$$

The order equations of higher order can be solved by a similar approach by first converting them to a system of equations.

For example, if we have a system:

$$y'' = c_1 y' + c_2 y,$$

then we can convert it to a system by setting $x_1(t) = y(t)$ and $x_2(t) = y'(t)$. Then we have

$$\begin{aligned} x_1'(t) &= x_2(t) \\ x_2'(t) &= c_2 x_1(t) + c_1 x_2(t), \end{aligned}$$

or in matrix form

$$x'(t) = \begin{bmatrix} 0 & 1 \\ c_2 & c_1 \end{bmatrix} x(t)$$

It turns out that diagonalization of the matrix A in this case equivalent to the standard method of solving these equations: find the roots of the polynomial $z^2 - c_1 z - c_2 = 0$. If the roots are λ_1 and λ_2 then the general solution is

$$y(t) = a_1 e^{\lambda_1 t} + a_2 e^{\lambda_2 t},$$

and the coefficients a_1 and a_2 can be found by fitting the initial conditions $y(0)$ and $y(1)$.

7.7.3 Markov Chains

Transition probabilities

Reading:

- This section relies heavily on the book "Markov Chains" by James Norris. In particular, Sections 1.1 and 1.7.
- Section 5.3 in Strang's book

A *distribution* μ on a finite state space S is a non-zero $|S| \times 1$ vector with non-negative entries . We call it a *probability distribution* if the sum of the vector entries is 1.

The interpretation of the component μ_x is that it is a probability to find a random system X in a state x.

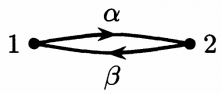
A Markov chain is a model that describes how the probability distribution μ evolves through time. Let us explain this in detail.

Let S be a finite set, and X_n, $n \geq 0$, be a sequence of random variables that take values in the state space S. (We will often identify S with a subset of integers $\{1, \ldots, m\}$.) We interpret X_n as the (random) state of a system X at time n.

We say that X_n is a *discrete-time Markov chain* with the initial probability distribution μ on S, and transition matrix P if

1. $\mathbb{P}(X_0 = x) = \mu_x$;
2. $\mathbb{P}(X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n) = P_{x_n, x_{n+1}}$.

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$



$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{pmatrix}$$

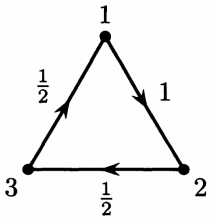


Figure 7.1

This is illustrated by diagrams in Figure 7.1.

The transition matrix P is a square $m \times m$ matrix. It has the properties that all its entries are non-negative and the sum of the entries in every row equals to 1. Such matrices are called *stochastic* matrices.

Example 7.7.2 (Random walk on a graph). Recall that an *(undirected) graph* $G = (V, E)$ is a set of vertices V and a set of edges E, which are simply unordered pairs of vertices: $E \subset V \times V/(v_1, v_2) \sim (v_2, v_1)$. A graph is *simple* if there are no multiple edges (edges with the same endpoints) and that there are no loops, i.e., edges that have the same vertex as both endpoints. A degree of a vertex v, denoted $d(v)$, is the number of edges which are incident to v, that is, that have v as one of its endpoints.

Now we can define a Markov chain which is called a *simple random walk* on G. The states are vertices and the transition probability $P_{uv} = 1/d(u)$. The interpretation is that if there is a particle at vertex u, it has equal probabilities move along each of the edges incident to u.

If we know the initial probability distribution μ_0, we can calculate the distribution at later times of a Markov chain by multiplying the distribution μ_0 by the transition matrix P *on the right*. Indeed, for every sequence of states, $(x_0, \ldots, x_n)$, we can calculate the probability that the system will go through this sequence of states as follows:

$$\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) = \mu_{x_0} P_{x_0,x_1} P_{x_1,x_2} \ldots P_{x_{n-1},x_n}.$$

In particular if we sum over all $x_0, \ldots x_{n-1}$, we will find the marginal distribution of X_n,

$$\mathbb{P}(X_n = x_n) = (\mu P^n)_{x_n}.$$

Here P^n is the n-th power of the matrix P, μP^n denote the product of vector μ by matrix P^n, and $(\mu P^n)_j$ is the j-th component of this product.

In particular, if we assume that $\mu = e_x$ (the initial probability vector is 1 for the state x and zero for all other states), then follows that

$$\mathbb{P}(X_n = y | X_0 = x) = (P^n)_{xy}.$$

That is, the probability that at time n the chain is at the state y conditional that originally it was at state x is the xy-th entry of the matrix P^n (i.e., the entry in the row x and column y).

We will often write the conditional probabilities $\mathbb{P}(A|X_0 = x)$ as $\mathbb{P}_x(A)$, so, for example, the previous result is $\mathbb{P}_x(X_n = y) = (P^n)_{xy}$.

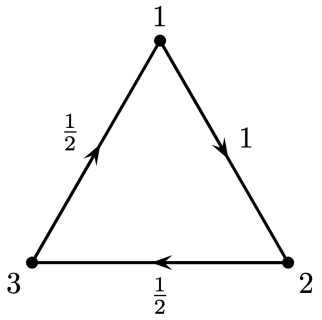


Figure 7.2

Example 7.7.3. Consider the three-state chain with diagram in Figure 7.2. The transition matrix is

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}.$$

We seek to find a general formula for the probability that a particle that starts at 1 will be at 1 after n steps.

This means that we want to calculate $(P^n)_{11}$.

By writing down the characteristic equations, we can find the eigenvalues.

$$\det(zI - P) = \frac{1}{4}(x - 1)(4x^2 + 1) = 0.$$

Therefore the eigenvalues are $1, i/2, -i/2$ and the matrix P is diagonalizable:

$$P = X \begin{bmatrix} 1 & 0 & 0 \\ 0 & i/2 & 0 \\ 0 & 0 & -i/2 \end{bmatrix} X^{-1}.$$

Then,

$$P^n = X \begin{bmatrix} 1 & 0 & 0 \\ 0 & (i/2)^n & 0 \\ 0 & 0 & (-i/2)^n \end{bmatrix} X^{-1}.$$

From this we can conclude that

$$P_{11}^n = a + b\Big(\frac{i}{2}\Big)^n + c\Big(\frac{-i}{2}\Big)^n,$$

where a, b, and c are some constants. These constants can be found from the cases $n = 0, 1, 2$:

$$\begin{aligned} P_{11}^0 &= 1 = a + b + c, \\ P_{11}^1 &= 0 = a + i(b - c)/2, \\ P_{11}^2 &= 0 = a - (b + c)/4, \end{aligned}$$

From the first and the third equations $1 = 5a$ and so $a = 1/5$. So, $b - c = 2i/5$ and $b + c = 4/5$ which gives $b = 2/5 + i/5$ and $c = 2/5 - i/5$. Then,

$$P_{11}^n = \frac{1}{5} + \frac{2+i}{5}\Big(\frac{i}{2}\Big)^n + \frac{2-i}{5}\Big(\frac{-i}{2}\Big)^n.$$

Alternatively, if we want to avoid complex-valued expressions in the final answer, then we note that

$$\Big(\pm\frac{i}{2}\Big)^n = \frac{1}{2^n}e^{\pm n\frac{i\pi}{2}} = \frac{1}{2^n}\Big(\cos\frac{n\pi}{2} \pm i\sin\frac{n\pi}{2}\Big)$$

and so it make sense to search for P_{11}^n in the form:

$$P_{11}^n = \alpha + \frac{1}{2^n}\Big(\beta\cos\frac{n\pi}{2} + \gamma\sin\frac{n\pi}{2}\Big).$$

Then, from initial conditions one can find that $\alpha = 1/5$, $\beta = 4/5$ and $\gamma = -2/5$ and so

$$P_{11}^n = \frac{1}{5} + \frac{1}{2^n}\Big(\frac{4}{5}\cos\frac{n\pi}{2} - \frac{2}{5}\sin\frac{n\pi}{2}\Big).$$

Invariant Distribution

If P is the transition matrix of a Markov chain then a probability distribution π is called *invariant* if

$$\pi P = \pi.$$

The terms *equilibrium* or *stationary* distribution are also used.

The definition of the invariant distribution implies that if X_n is distributed according to π then X_{n+1} is also be distributed according to π.

From an algebraic viewpoint an invariant measure is a *left* eigenvector of the matrix P with eigenvalue 1.

In general, a left eigenvector of a matrix A is a non-zero vector x such that $xA = \lambda x$. The corresponding λ is a left eigenvalue. However, it is easy to see that the left eigenvalues of A are the same as the usual eigenvalues. Indeed, they are solutions to the equation $\det(zI - A^t) = \det\left((zI - A)^t\right) = 0$, which is the same as the equation $\det(zI - A) = 0$ by a property of the determinant.

In contrast, the left *eigenvectors* are typically different from right eigenvectors if A is not symmetric.

One obvious property of the stochastic matrices is that they always have an eigenvalue $\lambda = 1$, which corresponds to the (right) eigenvector $v = [1, 1, \dots 1]$. So we know that there exists also a left eigenvector with $\lambda = 1$ and this gives us a practical method for computation of the invariant distribution if the state space is finite.

Example 7.7.4. Find the invariant distribution for the Markov chain from Example 7.7.3.

The matrix is

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}.$$

In order to find the left eigenvector corresponding to the eigenvalue $\lambda = 1$, we need to find the null-space of the matrix

$$P^t - I = \begin{bmatrix} -1 & 0 & 1/2 \\ 1 & -1/2 & 0 \\ 0 & 1/2 & -1/2 \end{bmatrix}$$

Since the dimension of the nullspace is 1 in this example (all eigenvalues are distinct), the row 3 is a linear combination of rows 1 and 2, which

are linearly independent. In addition, we have the normalization condition $\pi_1 + \pi_2 + \pi_3 = 1$. Hence in order to find the invariant distribution we need to solve the system:

$$\begin{bmatrix} -1 & 0 & 1/2 \\ 1 & -1/2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

We solved by the Gaussian elimination method:

$$\begin{bmatrix} -1 & 0 & 1/2 & | & 0 \\ 1 & -1/2 & 0 & | & 0 \\ 1 & 1 & 1 & | & 1 \end{bmatrix} \to \begin{bmatrix} -1 & 0 & 1/2 & | & 0 \\ 0 & -1 & 1 & | & 0 \\ 0 & 1 & 3/2 & | & 1 \end{bmatrix} \to \begin{bmatrix} -1 & 0 & 1/2 & | & 0 \\ 0 & -1 & 1 & | & 0 \\ 0 & 0 & 5/2 & | & 1 \end{bmatrix}$$
$$\to \begin{bmatrix} -1 & 0 & 1/2 & | & 0 \\ 0 & -1 & 1 & | & 0 \\ 0 & 0 & 1 & | & 2/5 \end{bmatrix} \to \begin{bmatrix} -1 & 0 & 0 & | & -1/5 \\ 0 & -1 & 0 & | & -2/5 \\ 0 & 0 & 1 & | & 2/5 \end{bmatrix}$$

So we find that the invariant distribution is $\pi_1 = 1/5$, $\pi_2 = 2/5$, and $\pi_3 = 2/5$.

Note from Example 7.7.3 that $P_{11}^n \to 1/5 = \pi_1$. This is a particular case of the general fact that under suitable assumptions about the Markov chain, for every i the transition probabilities P_{ij}^n converge to π_j as $n \to \infty$.

Now, let us ask the question about the existence of the invariant distribution π.

Theorem 7.7.5. *A finite-state Markov chain with transition matrix P always has an invariant distribution.*

Since 1 is an eigenvalue of P, therefore the left eigenvector with eigenvalue 1 exists. However, how do we know that it has non-negative entries?

It turns out that a stochastic matrix X always has a left eigenvector with eigenvalue 1 and with non-negative entries. The proof of this is not straightforward.

Is the invariant distribution unique? Not always.

One of the proofs of the existence and uniqueness of the invariant distribution is based on the Perron-Frobenius theorem. It holds not only for stochastic matrices but for a more general case of non-negative matrices with some additional restrictions. (I follow the book "Non-negative matrices" by Seneta here.)

The Perron-Frobenius theorem can be formulated in different degrees of generalities and we do it for an easier case that guarantees that the multiplicity of the eigenvalue $\lambda = 1$ is exactly 1 and the invariant distribution is unique.

Definition 7.7.6. A square non-negative matrix A is called primitive if for some positive integer k, all entries of the matrix A^k are positive.

Theorem 7.7.7. *Suppose A is a square non-negative primitive matrix. Then,*

1. *There exists a positive real eigenvalue λ such that it is strictly greater than the absolute value of any other eigenvalue.*
2. *The algebraic and geometric multiplicities of λ equal* 1.
3. *The left and right eigenvectors corresponding to λ are strictly positive.*

We refer to Seneta for the proof of this theorem. For stochastic primitive matrices, with some additional effort, it is possible to show that the largest eigenvalue equals 1. Hence, a consequence of the Perron-Frobenious theorem is that there exists a unique invariant distribution for every Markov Chain with primitive transition matrix.

Is a stochastic matrix always diagonalizable? No. There are examples when it is not, however we will see that a certain class of stochactic matrices (reversible matrices) is always diagonalizable.

The largest eigenvalue is always 1 and the other eigenvalues are also important. Consider the simple case when $\lambda_1 = 1 > \lambda_2 \geq \ldots \geq \lambda_n$, and assume that there is a basis that consists of left eigenvectors of P. (That is, P is diagonalizable). Then, if $\mu^{(0)}$ is the initial distribution then we can expand it in the basis:

$$\mu^{(0)} = c_1\pi + \sum_{k=2}^{n} c_k v_k,$$

where π is the invariant distribution and v_k are other left eigenvectors. Hence, the distribution at step t of the Markov Chain is

$$\mu^{(t)} = \mu^{(0)}P^t = c_1\pi + \sum_{k=2}^{n} \lambda_k^t c_k v_k.$$

This implies that

$$\lim_{t\to\infty} \mu^{(t)} = c_1\pi.$$

Since $\mu^{(t)}$ and π are probability distributions hence $c_1 = 1$ and we see that the distribution $\mu^{(t)}$ converges to the stationary distribution π. In addition, if λ_2 is the eigenvalue that has the second-largest absolute value, and if $c_2 \neq 0$, then

$$|\mu^{(t)} - \pi| \sim c_2 \|v_2\| |\lambda_2|^t.$$

That is, the speed of the convergence to the stationary distribution depends on the second largest eigenvalue λ_2.

Since the speed of the convergence to the stationary distribution is important in many applications of Markov Chains, it is often an important question if some good estimates of the second largest eigenvalue exist.

An example of application

Markov chains are often used to sample from a particular distribution on a state space. This method is called Markov Chain Monte Carlo simulation method (MCMC).

The idea is that if a distribution π on the state space is given then one can build a Markov Chain on this state space so that π is invariant distribution for this Markov Chain. Then one starts with arbitrary initial state and runs the chain for a sufficiently long time to ensure the convergence to the invariant distribution. The resulting state is considered to be a sample from the invariant distribution.

For this algorithm two important questions are relevant:

1. How can we build a Markov Chain on a state space with specified transitions and invariant distribution? That is, which transition probabilities should be assigned to these transitions, so that the invariant distribution of the chain equals to the target distribution?

2. Is it possible to give bounds on the time needed for convergence to invariant distribution?

For both questions, a special class of Markov chains called reversible Markov chains is useful. For the first question, it is often a relatively easy method to build the required chain.

For the second question, the theoretical results are unfortunately scarce. However, they are easier to come by for reversible chains because they have more structure imposed on their eigenvectors.

Example (Ising model on a finite graph).

Consider the Ising model on a graph G. It is useful to keep in mind a big finite sub-graph of $\mathbb{Z}^2$ where two vertices are connected if the distance between them is 1.

Then a state (or configuration) is the collection of $\pm$ spins assigned to each vertex of v. Formally, a state is a function $x : G \to \{\pm 1\}$. We write x_v to denote a spin at vertex v. One can also think about a state as a vector that has length $|G|$ and each element of this vector is either $+1$ or -1. So there are $2^|G|$ states. If G is large, this state space is enormous.

We can introduce a probability distribution on this state space:

$$\mathbb{P}(s = \{x_v\} = \frac{1}{Z} \exp(\beta H(s)),$$

where β is a parameter (inverse temperature), and Z is a normalizing constant. The function $H(s)$ (negative of the potential energy) is defined as follows:

$$H(s) = \sum_{u \sim v} x_u x_v + \sum_u \mu x_u.$$

Here the first summation is over all pairs of vertices, which are connected by an edge (denoted as $u \sim v$).

This distribution is called the Gibbs distribution.

One can see that the probability of a state s is larger if the spins at the neighboring vertices are aligned. In addition, the second term reflects the presence of a magnetic field: the probability of s is larger if the spins align along the external magnetic field.

One is often interested in sampling states from this system. The natural transitions between states is when a spin at a particular vertex is updated: it is either change to a negative or stays as it is.

A MCMC algorithm construct a Markov Chain such that the vertex v is chosen randomly and the probability of transition at vertex v only depends on the neighboring vertices. The transitions probabilities are chosen in such a way that the invariant distribution is the Gibbs distribution. Unfortunately, not much theoretical results are available for the speed of convergence.

Reversible Markov Chains

In MCMC the Markov chains often have the property which is called the reversibility. In this section we will discuss this property and some of its consequences.

We have seen previously that symmetric matrices have some additional useful properties. While a stochastic matrix is rarely symmetric, there is a subclass of stochastic matrices which shares many good properties of symmetric matrices.

A Markov chain with transition matrix P is called *reversible* if for some probability distribution μ and all states i, j.

$$\mu_j P_{ji} = \mu_i P_{ij} \tag{7.6}$$

In other words if we multiply rows of matrix P by numbers μ_i, $i = 1, \ldots, n$, respectively, then the resulting matrix will be symmetric.

These equations are called the *detailed balance equations*. The name is related to the fact that if initial distribution is the invariant distribution, then for reversible chain it is is not possible to distinguish statistically between sequences $X_0, \ldots, X_n$ and $X_n, \ldots, X_0$.

In terms of matrices, the detailed balance equations can be written as

$$DP = (DP)^t = P^t D, \tag{7.7}$$

where D is a real diagonal matrix with the entries $D_{ii} = \mu_i$. (We can write it as $diag(\mu)$.)

Example 7.7.8. Consider the matrix $P = \begin{bmatrix} 1/3 & 2/3 \\ 3/7 & 4/7 \end{bmatrix}$. Then if we multiply the first row by $\mu_1 = 3/7$ and the second row by $\mu_2 = 2/3$, we find that the resulting matrix is symmetric. This is not quite what we want since $(\mu_1, \mu_2) = (3/7, 2/3)$ is not a probability distribution. However, we can multiply this vector (μ_1, μ_2) by an appropriate constant (namely $(3/7 + 2/3)^{-1}$) to make sure that the result $(\widehat{\mu}_1, \widehat{\mu}_2)$ is a probability distribution. It is easy to see that $diag(\widehat{\mu}_1, \widehat{\mu}_2)P$ is still symmetric. Hence, P is reversible.

It turns out that the solution μ of the equation (7.6) is an invariant distribution.

Lemma 7.7.9. *If the probability distribution μ satisfy (7.6) , then μ is invariant.*

Proof. We need to check that $\mu P = \mu$. We write:

$$(\mu P)_i = \sum_j \mu_j P_{ji} = \sum_j \mu_i P_{ij} = \mu_i.$$

□

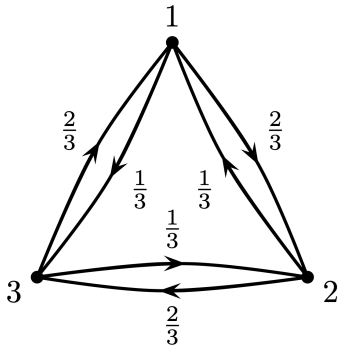


Figure 7.3: An example of a non-reversible chain: a random walk with a bias.

Often this property gives us a convenient tool for finding the invariant distribution of a chain. We can also use this property to give an example of a non-reversible chain.

Example 7.7.10. Consider a Markov chain with transition matrix

$$P = \begin{bmatrix} 0 & 2/3 & 1/3 \\ 1/3 & 0 & 2/3 \\ 2/3 & 1/3 & 0 \end{bmatrix}$$

The diagram for this chain is shown in 7.3. Clearly $\pi = (1/3, 1/3, 1/3)$ is invariant distribution, however matrix P is not symmetric. Hence, the Markov chain is not reversible.

(Intuitively, this chain is not reversible because even in the steady state an observer would be able to detect that the movement is mostly clock-wise, while after the time reversal, the movement would be counter-clockwise.)

Example 7.7.11 (Random walk on a graph). Consider a graph G with vertices $v \in V$. The *degree* d_v of a vertex v is the number of edges incident with v. A random walk on the graph G has the transition matrix P with entries $P_{uv} = 1/d_u$ if (u, v) is an edge, and $P_{uv} = 0$ otherwise. It is easy to check that P satisfies the detailed balance condition with $\mu_u = d_u$. It follows that the random walk is reversible with the invariant measure $\pi = d_u$.

If the graph G is not regular, that is, if it has vertices of differing degrees, then this invariant measure is not uniform. Vertices with larger degree will be visited more often than vertices with smaller degree. What if we want to have at our disposal a Markov chain on the graph G that would have the same transitions, – from a vertex to their neighbors, – but that would have a uniform distribution on vertices?

In this case, we can use a *lazy random walk*. Namely, suppose $d = \max\{d_1, \ldots, d_{|V|}\}$ is the maximum vertex degree in the graph. Then we set $P_{uv} = 1/d$ if (u, v) is an edge, and $P_{uu} = 1 - d_u/d$. In other words, if $d_u < d$ then with positive probability the particle will stay at vertex u and wait for the next time period. Since the resulting matrix P is symmetric, the uniform distribution is invariant for this chain.

Now we are going to justify our claim that the transition matrices for reversible chains have some properties which are similar to the properties of the symmetric matrices. Namely, they have an analogue of the orthogonal diagonalization which is the trademark of symmetric matrices.

Theorem 7.7.12. *The eigenvalues of the transition matrix P of a reversible Markov chain are real and P has the following factorization:*

$$P = D^{-1/2} Q \Lambda Q^t D^{1/2}.$$

where D is a diagonal matrix whose diagonal entries are the elements of the invariant distribution, and Q is an orthogonal matrix.

The fact that the eigenvalues of P are real for a system invariant to time-reversal is an important general fact. In addition, the decomposition stated in the theorem is useful in the analysis of properties of P.

Proof. The matrix form of the detailed balance equations (7.7) can be written as

$$\begin{aligned} D^{1/2} P D^{-1/2} &= D^{-1/2} P^t D^{1/2} \\ &= \left(D^{1/2} P D^{-1/2} \right)^t. \end{aligned}$$

In other words, the matrix

$$\widehat{P} = D^{1/2} P D^{-1/2}$$

is symmetric. Therefore, it has an orthogonal diagonalization $Q \Lambda Q^t$ and its eigenvalues are real. Since P is similar to $\widehat{P}$ its eigenvalues are also real and it has diagonalization

$$P = D^{-1/2} Q \Lambda Q^t D^{1/2}.$$

□

7.8 Exercises

Exercise 7.8.1. Consider the 2×2 matrix $A = \begin{bmatrix} 1 & 5 \\ -2 & 3 \end{bmatrix}$.

a. Calculate the eigenvalues of A.

b. If possible, construct matrices P and C such that $A = PCP^{-1}$, where C is diagonal.

Exercise 7.8.2. Let

$$A = \begin{bmatrix} 0 & 1 & -1 \\ 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix},$$

A has exactly two distinct eigenvalues, which are -2, and 1.

If possible, construct matrices P and D such that $A = PDP^t$, P is a matrix with orthonormal columns, and D is a diagonal matrix.

Exercise 7.8.3. a. If $A^2 = I$, what are possible eigenvalues of A?

b. If this A is 2×2 and not I or $-I$, find its trace and determinant.

c. If the first row of this matrix is $(3, -1)$, what is the second row?

Exercise 7.8.4. (a) A 2×2 matrix A satisfies $\text{tr}(A^2) = 5$ and $\text{tr}(A) = 3$ (where $\text{tr}(X)$ denotes the trace of X). Find $\det(A)$.

(b) A 2×2 matrix A has two proportional columns and $\text{tr}(A) = 5$. Find $\text{tr}(A^2)$.

(c) A 2×2 matrix A has $\det(A) = 5$ and positive integer eigenvalues. What is the trace of A?

Exercise 7.8.5. For each of the following statements, prove that it is true or give an example to show it is false. Throughout, A is a complex $m \times m$ matrix unless otherwise indicated.

a. If λ is an eigenvalue of A and $\mu \in \mathbb{C}$, then $\lambda - \mu$ is an eigenvalue of $A - \mu I$.

b. If A is real and λ is an eigenvalue of A, then so is $-\lambda$.

c. If A is real and λ is an eigenvalue of A, then so is $\overline{\lambda}$.

d. If λ is an eigenvalue of A and A is non-singular, then λ^{-1} is an eigenvalue of A^{-1}.

e. If all the eigenvalues of A are zero, then $A = 0$.

f. If A is diagonalizable and all its eigenvalues are equal, then A is diagonal.

g. If A is invertible and diagonalizable, then A^{-1} is diagonalizable.

h. Matrices A and A^t have the same eigenvalues.

Exercise 7.8.6. Suppose each "Gibonacci" number G_{k+2} is the average of the two previous numbers G_{k+1} and G_k. Then $G_{k+2} = \frac{1}{2}(G_{k+1} + G_k)$. In matrix form this can be written as

$$\begin{bmatrix} G_{k+2} \\ G_{k+1} \end{bmatrix} = A \begin{bmatrix} G_{k+1} \\ G_k \end{bmatrix}.$$

a. Find the eigenvalues and eigenvectors of A.

b. Find the limit of the matrices A^n as $n \to \infty$.

c. If $G_0 = 0$ and $G_1 = 1$, which number do the Gibonacci numbers approach?

Exercise 7.8.7. A flea hops about at random on the vertices of a triangle with all jumps equally likely. (So if the vertices are labeled 1, 2, 3 and the flea is at vertex 1 then it jumps to vertices 2 and 3 with probabilities 1/2 and 1/2, respectively.) Find the probability that after n hops the flea is back where it started.

A second flea also hops about on the vertices of a triangle, but this flea is twice as likely to jump clockwise as anti-clockwise. What is the probability that after n hops this second flea is back where it started. [Recall that $e^{\pm i\pi/6} = \sqrt{3}/2 \pm i/2$.]

Exercise 7.8.8. Let X_n, $n = 0, 1, \ldots$, be a Markov chain on $\{1, 2, 3\}$ with transition matrix

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 2/3 & 1/3 \\ 1/2 & 1/2 & 0 \end{bmatrix}.$$

Calculate the invariant distribution for this chain.

Exercise 7.8.9. In each of the following cases determine whether the stochastic matrix P is reversible:

1.

$$\begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix};$$

$(0 < p < 1$ and $0 < q < 1.)$

2.

$$\begin{bmatrix} 0 & p & 1-p \\ 1-p & 0 & p \\ p & 1-p & 0 \end{bmatrix};$$

$(0 < p < 1)$

3. The state space is $\{0, 1, \ldots, N\}$ and $p_{ij} = 0$ if only if $|j - i| \geq 2$.

Exercise 7.8.10. (*) The *spectral radius* $\rho(A)$ of a square matrix A is the largest absolute value of an eigenvalue λ of A.

For an arbitrary $m \times m$ complex matrix A and the operator norm $\|\cdot\|$, prove using the Schur decomposition:

$$\lim_{n\to\infty} \|A^n\| = 0 \text{ if and only if } \rho(A) < 1.$$

7.9 Appendix: Complex numbers

Complex number is a pair of real numbers (x, y). So, it is essentially a vector in $\mathbb{R}^2$. The addition of complex numbers is the addition of vectors. However, the wonderful fact is that there is also a multiplication operation. This operation has no analogue for vectors in $\mathbb{R}^n$ for general n.

It is easier to remember this operation, if we write complex numbers as $z = x + iy$, in which case the product of two complex number is defined as $z_1 z_2 := x_1 x_2 - y_1 y_2 + i(x_1 y_2 + x_2 y_1)$. This is the same as if we thought about i as a special kind of number with property $i^2 = -1$.

It turns out that this operation is associative and commutative and satisfies the distributive law with respect to the addition.

Formally, we converted the linear space $\mathbb{R}^2$ to a commutative algebra over $\mathbb{R}$.

One important new operation is that of conjugation: $\overline{x + iy} = x - iy$. Note that $z\overline{z} = x^2 + y^2 = \|z\|^2$. In the context of complex numbers, $\|z\|$ is called the *absolute value*, or the *modulus* of z.

Since we have multiplication and addition, we can define polynomials and power series using complex numbers. The convergence for series is defined

using the norm in $\mathbb{R}^2$. In particular, we can define the exponential function e^z as the convergent series $\sum_{k=0}^{\infty} z^k/k! = 1 + z + z^2/2! + \ldots$. This function preserves the important property of the standard exponential function:

$$e^{z_1+z_2} = e^{z_1}e^{z_2}.$$

In particular, $e^{x+iy} = e^x e^{iy}$. In addition, directly from definition of e^z, we can obtain

$$e^{iy} = \cos y + i \sin y.$$

Therefore,

$$e^{x+iy} = e^x(\cos y + i \sin y).$$

This formula allows us to give a geometric meaning to the product operation. For this we need to represent the vector $z = (x, y)$ in polar coordinates as $(r\cos\alpha, r\sin\alpha)$. Here $r = \sqrt{x^2 + y^2} = \|z\|$, and α is called the *argument* of z.

Then, $z = x + iy = e^{\log r + i\alpha}$. If we have another complex number $z' = x' + y' = e^{\log r' + i\alpha'}$ then the addition formula for the exponential gives us:

$$zz' = e^{\log r + \log r' + i(\alpha+\alpha')} = rr'e^{i(\alpha+\alpha')}.$$

In other words, when we multiply z and z', their absolute values are multiplied, and their arguments are added.

The fundamental theorem of algebra says that every equation of degree n (with coefficient in real or complex numbers) has at least one solution in complex numbers. This can be strengthened to the statement that it has exactly n solutions if we count the solutions with multiplicities. The proof of this remarkable theorem is quite non-trivial.

Chapter 8

Bilinear and Quadratic Forms

Reading for this Chapter

- Strang: Chapter 6.

8.1 Definitions and diagonalization

A bilinear form is a map that sends a pair of vectors to a number, $B : \mathbb{R}^n \times \mathbb{R}^n \to R$. This map is required to be linear in both arguments:

$$B(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1 B(x_1, y) + \alpha_2 B(x_2, y)$$
$$B(x, \alpha_1 y_1 + \alpha_2 y_2) = \alpha_1 B(x, y_1) + \alpha_2 B(x, y_2)$$

A bilinear symmetric form has an additional property $B(x, y) = B(y, x)$.

Remark: the bilinear forms can also be defined for complex numbers, however, a more useful concept in that setting is the concept of Hermitian forms, when $B(\lambda x, y) = \overline{\lambda} B(x, y)$, $B(x, y) = \overline{B(y, x)}$ which implies $B(x, \lambda y) = \lambda B(x, y)$. (So, the Hermitian form is linear in the second argument and "conjugate-linear" or antilinear in the first argument. In the following, for concretenes we focus the discussion on bilinear forms.

For a bilinear form, one can define a quadratic form $Q(x) = B(x, x)$. Conversely, if we are given a quadratic form $Q(x)$ then we can define a symmetric bilinear form as $B(x, y) = [Q(x + y) - Q(x - y)]/4$.

To every bilinear form $B(x, y)$ we can associate a matrix $B_{ij} = B(e_i, e_j)$. If the bilinear form is symmetric then the matrix is also symmetric. If

$x = (x_1, \ldots, x_n)$ in the standard basis, then the bi-linearity of the form implies that

$$B(x, y) = x^t B y.$$

We can also approach this topic from another, more elementary angle. *Quadratic form* in variables $x_1, \ldots, x_n$ is a polynomial $Q(x_1, \ldots, x_n)$ whose monomials have the degree exactly 2. That is, it is a *homogeneous* polynomial of degree 2, and we can write it as

$$Q(x_1, \ldots, x_n) = \sum_{i=1}^{n} b_{ii} x_i^2 + 2 \sum_{1 \le i < j \le n} b_{ij} x_i x_j.$$

We can write this expression using matrices:

$$Q(\boldsymbol{x}) = \boldsymbol{x}^t B \boldsymbol{x},$$

where B is a symmetric matrix with entries $B_{ij} = B_{ji} := B_{ij}$. In this section we assume that B_{ij} are real.

This matrix B is exactly the matrix of the symmetric bilinear form that corresponds to quadratic form Q.

Example 8.1.1. What is the matrix for the form: $x_1^2 + 3x_2^2 + 5x_3^2 + 4x_1x_2 - 16x_1x_3 + 7x_2x_3$?

If we change the variables $\boldsymbol{x} = R\boldsymbol{y}$, where R is an invertible matrix, then in the new variables this form will be

$$Q(\boldsymbol{y}) = \boldsymbol{y}^t R^t B R \boldsymbol{y}.$$

The transformation

$$B \to R^t B R$$

is called the *congruence transformation* on matrices. Compare this with the *similarity transformation* $B \to X^{-1} B X$.

We are interested in the properties of quadratic forms and associated matrices which do not depend on the change of variables, that is, which are invariant with respect to congruence transformations.

The main fact here is that every symmetric matrix B can be brought to the diagonal form by a suitable congruence transformation. In fact, there are many congruence transformations that accomplish this task. The most

straightforward method is based on orthogonal diagonalization. Indeed, since B is a real symmetric matrix we can write it as

$$B = Q\Lambda Q^t,$$

where Q is an orthogonal matrix. Hence,

$$Q^t B Q = \Lambda,$$

and we are done.

Example 8.1.2. Consider the quadratic form $Q(x) = 2x_1^2 + 2x_1x_2 + 2x_2^2$. Bring it to the diagonal form.

There are some other methods, which are simpler computationally and involve only algebraic operations.

One of them is based on elementary row and column operations. Suppose that we reduce B by row operations to the upper-diagonal form as we did in the algorithm for LU decomposition. Note that B is symmetric and so when we perform row operations in the row reduction procedure, we can also do analogous operations on columns. As a result we will get a decomposition of the matrix B:

$$B = LDL^t,$$

where L is a lower-triangular matrix with ones on the main diagonal and D is the diagonal matrix with the pivots on the main diagonal.

Example 8.1.3. Let us find a "congruence" diagonalization of a matrix B by using the algorithm that we just described. We are looking for C such that $C^*AC = D$, where D is diagonal and C is non-singular. Let

$$B = \begin{bmatrix} 1 & 2 & -3 \\ 2 & 5 & -4 \\ -3 & -4 & 8 \end{bmatrix}$$

Then, we can do the following sequence of row and column transformations. [We perform column operations only on the left hand side of the augmented

matrix.]

$$\left[\begin{array}{ccc|ccc} 1 & 2 & -3 & 1 & 0 & 0 \\ 2 & 5 & -4 & 0 & 1 & 0 \\ -3 & -4 & 8 & 0 & 0 & 1 \end{array}\right] \xrightarrow{R_2-2R_1} \left[\begin{array}{ccc|ccc} 1 & 2 & -3 & 1 & 0 & 0 \\ 0 & 1 & 2 & -2 & 1 & 0 \\ -3 & -4 & 8 & 0 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{C_2-C_1} \left[\begin{array}{ccc|ccc} 1 & 0 & -3 & 1 & 0 & 0 \\ 0 & 1 & 2 & -2 & 1 & 0 \\ -3 & 2 & 8 & 0 & 0 & 1 \end{array}\right] \xrightarrow{R_3+3R_1} \left[\begin{array}{ccc|ccc} 1 & 0 & -3 & 1 & 0 & 0 \\ 0 & 1 & 2 & -2 & 1 & 0 \\ 0 & 2 & -1 & 3 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{C_3+3C_1} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 2 & -2 & 1 & 0 \\ 0 & 2 & -1 & 3 & 0 & 1 \end{array}\right] \xrightarrow{R_3-2R_2} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 2 & -2 & 1 & 0 \\ 0 & 0 & -5 & 7 & -2 & 1 \end{array}\right]$$

$$\xrightarrow{C_3-2C_2} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -2 & 1 & 0 \\ 0 & 0 & -5 & 7 & -2 & 1 \end{array}\right]$$

This means that

$$C^* = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 7 & -2 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -5 \end{bmatrix}$$

(From the practical point of view it is enough to do the row operations, the column operations are done only to illustrate that the matrix is indeed reduced to the diagonal form.)

Note that this algorithm fails if one needs to do an exchange of rows as for example for matrix

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

(In this case one can proceed by introducing the non-zero element by a row operation. For example:

$$\left[\begin{array}{cc|cc} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{array}\right] \xrightarrow{R_1+\frac{1}{2}R_2} \left[\begin{array}{cc|cc} 1/2 & 1 & 1 & 1/2 \\ 1 & 0 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{C_1+\frac{1}{2}C_2} \left[\begin{array}{cc|cc} 1 & 1 & 1 & 1/2 \\ 1 & 0 & 0 & 1 \end{array}\right] \to \dots$$

but the resulting matrix C^* is no longer lower-triangular.

8.2 Positive definite forms

Let Q be a quadratic form and A a corresponding symmetric matrix: $Q(x,x) = x^tAx$. (I have changed the notation from B to A here, sorry.) A quadratic form Q and the corresponding matrix A are called *positive definite* if $Q(\boldsymbol{x}) = \boldsymbol{x}^tA\boldsymbol{x} > 0$ for every $\boldsymbol{x} \neq 0$. It is clear that this property does not depend on the change of variables, so for invertible matrices R, matrix A is positive definite if and only if matrix R^tAR is positive definite.

In applications it is often needed to check whether a matrix is positive definite. In particular, $Q(\boldsymbol{x})$ has a strict minimum at 0 if and only $Q(\boldsymbol{x})$ is positive definite. We can check whether a quadratic form is positive-definite by using one of the criteria given by the following theorem.

Theorem 8.2.1. *Suppose $Q(\boldsymbol{x}) = \boldsymbol{x}^tA\boldsymbol{x}$ where A is a real symmetric matrix. Then, each of the following tests is a necessary and sufficient condition for the form $Q(\boldsymbol{x})$ to be positive definite:*

1. *All the eigenvalues of A satisfy $\lambda_i > 0$.*
2. *A can be reduced to the upper diagonal form without row exchanges and all the pivots (without row exchanges) satisfy $d_k > 0$.*
3. *All the upper left $k \times k$ sub-matrices A_k have positive determinants.*

Proof. Matrix A is real symmetric, so we can write an orthogonal diagonalization: $A = Q\Lambda Q^t$, so A is positive-definite if and only if Λ is positive definite, and the form for Λ is $Q(x) = \lambda_1x_1^2 + \ldots \lambda_nx_n^2$, and so it is clear that it is positive definite if and only if $\lambda_i > 0$ for every i.

Next, we are going to prove that if A is positive-definite, then (2) holds. We perform the algorithm described above. At every stage, after we perform a row operation and a corresponding row operation, the matrix remains positive definite. In particular, there can be no zero elements on the main diagonal, so a row exchange is never required. Eventually, we will get a diagonal matrix and all the diagonal elements d_k (pivots) must be positive.

Conversely, (2) implies that the matrix A is positive-definite. Indeed, condition (2) implies that we can find a decomposition of the matrix A:

$$A = LDL^t,$$

where L is a lower-triangular matrix with ones on the main diagonal and D is the diagonal matrix with the pivots on the main diagonal. Since $d_k > 0$, it follows that A is positive definite.

Finally we claim that (2) is equivalent to (3). Indeed, the row operations do not change the determinants of A_k. So if (2) holds then all of these determinants must be positive: $\det(A_k) = d_1 \dots d_k$. Conversely, if all of these determinants are positive, then the row exchange is never required (otherwise, one of the determinants at this stage would be equal to zero) and all pivots must be positive (or one of the determinants would be negative. □

Another useful fact is that a positive definite matrix A can be factorized as $R^t R$.

Theorem 8.2.2. *The symmetric $n \times n$ matrix A is positive definite if and only if there is a non-singular n matrix R with independent columns such that $A = R^t R$.*

Proof. Suppose $A = R^t R$, where R is non-singular. Then, $x^t A x = (Rx)^t Rx = \|Rx\| \geq 0$. If this quantity is zero, then $Rx = 0$, hence $x = 0$ because R is non-singular.

In the other direction, we can write $A = LDL^t$, where L is lower diagonal and D is a diagonal matrix with positive entries on the diagonal. Then we can take $R = \sqrt{D}L^t$, and observe that the columns of R are linearly independent. □

The decomposition $A = R^t R$, where R is upper-diagonal is often called the *Cholesky decomposition* of a positive definite matrix.

8.3 Law of Inertia

What can be said about more general situation, when the form $Q(x)$ represented by a symmetric matrix A is not necessarily positive definite?

It turns out that in this case, it can be reduced by a suitable change of variable to a form represented by a diagonal matrix that have only ± 1 or 0 on the main diagonal. Moreover, the number of positive, negative and zero items on the main diagonal of this diagonal matrix does not depend on the particular choice of this change of variables. This statement is called Sylvester's law of inertia and it follows from the following result.

Theorem 8.3.1. *Let A be a real symmetric matrix and C be a real invertible matrix. Then, matrix $C^t AC$ has the same number of positive eigenvalues, negative eigenvalues, and zero eigenvalues as A.*

So we can define the *signature* of a quadratic form as a triple (k_p, k_n, k_0), where the k_p, k_n, and k_0 is the number of positive, negative and zero entries on the main diagonal of any of these diagonal matrices. (Sometimes $k_p - k_n$ is also called the signature of the form.)

Proof of Theorem 8.3.1. We give a sketch of a proof for a simpler situation in which A is non-singular, so we do not need to worry about zero eigenvalues.

Let $C(u)$, $u \in [0,1]$, be a family of matrices such that $C(0) = C$, $C(1) = Q$, where Q is an orthogonal matrix, and $C(u)$ is never singular. (We will prove that it is possible to find such family of matrices below.) Then the matrix $C(u)^* AC(u)$ is never singular, so its determinant is never zero, and therefore, its eigenvalues are never zero. In addition, the eigenvalues of $C(u)^* AC(u)$ are continuous in u. (We skip the proof of this claim.) It follows that they can never change sign, when u changes from 0 to 1, and therefore, the number of positive eigenvalues of $C^* AC$ is the same as the number of positive eigenvalues of $Q^* AQ$. However, $Q^* AQ$ has the same eigenvalues as A.

In order to prove that there is a required $C(u)$ we can take Q from the QR decomposition $C = QR$. We choose the decomposition in such a way that R has positive entries on the main diagonal. Then we can write $C(u) = Q\big(uI + (1-u)R\big)$, and this matrix is always non-singular because the matrix $\big(uI + (1-u)R\big)$ is upper-diagonal and has positive entries on its diagonal.

□

8.4 Exercises

Exercise 8.4.1. Let

$$A = \begin{bmatrix} 1 & -3 & 2 \\ -3 & 7 & -5 \\ 2 & -5 & 8 \end{bmatrix}$$

Find a nonsingular real matrix C, such that $D = C^* AC$ is diagonal, and find $sign(A)$, the signature of A.

Exercise 8.4.2. Determine whether each of the following quadratic forms Q is positive definite:

(a) $Q(x, y, z) = x^2 + 2y^2 - 4xz - 4yz + 7z^2$.

(b) $Q(x, y, z) = x^2 + y^2 + 2xz + 4yz + 3z^2$.

Chapter 9

Singular Value Decomposition (SVD)

Reading for this Chapter

- Strang: Section 6.3
- Trefethen, Bau: Lectures 3, 4, and 5

9.1 Matrix norms

Matrices form a linear space so we can talk about norms of matrices. Since matrices also have some additional structure: for example, they act on vectors, – there are some additional issues for matrix norms.

The two most popular matrix norms are *Frobenius* and *operator* norms. The *Frobenius norm* is defined as follows:

$$\|A\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}|^2} = \sqrt{\mathrm{Tr}(A^*A)},$$

where Tr is the trace: $\mathrm{Tr}M = \sum_{i=1}^{n} M_{ii}$.

It is easy to see that the Frobenius norm of A is simply the norm of the long vector formed by stacking all column vectors of A together. The benefit of this norm is that it is essentially our familiar vector norm, in particular, there is an associated scalar product: $\langle A, B\rangle = \mathrm{Tr}(A^*B)$. One of the big advantages of the Frobenius norm is that it is easy to calculate. Another

useful norm, or rather a family of norms, is called the *operator norm* and it is defined by the following formula:

$$\|A\| := \sup_{v \neq 0} \frac{\|Av\|}{\|v\|} = \sup_{v:\|v\|=1} \|Av\|. \tag{9.1}$$

The operator norm depends on which vectors norms we choose to use to measure $\|v\|$ and $\|Av\|$. The most frequent situation is when both are usual Euclidean norms, that is, ℓ^2 vector norms: $\|v\| = (\sum v_j^2)^{1/2}$. Sometimes, to make this clear, the operator norm can be denoted $\|A\|_{(2,2)}$ or $\|A\|_2$. Below, if we say "operator norm" without qualifier we mean the 2-norm $\|A\|_2$.

From (9.1), it is clear that the operator norm equals the maximum increase in the length of a vector which is achieved by the linear transformation that have matrix A. Obviously, this is a useful quantity but it is more difficult to calculate.

Example 9.1.1. Let D is an $m \times n$ diagonal matrix with diagonal elements $d_1 \geq d_2 \geq \ldots \geq d_n \geq 0$. (We assume $m \geq n$.) What are the Frobenius and the operator norms of this matrix?

So far we talked about matrix norms as functions on matrices that satisfy the axioms of vector norms. However, sometimes additional requirements are imposed on matrix norms, which are related to such operations on matrices such as taking the adjoint (or transposition) and the multiplication. In particular, it is usually required that

$$\|A^*\| = \|A\|,$$

and

$$\|AB\| \leq \|A\|\|B\|.$$

Both the operator norm and the Frobenius norm satisfy these properties. For the operator norm it is essentially by definition and for the Frobenius norm it is an exercise based on the Cauchy-Schwarz inequality. (See Trefethen-Bau textbook, p.23, for a derivation.)

Another important property of these two norms is that they are invariant relative to unitary transformations.

Theorem 9.1.2. *For every $m \times n$ matrix A and every unitary $m \times m$ matrix Q, we have*

$$\|QA\|_2 = \|A\|_2, \text{ and}$$
$$\|QA\|_F = \|A\|_F.$$

9.2 Definition and existence of SVD

The motivation for the eigenvalue decomposition of a square matrix A is to find a basis $\{v_i\}_{i=1}^n$, in which the linear transformation A has the simplest possible form: $A : v_i \to \lambda_i v_i$. When we can find such decomposition, it gives an enormous insight in the properties of A. However, there are some problems with this approach.

1. Matrix A must be square, that is, the linear transformation must be an endomorphism: it maps the linear space to itself.

2. In many cases we encounter complex eigenvalues and eigenvectors

3. The basis of eigenvectors is not always orthogonal, which means that it is not easy to measure distances in this basis.

4. The eigenvalue decomposition does not always exist and we need to use the Jordan matrices instead of diagonal matrices.

These problems disappear for symmetric matrices but it is a big restriction.

The SVD decomposition is a different approach to the study of properties of linear transformations. Suppose $A : \mathbb{R}^n \to \mathbb{R}^m$. Then we look for two orthonormal bases $\{v_i\}_{i=1}^n$ and $\{u_i\}_{i=1}^m$ such that

$$Av_i = \sigma_i u_i, \text{ for all } i \leq \min\{n, m\}, \tag{9.2}$$

where $\sigma_i \geq 0$ are some real non-negative numbers. If $n > m$ we also require that $Av_i = 0$ for all $i > \min\{n, m\}$. (In fact this requirements is satisfied automatically.)

In matrix form, this is equivalent to the following definition.

Definition 9.2.1. A singular value decomposition (SVD) of an $m \times n$ matrix A is the following product

$$A = U\Sigma V^*, \tag{9.3}$$

where U is an $m \times m$ unitary matrix, V^* is an $n \times n$ unitary matrix and Σ is an $m \times n$ diagonal matrix with real *non-negative* entries. That is, if $i \neq j$ then $\Sigma_{ij} = 0$, otherwise $\Sigma_{ii} \geq 0$.

The diagonal elements of the matrix Σ are called singular values and denoted σ_i. For a real matrix A all elements in matrices U and V can be chosen to be real (so in particular, U and V are orthogonal matrices).

By convention, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$. This is always can be achieved by re-arranging columns of U and V.

The columns of matrices V and U are the orthonormal bases $\{v_i\}_{i=1}^n$ and $\{u_i\}_{i=1}^m$ that we introduced above and it is easy to see that property (9.4) holds. Moreover, we also have the following property:

$$A^* u_i = \sigma_i v_i, \text{ for all } i \leq \min\{n, m\}. \tag{9.4}$$

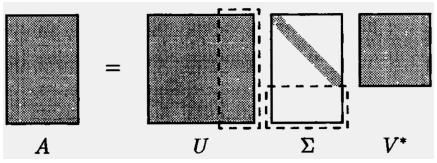


Figure 9.1: Full SVD decomposition, $m > n$

Intuitively, for $m \geq n$, if A represent a linear transformation, then we can write it as a rotation in $\mathbb{R}^n$, represented by V^*, followed by a map Σ that stretches the result and imbeds it isometrically to $\mathbb{R}^m$, and completed by another rotation in $\mathbb{R}^m$, represented by U.

In particular, this interpretation suggests that a unit sphere in $\mathbb{R}^n$ will be mapped to an ellipsoid in $\mathbb{R}^m$ and the half-lengths of the ellipsoid's principal axes will be equal to the singular values $\sigma_i := \Sigma_{ii}$.

The decomposition is clearly not unique if $m > n$. In the picture, the portion of the matrix U selected by dashed lines will be multiplied by zeros in the matrix Σ. Therefore, this portion can be chosen arbitrarily. Intuitively, we can rotate the orthogonal complement to the range of the map A in arbitrary way.

If we want to remove this source of non-uniqueness, then it is useful to define a *reduced singular value decomposition*. Assume that $m \geq n$ and that A is full rank, so that its range space has dimension n. Then the reduced SVD is

$$A = \widehat{U}\widehat{\Sigma}V^*, \tag{9.5}$$

where $\widehat{U}$ is an $m \times n$ matrix that has an orthonormal set of columns. Matrix $\widehat{\Sigma}$ is a square $n \times n$ diagonal matrix. And matrix V^* is the same as in full SVD, that is, it is an $n \times n$ unitary matrix.

In the reduced SVD, $\widehat{U}$ is not is not square (if $m \neq n$) and therefore it is not unitary. However, $\widehat{U}^*\widehat{U} = I_n$. Intutitively, the matrix $\widehat{U}$ is an isometric embedding of $\mathbb{R}_n$ in $\mathbb{R}_m$. Its columns give an orthonormal basis in the image of this embedding.

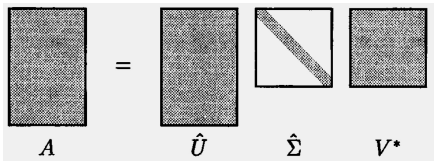


Figure 9.2: Reduced SVD decomposition, $m > n$

The reduced SVD is still not unique.
However, this non-uniqueness is mild.
It is up to permutation of certain columns and rows in these matrices and up to multiplication of columns and rows by ± 1. It can be almost fixed by requiring that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$ and that the first elements in columns of U and rows of V^* are positive. In exceptional cases when some σ_i are equal, some additional effort may be needed to get the uniqueness, however, this rarely happens in practice.

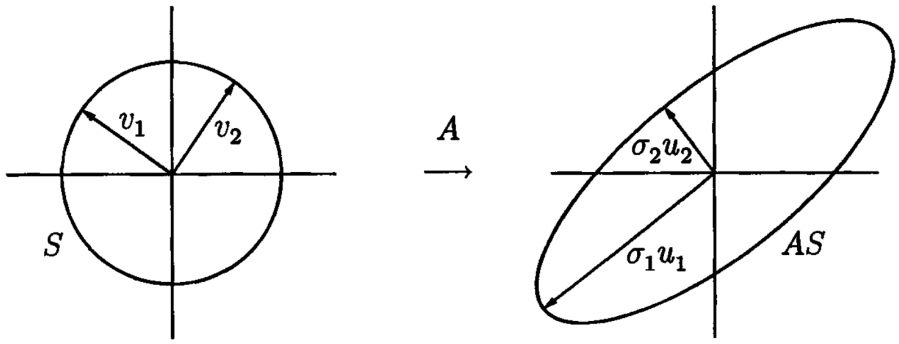


Figure 9.3: SVD decomposition of a 2×2 matrix

Geometric meaning of matrices $\widehat{U}, \widehat{\Sigma}, V$

If $v_1, v_2, \ldots v_n$ are the columns of V, $u_1, u_2, \ldots u_n$ are the columns of $\widehat{U}$, and $\sigma_1, \ldots, \sigma_n$ are the diagonal entries of $\widehat{\Sigma}$, then matrix A sends $v_i \to \sigma_i u_i$. See illustration in Figure 9.3.

The existence of the SVD decomposition

Here is Theorem 4.1 from Trefethen - Bau.

Theorem 9.2.2. *Every matrix A has a singular value decomposition (9.5). Furthermore the singular values σ_i are uniquely determined. If A is square and the σ_i are distinct then the corresponding column vectors in U and V are uniquely determined up to a multiplication by a scalar that have absolute value* 1.

For the complete proof, see the Trefethen-Bau book. Here is a sketch of the proof of the existence claim for $m \geq n$.

Proof of the existence claim. For concreteness, let us work with real matrices.

We will use induction on the size of the matrix and leave the base of the induction (the case when $n = 1$) as an exercise.

By a compactness argument, the supremum in the definition of the matrix norm (9.1) is attained on a vector v_1, and so there exist vectors u_1 and v_1 such that $u_1 = Av_1/\|A\|$, $|v_1| = 1$, $|u_1| = 1$. (In addition it can be proved that for a real matrix A, the maximizing vector v_1 can be chosen to be real.)

Let us define $\sigma_1 = \|A\|$, so we have $u_1 = \sigma_1 A v_1$. Complete the vectors u_1 and v_1 to a pair of orthonormal bases $\{u_i\}$ and $\{v_j\}$ in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. Let U_1 and V_1 be the matrices with columns u_i and v_i, respectively.

Then from $Av_1 = u_1$ we have that

$$U_1^* A V_1 = S = \begin{bmatrix} \sigma_1 & w^* \\ 0 & B \end{bmatrix}.$$

We claim that in fact if the norm of A is attained on v_1, then the vector w must be zero.

Indeed, S is obtained from A by a multiplication by two orthogonal matrices on both sides, so it has the same norm as A, that is, $\|S\| = \sigma_1$. Then, we notice that the first element of the vector

$$S \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} = \begin{bmatrix} \sigma_1 & w^* \\ 0 & B \end{bmatrix} \begin{bmatrix} \sigma_1 \\ w \end{bmatrix}$$

is $\sigma_1^2 + w^* w$. Hence

$$\left\| S \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\| \geq \sigma_1^2 + w^* w = \left(\sigma_1^2 + w^* w\right)^{1/2} \left\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|$$

So $\|S\| \geq \left(\sigma_1^2 + w^* w\right)^{1/2}$, so it must be that $w = 0$.

Also note that $\|B\| \leq \|A\|$.

However, then we can apply the induction hypothesis to the matrix B and notice that it can be written as $B = U_2 \Sigma_2 V_2^*$.

This leads to the decomposition

$$A = U_1 \begin{bmatrix} 1 & 0 \\ 0 & U_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V_2 \end{bmatrix}^* V_1^*,$$

which gives an SVD for matrix A. □

Note that in fact, the proof gave us more than the existence of the SVD. It also showed that the largest singular value σ_1 equals to the maximum of $\|Ax\|$ subject to the constraint $\|x\| = 1$, and that the maximum is achieved at the right singular vector v_1.

Moreover, by analyzing the proof, we see that $\sigma_2 = \max \|Ax\|$ subject to $\|x\| = 1$ and an additional constraint $x \perp v_1$ and that this maximum is achieved at v_2. We can continue this and find that $\sigma_k = \max \|Ax\|$ subject to the constraints that $\|x\| = 1$ and that $x \perp \operatorname{span}(v_1, \ldots, v_{k-1})$.

9.3 Relation to eigenvalue decomposition

Previously, we learned that many matrices can be diagonalized and represented in the form $A = X\Lambda X^{-1}$ where Λ is the diagonal matrix of eigenvalues and X is the matrix, whose columns are eigenvectors.

For general matrices, the connection between eigenvalues and singular values is not straightforward. There is a bunch of inequalities between the singular values and absolute values of eigenvalues. There is also a wonderful connection between them for large random matrices, however, we are not going to talk about it here.

The eigenvalue diagonalization is very useful when matrix A is symmetric (or Hermitian in the complex case). In this case, all eigenvalues are real and one can choose eigenvectors in such a way that they form an orthonormal set, so that matrix X is orthogonal. This is very close to the SVD decomposition and the difference is that some eigenvalues may happen to be negative, while all singular values must be non-negative.

Theorem 9.3.1. *If A is a symmetric $n \times n$ matrix, then the singular values of A are the absolute values of the eigenvalues of A, $\sigma_i = |\lambda_i|$, for $i = 1, \ldots, n$.*

Proof. In the case of symmetric (or Hermitian) matrices, we have the eigenvalue decomposition:

$$A = Q\Lambda Q^*,$$

where Λ and Q are diagonal and orthogonal (or unitary) matrices, respectively. We can easily convert it to the SVD decompositions by multiplying some of the columns by -1,

$$A = Q|\Lambda|\operatorname{sign}(\Lambda)Q^*,$$

where $|\Lambda|$ is the diagonal matrix with $|\lambda_i|$ on the main diagonal and $\operatorname{sign}(\Lambda)$ is the diagonal matrix with the diagonal entries $\operatorname{sign}(\lambda_i)$. We can also choose

the ordering of λ_i in such a way that its absolute values decrease: $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. This decomposition shows that $\sigma_i = |\lambda_i|$. □

The proof also shows that in the SVD decomposition of a symmetric or an Hermitian matrix, $U = Q$ and V equals Q with some of the columns multiplied by -1.

For more general matrices, we can still use the eigenvalue diagonalization to calculate the SVD.

Theorem 9.3.2. *Let A be an $m \times n$ matrix and $t = \min\{m, n\}$. Matrices A^*A and AA^* have the same sets of t eigenvalues with the largest absolute value, and the t singular values of A are the square roots of these eigenvalues.*

Proof. Let the (full) singular value decomposition for A be

$$A = U\Sigma V^*,$$

where Σ is $m \times n$ matrix and U and V are orthogonal. Then,

$$A^*A = V(\Sigma^*\Sigma)V^*,$$
$$AA^* = U(\Sigma\Sigma^*)U^*$$

where A^*A is $n \times n$ matrix and A^*A is $m \times m$. The matrices $\Sigma^*\Sigma$ and $\Sigma\Sigma^*$ are diagonal and its first t diagonal elements are $\sigma_1^2, \ldots, \sigma_t^2$.

Hence the first t eigenvalues of A^*A with the largest absolute value are the same as the first t eigenvalues of AA^* with the largest absolute value, and both sets are equal to $\{\sigma_1^2, \ldots, \sigma_t^2\}$. □

Note that the proof also shows that V corresponds to eigenvectors of matrix A^*A. And U of the full SVD can be calculated as the matrix of eigenvectors of AA^*.

In the situation when $m > n$ we are typically interested in the reduced SVD and we can observe that A maps column vectors of V to column vectors of $\widehat{U}$ (the U matrix of the reduced decomposition, except it stretches them by the singular values. Hence, we can calculate $u_i = (1/\sigma_i)Av_i$. The situation with $\sigma_i = 0$ is special. In this case one can simply take a unit vector u_i which is perpendicular to all other left-singular vectors. (See an example below for an illustration.)

Note also that this theorem gives another proof of Theorem 9.3.1, since for a real symmetric matrix A, we have $A^*A = A^2$ and the eigenvalues of A^2 are equal to the squares of eigenvalues of A. So, by Theorem 9.3.2, singular values of A are equal to $\sqrt{\lambda_i^2} = |\lambda_i|$, absolute values of eigenvalues of A.

Example 9.3.3. Find the (reduced) SVD decomposition of the matrix

$$A = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix}.$$

We have

$$A^*A = \begin{bmatrix} 9 & -9 \\ -9 & 9 \end{bmatrix}$$

Then we can calculate the eigenvalues $\lambda_1 = 18$, $\lambda_2 = 0$. (This can be done either by finding the roots of the characteristic polynomial, or by noticing that the matrix is singular, so one of the eigenvalues must be zero, and finding the second one from the fact that the trace of the matrix is equal to the sum of the eigenvalues.) Hence the singular values are $\sigma_1 = \sqrt{18} = 3\sqrt{2}$, $\sigma_2 = 0$. The eigenvectors of A^*A are $v_1 = \frac{1}{\sqrt{2}}[1, -1]^*$ and $v_2 = \frac{1}{\sqrt{2}}[1, 1]^*$. These are right singular vectors.

We calculate the first left singular vector as

$$u_1 = \frac{1}{3\sqrt{2}} A v_1 = \frac{1}{3\sqrt{2}} \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix}$$

Since $\sigma_2 = 0$, we can take any unit vector perpendicular to u_1 as the second left singular vector. For example, $u_2 = \frac{1}{\sqrt{5}}[2, 1, 0]^*$ will do.

So, one possible reduced SVD of A is

$$A = \begin{bmatrix} 1/3 & 2/\sqrt{5} \\ -2/3 & 1/\sqrt{5} \\ 2/3 & 0 \end{bmatrix} \begin{bmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

9.4 Properties of the SVD and singular values

Theorem 9.4.1. *Let $A = U\Sigma V^*$ be the full SVD of A and let r be the number of non-zero singular values. Then*

$$\begin{aligned} \text{Range}(A) &= \text{span}\{u_1, \ldots, u_r\}, \\ \text{Null}(A) &= \text{span}\{v_{r+1}, \ldots, v_n\}, \end{aligned}$$

where u_i and v_j are columns of matrices U and V respectively. In particular the rank of A equals r.

Proof. The matrices U and V are full rank orthogonal matrices. Essentially they simply rotate $\mathbb{R}^m$ and $\mathbb{R}^n$. What is important is that the $\text{Range}(\Sigma) = \text{span}\{e_1, \ldots, e_r\}$ in $\mathbb{R}^m$ and $\text{Null}(\Sigma) = \text{span}\{e_{r+1}, \ldots, e_n\}$ in R^n. □

The operator and Frobenius norms of a matrix can be written in terms of its singular values.

Theorem 9.4.2. *Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ be non-zero singular values of matrix A. Then,*

$$\|A\|_2 = \sigma_1,$$
$$\|A\|_F = \sqrt{\sigma_1^2 + \ldots + \sigma_r^2}.$$

Proof. Note that multiplication by an orthogonal (or unitary) matrix does not change the norm of a vector. This implies that $\|A\|_2 = \|\Sigma\|_2$, and it is easy to check that $\|\Sigma\|_2 = \sigma_1$. For the Frobenius norm, we calculate:

$$\begin{aligned}\|A\|_F^2 = \text{Tr}(A^*A) &= \text{Tr}\Big((V\Sigma^*U^*)(U\Sigma V^*)\Big) \\ &= \text{Tr}\Big(V\Sigma^*\Sigma V^*\Big) = \text{Tr}(\Sigma^*\Sigma),\end{aligned}$$

where the last step is by the property of the trace: $\text{Tr}(AB) = \text{Tr}(BA)$.

And the last quantity is easy to calculate:

$$\text{Tr}(\Sigma^*\Sigma) = \sigma_1^2 + \ldots + \sigma_r^2.$$

□

Now let us consider the relation of eigenvalues and singular values to the determinant. For eigenvalues, we have seen that $\det(A) = \prod_{i=1}^n \lambda_i$. If matrix A has an eigenvalue decomposition, then

$$\begin{aligned}\det(A) = \det(X\Lambda X^{-1}) &= \det(X)\det(\Lambda)\det(X)^{-1} \\ &= \det(\Lambda) = \prod_{i=1}^n \lambda_i.\end{aligned}$$

In general it follows because $\det(zI - A) = (z - \lambda_1)\ldots(z - \lambda_n)$ by setting $z = 0$.

It turns out that we can also write a similar formula using the singular values, except that we lose the information about the sign of the determinant.

Theorem 9.4.3. *For an $m \times m$ matrix A,*

$$|\det(A)| = \prod_{i=1}^{m} \sigma_i,$$

where σ_i are singular values of the matrix A.

Proof. By using the multiplicative property of the determinant, we write:

$$\det(A) = \det(U\Sigma V^*) = \det(U)\det(\Sigma)\det(V^*)$$

Now we use the fact that the determinant of a unitary matrix has absolute value 1. (This holds because (i) $\det(U)\det(U^*) = \det(UU^*) = 1$, and (ii) $\det(U^*) = \overline{\det(U)}$. Hence $|\det(U)|^2 = 1$, and therefore $|\det(U)| = 1$.) Therefore

$$|\det(A)| = |\det(\Sigma)| = \prod_{i=1}^{m} \sigma_i.$$

□

9.5 Low-rank approximation via SVD

The SVD is useful because it allows us to construct low-rank approximations to a matrix which are optimal both in the Frobenius and operator norms.

Given an integer $\nu \geq 1$, a *rank-ν approximation* to a matrix A in a norm $\|\cdot\|$ is a matrix B that has rank ν and minimizes the norm of the difference $A - B$.

Theorem 9.5.1. *Let an $m \times n$ matrix A has rank r, and let $A = U\Sigma V^*$ be its SVD, with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$. Then*

$$A_\nu = \sum_{j=1}^{\nu} \sigma_j u_j v_j^*$$

is a rank-ν approximation to A in the operator norm. Moreover, for $\nu < r$ the error of the approximation

$$\inf_{B:\mathrm{rank}(B)\leq\nu} \|A - B\| = \|A - A_\nu\| = \sigma_{\nu+1}.$$

(For $\nu \geq r$, $A_\nu = A$.)

Proof. Suppose that there is some matrix B with the rank $\leq \nu$, which outperform A_ν. Namely, suppose that $\|A - B\|_2 < \|A - A_\nu\|_2 = \sigma_{\nu+1}$. Since the matrix B has rank $\leq \nu$, therefore its null-space W has dimension $\geq n - \nu$. For every vector in $w \in W$, we have

$$\|Aw\| = \|(A - B)w\| < \sigma_{\nu+1}\|w\|.$$

On the other hand, if V is the linear subspace spanned by the first $\nu + 1$ singular vectors of A, then we have that for every $v \in V$,

$$\|Av\| \geq \sigma_{\nu+1}\|v\|.$$

Since the sum of the dimensions of W and V exceeds n, they must have a non-zero vector in common. This gives a contradiction. □

An analogous result holds also for the Frobenius norm.

9.6 Applications

9.6.1 Linear regression and pseudoinverse

Consider the linear regression problem $Xb = y$, where X is an $m \times n$ matrix.

If $m > n$ and the data matrix X has the reduced SVD

$$X = \widehat{U}\widehat{\Sigma}V^*,$$

then

$$X^*X = V\widehat{\Sigma}^*\widehat{U}^*\widehat{U}\widehat{\Sigma}V^* = V\widehat{\Sigma}^2V^*,$$

where V is the orthogonal matrix, with $V^{-1} = V^*$.

In particular we can write the normal equations $X^*Xb = X^*y$ as

$$V\widehat{\Sigma}^2V^*b = V\widehat{\Sigma}U^*y,$$

and by mulitplying by V^* on the left:

$$\widehat{\Sigma}^2V^*b = \widehat{\Sigma}U^*y,$$

Assuming that all diagonal entries of $\widehat{\Sigma}$ are positive we can reduce the system even further to

$$\widehat{\Sigma}V^*b = U^*y,$$

and then the algorithm is simple:

1. Calculate U^*y.

2. Solve $\widehat{\Sigma}w = U^*y$ for w. (This is simple because $\widehat{\Sigma}$ is diagonal.)

3. Set $x = Vw$.

The most work in this algorithm goes into calculating the singular value decomposition. According to Trefethen and Bau, this method has some advantages over other methods if some of the singular values of the matrix X are small.

What if $m > n$? In this case the equation $Xb = y$ has more than one solution. To select one of them with some nice properties, one can use the full decomposition $X = U\Sigma V^*$ and define the pseudoinverse $X^+ = V\Sigma^+ U^*$, where Σ^+ is $n \times m$ matrix with

$$\Sigma^+ = \begin{bmatrix} \sigma_1^{-1} & 0 & \dots & 0 \\ 0 & \sigma_2^{-1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

(In other words, all non-zero diagonal elements of Σ are inverted and the matrix is also transposed.)

Then $\Sigma^+\Sigma = I_n$ and therefore $X^+X = I_n$, so that X^+ is the left inverse of X. In particular we can calculate one of possible solutions as $b = X^+y$.

9.6.2 Principal Component Analysis (PCA)

The singular value decomposition is often used in data analysis for dimension reduction. The basic idea that we are trying to approximate a matrix of data with a low-rank matrix.

Suppose X is the matrix of data. The rows of this matrix are data points and the columns give values of various variables (also called features) for these datapoints. For example, rows can correspond to different individuals and columns to different characteristics of the individual. For another example, rows can correspond to dates and the columns to different financial stocks while the entries are the stock returns recorded on that day.

One statistical technique to analyze data X is called the principal component analysis. It is essentially the SVD of matrix X.

If we write the reduced SVD of $m \times n$ matrix X: $X = U\Sigma V^\star$ then the j-th column of matrix V is called the j-th *principal component* and its elements are called the *loadings* of the the j-th component.

The elements of the j-th column of matrix U are connected to the "scores" of the j-th principal component for a particular observation. So for example, $\sigma_j U_{ij}$ is the score of the j-th component for i-th observation.

We can also write:

$$X = U\Sigma V^* = \sum_{k=1}^{n} \sigma_k u^{(k)} \left(v^{(k)}\right)^*, \tag{9.6}$$

where $u^{(k)}$ and $v^{(k)}$ are k-th columns of the matrices U and V, respectively.

In particular the scores of the k-th principal component can be calculated as

$$\sigma_k u^{(k)} = X v^{(k)},$$

which means that the vector of scores is a linear combination of columns of X ("features"), with coefficients given by entries of the vector $v^{(k)}$

In components, this can be written as

$$x_{ij} = \sum_{k=1}^{r} \sigma_k u_{ik} v_{jk}$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

Note that V is the matrix of eigenvectors of matrix X^*X which has the meaning of the empirical covariance matrix for the data. Statistically one can think about the first column V (i.e., the first eigenvector) as the coefficients of the linear combination of variables that has the largest variance (that is, for which the quadratic form v^*X^*Xv achieves its maximum, assuming that v has unit length. In other words this column gives the coefficients of the linear combination of characteristics with the largest variation across individuals.

Similar interpretations can be given for other columns of matrix V.

Often for visualization purposes only the first two principal components are used and the observation vector $x_{i1}, \ldots x_{ip}$ is replaced with the scores on the first two principal components: $\sigma_1 U_{i1}$ and $\sigma_2 U_{i2}$.

We also know that the best approximation to the matrix X with rank r is given by

$$X = U\Sigma V^* = \sum_{k=1}^{r} \sigma_k u^{(k)} \left(v^{(k)}\right)^*, \tag{9.7}$$

where the sum in (9.6) is cut at $r \leq n$. This is the basis for the dimension reduction technique when X is replaced with the matrix of the scores for

the first r components, that is with matrix whose columns are $\sigma_k u^{(k)}$, $k = 1, \ldots, r$.

This technique is very popular. One example is the data on financial stock returns. It turns out that the empirical covariance matrix exhibit three important factors (which are principal components with large singular values).

9.6.3 Face recognition

The SVD can be used for face recognition. This is one of the applications of PCA. Currently, this method is less popular than methods based on neural networks.

In order to use PCA for face recognition, face images are vectorized (that is, an image is represented as a long vector of pixel values). Then a collection of these vectors for a large number of individuals is put together as a matrix. For example, let X be a matrix where columns represent individuals and rows are pixels in an image.

After the SVD is performed on this matrix, we have as before:

$$x_{ip} = \sum_{k=1}^{r} \sigma_k u_i^{(k)} v_p^{(k)}$$

where p stands for a person.

In this application, the vectors u_i^k are the principal components, or "eigenfaces".

Note that the decomposition above means that the $\sigma_k v_p(k)$, $k = 1, \ldots, r$ are coefficients in the expansion of the p-th column vector X_{ip} (which is the image of the person p) over the orthogonal basis given by eigenfaces $u^{(k)}$.

We can interpret the vector $\sigma_k v_p^{(k)}$, $k = 1, \ldots, r$ as the "signature" of the individual p. These signatures are stored in a database. When a new face image is presented, it is decomposed in the eigenface basis and compared to the signatures in the database. If a sufficiently close match is found, the face is recognized.

9.6.4 Image Processing

1. An SVD was suggested as a method for image compressing, however, the standard technologies use different compressing algorithms. In particular, JPEG uses the discrete cosine transform, which is a variant of the Fast Fourier Transform.

The SVD method is straightforward. An image can be represented as 3 matrices of pixels. Every matrix can be subjected to SVD and a low-rank approximation computed. Then it is only necessary to retain several largest singular values and the corresponding singular vectors. This gives a significant compressing ratio.

2. The SVD can be used in removing static background from videos. Videos can be converted to matrices by vectorizing each frame and stacking them together. In this case the background is the low-rank approximation to the matrix and can be removed by calculating the low-rank approximation and subtracting it from the matrix.

9.6.5 Other applications

- The SVD has some application in continuous mechanics and in robotics since it decomposes a matrix as a product of two rotations, both of which can be accomplished without stress, and a stretching matrix.
- Eigenvalue decomposition is used in the spectral clustering algorithms.

9.7 Exercises

Exercise 9.7.1. Example 3.6 in Trefethen-Bau shows that if A is an outer product of two vectors $A = uv^t$, then $\|A\|_2 = \|u\|_2\|v\|_2$, where $\|\cdot\|_2$ denotes both the 2-norm on vectors (the usual Euclidean norm) and the corresponding induced operator norm on matrices.

Is the same true for the Frobenius norm, that is, is $\|A\|_F = \|u\|_F\|v\|_F$? Prove it or give a counterexample.

Exercise 9.7.2. Determine the SVDs of the following matrices (by hand calculation):

$$\text{(a)} \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}, \text{(b)} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}, \text{(c)} \begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \text{(d)} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \text{(e)} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

(Note that the answers can be different up to some multiplication of columns of U and V by ± 1.)

Exercise 9.7.3. Determine the SVD of the following matrix (by hand calculation):

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

Exercise 9.7.4. Suppose A is an $m \times n$ matrix and B is the $n \times m$ matrix obtained by rotating A ninety degrees clockwise on paper. Do A and B have the same singular values? Prove that the answer is yes or give a counterexample.

Chapter 10

Further properties of eigenvalues and singular values

Reading for this Chapter

- Strang: Section 6.4, Chapter 7.
- Trefethen, Bau: Lectures 12, Chapter 5.

10.1 Condition Number

Let $y = Ax$, where A is an $m \times n$ matrix. In applications it is often important to know whether small changes in input can lead to big changes in output, and it is often useful to measure the size of the changes in relative terms. Then we define a *relative condition number* at input x, as

$$\kappa(x) := \sup_{\delta x \neq 0} \left[\frac{\|A(x+\delta x) - Ax\|}{\|Ax\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right] = \sup_{\delta x \neq 0} \left[\frac{\|A\delta x\|}{\|\delta x\|} \right]$$

By definition of the operator norm we see that

$$\kappa(x) = \|A\| \frac{\|x\|}{\|Ax\|}.$$

Now sometimes we want a bound on the relative condition number which would be independent of the input. Let us assume that A is a *square* non-singular matrix.

Then we have

$$\frac{\|x\|}{\|Ax\|} = \frac{\|A^{-1}Ax\|}{\|Ax\|} \leq \|A^{-1}\|.$$

Hence, we have $\kappa(x) \leq \|A\|\|A^{-1}\|$.

Theorem 10.1.1. *Let A be a square non-singular matrix and consider the equation $Ax = b$. The problem of computing b, given x, has the relative condition number*

$$\kappa = \|A\|\frac{\|x\|}{\|b\|} \leq \|A\|\|A^{-1}\|,$$

with respect to perturbations of x. The problem of computing x, given b, has the relative condition number

$$\kappa' = \|A^{-1}\|\frac{\|b\|}{\|x\|} \leq \|A\|\|A^{-1}\|,$$

with respect to perturbations of b.

Proof. We proved the first part above. For the second part, note that we can re-write the problem of computing x given b as $A^{-1}b = x$, and then we can apply the first part. □

We know that $\|A\| = \sigma_1(A)$ and $\|A^{-1}\| = 1/\sigma_n$, where $\sigma_1(A)$ and $\sigma_n(A)$ are the largest and the smallest singular values of A. So we can write a bound $\kappa(x) \leq \sigma_1/\sigma_n$.

The first part of this theorem can be easily generalized to non-square matrices. Indeed, if the matrix A is $m \times n$ with $m > n$, then we can replace A^{-1} in the arguments above with the pseudo-inverse A^+, and then $k(x) \leq \|A\|\|A^+\| = \sigma_1/\sigma_n$.

The quantity σ_1/σ_n is called the *condition number* of the matrix A. It is a universal bound for the relative condition number $\kappa(x)$ which is valid for all inputs $x \neq 0$.

In fact this number also controls the sensitivity of output to perturbations in the matrix.

Theorem 10.1.2. *Let b be fixed and consider the problem of computing $x = A^{-1}b$, where A is square and nonsingular. The relative condition number of this problem with respect to perturbations in A is*

$$\kappa = \|A\|\|A^{-1}\| = \frac{\sigma_1(A)}{\sigma_n(A)}.$$

Proof. If we perturb A in the equation $Ax = b$, we find that

$$(A + \delta A)(x + \delta x) = b.$$

By using the equality $Ax = b$ and dropping the second order term $\delta A \delta x$, we find $(\delta A)x + A\delta x = 0$, or $\delta x = -A^{-1}(\delta A)x$. This implies that $\|\delta x\| \leq \|A^{-1}\|\|\delta A\|\|x\|$, which we can re-write as:

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta A\|}{\|A\|}.$$

This shows that the relative condition number is bounded from above by $\|A\|\|A^{-1}\|$.

In fact it is possible to show that the bound is achieved for some δA. We omit the proof of this fact. See the book by Trefethen and Bau for details. □

10.2 Rayleigh quotient

Rayleigh quotient is an important way to characterize eigenvalues as the maximum of a quadratic form.

In order to motivate this property, note that the largest singular value $\sigma_1(A)$ equals the norm of the matrix A, which is the maximum of the quotient

$$\frac{\|Ax\|}{\|x\|}.$$

over all possible non-zero x. The corresponding left singular vector is the vector at which this maximum is achieved. The square of this expression can be rewritten as

$$\frac{\|Ax\|^2}{\|x\|^2} = \frac{(Ax, Ax)}{(x, x)} = \frac{(x, A^*Ax)}{(x, x)}.$$

Hence the square of the largest singular value is the maximum of the expression (x, A^*Ax) given that $(x, x) = 1$.

The Rayleigh quotient is a modification of this idea, which focuses on eigenvalues instead of singular values. By definition, the *Rayleigh quotient* of a vector x is the ratio:

$$R(x) = \frac{(x, Ax)}{(x, x)}.$$

Theorem 10.2.1 (Rayleigh-Ritz)**.** *If a symmetric matrix A has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, then λ_1 and λ_n are the maximum and the minimum, respectively, of the Rayleigh quotient $R(x)$ over all $x \neq 0$.*

Proof. We need to check that

$$\lambda_n(x,x) \leq (x, Ax) \leq \lambda_1(x,x) \tag{10.1}$$

holds and that the bounds can be achieved by a suitable choice of $x \neq 0$.

The inequalities hold because $A = U\Lambda U^*$ and so

$$(x, Ax) = (U^*x, \Lambda U^*x) = \lambda_1 y_1^2 + \ldots + \lambda_n y_n^2.$$

where $y = (y_1, \ldots, y_n)^* = U^*x$. The last expression is between $\lambda_n\|y\|^2$ and $\lambda_1\|y\|^2$ and we know that $\|y\|^2 = \|x\|^2$.

It is also clear that the bounds in the inequalities (10.1) are achieved if we set x equal to the eigenvectors corresponding to eigenvalues λ_1 and λ_n. □

For example, for the largest eigenvalue we have

$$\lambda_1 = \max_{x \neq 0} \frac{(x, Ax)}{(x,x)}.$$

Alternatively we can write:

$$\lambda_1 = \max_{x:\|x\|=1} (x, Ax),$$

and the maximum is achieved on an eigenvector of A that corresponds to the eigenvalue λ_1.

Note that if $Q(x)$ is the quadratic form associated to the symmetric matrix A, then this gives us ability to find the maximum of $Q(x)$ on the set of all vectors x that have unit length.

Similarly, for the smallest eigenvalue we have:

$$\lambda_n = \min_{x:\|x\|=1} (x, Ax),$$

and again the minimum is achieve at the eigenvector that corresponds to the smallest eigenvalue λ_n.

Corollary 10.2.2. *The diagonal entries of any symmetric matrix are between λ_1 and λ_n.*

Proof. This is a consequence of Theorem 10.2.1 because the diagonal entry $a_{ii} = R(\boldsymbol{e}_i)$, where $\boldsymbol{e}_i = (0, \ldots, 1, \ldots, 0)$ is the i-th coordinate vector. □

Characterization of the largest and smallest eigenvalues as the maximum and minimum, respectively, of a quadratic form can be extended to intermediate eigenvalues. Let V_{k-1} be the space spanned by orthonormal system of eigenvectors $u_1, u_2, \ldots u_{k-1}$ that correspond to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{k-1}$. Then,

$$\lambda_k = \max_{x \neq 0, x \perp V_{k-1}} R(x).$$

In order to see this, note that the space $V_{k-1}^{\perp}$ orthogonal to V_{k-1} is invariant under the transformation A and spanned by the eigenvectors corresponding to the eigenvalues $\lambda_k, \ldots, \lambda_n$. Then the desired result can be obtained by restricting the linear transformation A to the linear space $V_{k-1}^{\perp}$ and applying the Rayleigh-Ritz theorem to this restriction.

For example, the second eigenvalue of A gives the following maximum:

$$\lambda_2 = \max_{x \neq 0, x \perp u_1} \frac{(x, Ax)}{(x, x)},$$

where u_1 is the first eigenvector corresponding to λ_1. Alternatively we can write this expression as

$$\lambda_2 = \max_{x: \|x\| = 1, x \perp u_1} (x, Ax).$$

The maximum is achieved at u_2, an eigenvector that corresponds to λ_2.

A useful extension of this result is the Courant-Fisher Theorem. It says that instead of explicitly choosing V_k as the span of the first k eigenvectors, one can solve a minmax problem.

Theorem 10.2.3 (Courant-Fisher)**.** *If a symmetric matrix A has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, then for $1 \leq k \leq n$,*

$$\lambda_k = \min_{V_{k-1}} \max_{x \neq 0, x \perp V_{k-1}} R(x),$$

where the minimization is over all $k-1$ dimensional subspaces V_{k-1}, and

$$\lambda_k = \max_{V_{n-k}} \min_{x \neq 0, x \perp V_{n-k}} R(x),$$

where the maximization over all $n-k$ dimensional subspaces V_{n-k}.

The difference (and the benefit) of this theorem from our previous considerations is that it does not define V_{k-1} as the span of eigenvectors $u_1, \ldots, u_{k-1}$ but allows V_{k-1} to run over all possible $k-1$ subspaces and chooses the "worst" possible subspace. The worst here means that it leads to the smallest of the maximal Rayleigh ratios. (The second expression is similar but the maximum and minimum are exchanged in this expression.)

The Courant-Fisher Theorem allows proving several important theoretical results. One of the most useful is a theorem by Hermann Weyl. Let us write $\lambda_j(X)$ to denote the eigenvalues of an Hermitian matrix X arranged in decreasing order.

Theorem 10.2.4 (Weyl)**.** *Let A and B be two Hermitian $n \times n$ matrices. For each $k = 1, 2, \ldots, n$, we have*

$$\lambda_k(A) + \lambda_n(B) \leq \lambda_k(A+B) \leq \lambda_k(A) + \lambda_1(B)$$

Proof. For every vector x, we have

$$\lambda_1(B) \geq \frac{x^t B x}{x^t x} \geq \lambda_n(B).$$

Hence

$$\begin{aligned}
\lambda_k(A+B) &= \min_{V_{k-1}} \max_{x \neq 0, x \perp V_{k-1}} \frac{x^t(A+B)x}{x^t x} \\
&= \min_{V_{k-1}} \max_{x \neq 0, x \perp V_{k-1}} \left[\frac{x^t A x}{x^t x} + \frac{x^t B x}{x^t x} \right] \\
&\leq \min_{V_{k-1}} \max_{x \neq 0, x \perp V_{k-1}} \left[\frac{x^t A x}{x^t x} + \lambda_1(B) \right] = \lambda_k(A) + \lambda_1(B).
\end{aligned}$$

The lower bound can be established similarly. □

Corollary 10.2.5. *If matrix B is non-negative definite, then all eigenvalues of A increase when we add B:*

$$\lambda_k(A+B) \geq \lambda_k(A)$$

Another important and surprising result is as follows.

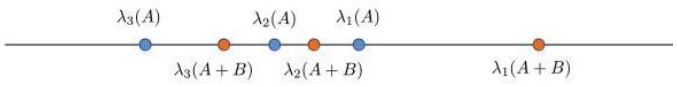


Figure 10.1

Theorem 10.2.6 (interlacing of eigenvalues I)**.** *If the matrix B is non-negative definite and has rank 1, then*

$$\lambda_{k+1}(A) \leq \lambda_{k+1}(A+B) \leq \lambda_k(A),$$

where $k = 0, \ldots, n-1$, with the convention that $\lambda_0(A) = +\infty$.

In other words the rank-one perturbation of matrix A cannot move the internal eigenvalues too much. This is illustrated in Figure 10.1.

Proof. Note that $\lambda_k(A+B) \geq \lambda_k(A)$ holds by Corollary 10.2.5, so we only need to prove the other inequality.

By the eigenvalue decomposition, every non-negative definite symmetric rank 1 matrix can be written as a outer product of a vector with itself. So let $B = vv^t$. For $1 \leq k \leq n-1$ we write the following sequence of inequalities (where $x \neq 0$ always):

$$\begin{aligned}
\lambda_k(A) &= \min_{V_{k-1}} \max_{x \perp V_{k-1}} \frac{x^t(A + vv^t - vv^t)x}{x^t x} \\
&\geq \min_{V_{k-1}} \max_{x \perp V_{k-1}, x \perp v} \frac{x^t(A + B - vv^t)x}{x^t x} \\
&= \min_{V_{k-1}} \max_{x \perp (V_{k-1} \oplus \langle v \rangle)} \frac{x^t(A+B)x}{x^t x} \\
&\geq \min_{V_k} \max_{x \perp V_k} \frac{x^t A x}{x^t x} = \lambda_{k+1}(A+B).
\end{aligned}$$

The inequality in the second line of this display holds because we added a new constraint to the maximization problem. The second inequality holds because in the constraint we used arbitrary V_k instead of those V_k that required to include v.

□

A closely related result is as follows.

Theorem 10.2.7 (interlacing of eigenvalues II)**.** *Let A be a Hermitian $n \times n$ matrix, and let A' be its $(n-1) \times (n-1)$ upper-left principal submatrix.*

$$\lambda_1(A) \geq \lambda_1(A') \geq \lambda_2(A) \geq \lambda_2(A') \geq \ldots \geq \lambda_{n-1}(A') \geq \lambda_n(A).$$

(In fact the result holds for any A' which by removing k-th column and k-th row from A, where $1 \leq k \leq n$.)

Example 10.2.8. The matrix

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

has eigenvalues $\lambda_1(A) = 2+\sqrt{2}$, $\lambda_2(A) = 2$, and $\lambda_3(A) = 2-\sqrt{2}$, and matrix

$$A' = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

has eigenvalues $\lambda_1(A') = 3$ and $\lambda_2(A') = 1$, and as the theorem claims, we have:

$$2 + \sqrt{2} \geq 3 \geq 2 \geq 1 \geq 2 - \sqrt{2}.$$

It is also interesting that if the eigenvalues of $n \times n$ symmetric matrices A and B are known, and n is large, then one can calculate approximately the distribution of eigenvalues of the matrix $A + UBU^*$, where U is a random unitary matrix. This was one found recently (around 20 years ago) in research that comprised the study of random matrices and results from a field in functional analysis called free probability theory.

10.3 Power iteration for calculation of the largest eigenvalue

Let A be an $n \times n$ symmetric matrix. The idea of power iteration is that the linear transformation A has the largest stretch in the direction of the eigenvector v_1 corresponding to the eigenvalue λ_1 that has the largest absolute value. So if we start with an arbitrary vector x and apply A^n to this vector, then the result will be close to a multiple of eigenvector v_1.

This is the basis for the following algorithm called the "power iteration".

- Start with a unit vector $v^{(0)}$. Let $\lambda^{(0)} = (v^{(0)})^* A v^{(0)}$.
- For $k = 1, 2, \ldots$, repeat the following steps until the change in $\lambda^{(k)}$ is small enough.
    1. Apply A: $w = Av^{(k)}$.

2. Normalize: $v^{(k)} = w/\|w\|$.
3. Calculate the Rayleigh quotient: $\lambda^{(k)} = (v^{(k)})^* A v^{(k)}$.

Theorem 10.3.1. *Let A be an $n \times n$ real symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$ and a corresponding orthonormal system of real eigenvectors $q_1, \ldots, q_n$. Suppose that $|\lambda_1| > |\lambda_2| \geq |\lambda_n| \geq 0$ and $q_1^* v^{(0)} \neq 0$. Then, as $k \to \infty$,*

$$\|v^{(k)} - (\pm q_1)\| = O\Big(\big|\frac{\lambda_2}{\lambda_1}\big|^k\Big),$$
$$\big|\lambda^{(k)} - \lambda_1\big| = O\Big(\big|\frac{\lambda_2}{\lambda_1}\big|^{2k}\Big).$$

The $\pm$ sign means that at each step k, one or the other choice of sign is to be taken, and then the indicated bound holds.

Note that convergence rate is $2k$ in the second bound.

Sketch of proof. We expand v^0 in the basis of eigevectors: $v^{(0)} = a_1 q_1 + a_2 q_2 + \ldots a_n q_n$, where $a_1 = q_1^t v^{(0)} \neq 0$ by assumption. Vector $v^{(k)}$ is a multiple of $A^k v^{(0)}$, so for some constants c_k we have:

$$\begin{aligned} v^{(k)} &= c_k(a_1 \lambda_1^k q_1 + \ldots + a_n \lambda_n^k q_n) \\ &= c_k(a_1 \lambda_1^k q_1 + \ldots + a_n \lambda_n^k q_n) \\ & c_k a_1 \lambda_1^k \Big(q_1 + \frac{a_2}{a_1}(\frac{\lambda_2}{\lambda_1})^k q_2 \ldots + \frac{a_n}{a_1}(\frac{\lambda_n}{\lambda_1})^k q_n\Big). \end{aligned}$$

The first claim of the lemma follows from this identity and the fact that both $v^{(k)}$ and q_1 have the unit length.

The second claim from the first one after the following estimate is proved: as $x \to q_1$,

$$\|R(x) - R(q_1)\| = O(\|x - q_1\|^2).$$

For the proof of this statement, see the Trethefen-Bau book. □

10.4 Inverse iteration method

The power iteration method is suitable for calculating the largest eigenvalue and the corresponding eigevector. What if we want to find other eigenvalues and eigenvectors?

For the smallest eigevalue we can do a simple adjustment and we get the *inverse iteration* method. Namely, in step 1, instead of calculating $w = Av^{(k)}$, we calculate $w = A^{-1}v^{(k)}$. Then, the sequence of vectors $v^{(1)}, v^{(2)}, \ldots,$ will converge to the eigenvector which corresponds to the largest eigenvalue of A^{-1}, which is $1/\lambda_n$. And the Rayleigh quotient $R(v^{(k)})$ converges to λ_n.

From the technical viewpoint, we usually do not invert A to calculate $w = A^{-1}v^{(k)}$ but rather solve the equation $Aw = v^{(k)}$ for w. All other steps are the same as in the original power iteration method.

This method can be generalized to estimation of other eigenvalues, provided that we have an estimate μ which is closer to λ_k than to other eigenvalues. Then the largest (in absolute value) eigenvalue of $(A - \mu I)^{-1}$ is $(\lambda_k - \mu)^{-1}$, and we can use $w = (A - \mu I)^{-1}v^{(k)}$ as the first iteration step of the algorithm. This is the *inverse iteration with a shift*.

In fact this method can even be used to improve the convergence of the power method for the largest eigenvalue, since it can potentially amplify the distance of the largest eigenvalue from other eigenvalues.

10.5 QR method for eigenvalue calculation

The QR method for eigenvalue calculation is a very interesting method for finding eigenvalues which is based on the following result.

Theorem 10.5.1. *Suppose that $A = QR$ is the QR factorization of a real symmetric matrix A, and let $A_1 = RQ$. Then A_1 is real symmetric and it has the same eigenvalues as A.*

Proof. Since Q is orthogonal, we can express R as $R = Q^*A$. If we plug this expression in the definition of A_1 we find $A_1 = Q^*AQ$ which implies the claims of the theorem. □

Then we can define matrices A_k recursively (with $A = A_0$). For $k = 1, 2, \ldots,$ if $A_{k-1} = Q_kR_k$, then we define

$$A_k = R_kQ_k = Q_k^*A_{k-1}Q_k.$$

For example $A = A_0 = Q_1R_1$ and so $A_1 = R_1Q_1 = Q_1^*AQ_1$. Then

$$\begin{aligned}
&A_1 = Q_2R_2,\\
&A_2 = R_2Q_2 = Q_2^*A_1Q_2 = Q_2^*Q_1^*AQ_1Q_2,\\
&\quad\ldots\\
&A_k = R_kQ_k = Q_k^*A_{k-1}Q_k = Q_k^*Q_{k-1}^*\ldots Q_1^*AQ_1\ldots Q_{k-1}Q_k,\\
&\quad\ldots,
\end{aligned}$$

By the previous theorem all A_k are real symmetric and have the same eigenvalues as A.

A useful identity that follows from the definitions is that $A_{k-1}Q_k = Q_kA_k$.

What we want to show is that if eigenvalues of A are all positive and distinct, then A_k converges to a diagonal matrix. In particular, the diagonal entries of A_k converge to the eigenvalues of A.

We will not prove this statement in detail but give some ideas why it is true.

Note that we have

$$A_k = (Q^{(k)})^* A Q^{(k)},$$

where $Q^{(k)} = Q_1Q_2 \dots Q_k$. Define also

$$R^{(k)} = R_k R_{k-1} \dots R_1$$

Theorem 10.5.2. *The matrices $Q^{(k)}$ and $R^{(k)}$ give the QR decomposition of the k-th power of the matrix A^k,*

$$A^k = Q^{(k)} R^{(k)}.$$

Proof. For $k = 1$, this simply means that $A = QR$. For large k, we proceed by induction. Suppose that the we already know that $A^{k-1} = Q^{(k-1)}R^{(k-1)}$. Multiply this equality by A on the left and use the identity $A_{k-1}Q_k = Q_kA_k$

$$\begin{aligned} AQ^{(k-1)} &= A_0Q_1Q_2 \dots Q_{k-1} = Q_1A_1Q_2 \dots Q_{k-1} \\ &= Q_1Q_2A_2 \dots Q_{k-1} \\ &\dots \\ &= Q_1Q_2Q_{k-1}A_{k-1} = Q^{(k-1)}Q_kR_k \end{aligned}$$

This implies that $A^k = Q^{(k)}R^{(k)}$. □

This theorem implies that the columns of $Q^{(k)}$ form the basis of the column space of A^k obtained as a result of Gram-Schmidt orthogonalization.

We can write $A^k = U\Lambda^k U^*$, where U is the matrix of eigenvectors of A and Λ is the diagonal matrix of eigenvalues. If the eigenvalues of A are all positive and distinct, then the columns of matrix A^k are linear combinations of eigenvectors u_i,

$$b_1\lambda_1^k u_1 + \dots + b_n\lambda_1^k u_n.$$

Moreover, if we assume $\lambda_1 > \lambda_2 > \ldots > \lambda_n$, then all columns of A^k, including the first one, are close to the vectors proportional to the eigenvector u_1. Hence the first column of $Q^{(k)}$ is also very close to u_1 The idea is that after orthogonalization, the second column of $Q^{(k)}$ will be close to the second eigenvector u_2, and so on.

Indeed, the second flag space V_2, spanned by the first and the second columns of A^k, is close to the space $\widehat{V}_2$ spanned by the first and the second eigenvectors u_1 and u_2. So, the second column of the matrix $Q_{(k)}$ obtained from the orthogonalization of the flag $V_1 \subset V_2$ will be close to the vector u_2. (This is because the first column of $Q_{(k)}$, as was just argued, is close to u_1 and u_2 is the only vector in $\widehat{V}_2$ orthogonal to u_1.)

In summary, the orthogonal matrix $Q^{(k)}$ will be close to the orthogonal matrix of eigenvectors of U and therefore $(Q^{(k)})^* U \approx I$. Then we have the eigenvalue decomposition $A = U\Lambda U^*$, and so we can write:

$$A_k = (Q^{(k)})^* A Q^{(k)} = (Q^{(k)})^* U \Lambda U^* Q^{(k)} \approx \Lambda.$$

That is, A_k is close to Λ.

The detailed implementation of this plan is omitted.

Chapter 11

Covariances and Multivariate Gaussian Distribution

11.1 Covariance of a linearly transformed vector

Suppose $x = (x_1, \ldots, x_m)^*$ be a column vector of random variables x_i. Then the covariance matrix C of x is the $m \times m$ matrix of covariances of the r.v.'s x_i:

$$C_{ij} = \mathrm{Cov}(x_i, x_j)$$

We will denote this matrix by $\mathbb{V}\mathrm{ar}(x)$. For example, if x_i are i.i.d random variables with variance σ^2, then the covariance matrix is a multiple of the identity matrix:

$$C \equiv \mathbb{V}\mathrm{ar}(x) = \sigma^2 I_{m\times m}$$

Obviously, the covariance matrix is symmetric. It has also another important property. First, let us define a symmetric positive definite matrix as a symmetric matrix that has the following property: $(x, Ax) = x^* Ax > 0$ for all real vectors $x \neq 0$. If a symmetric matrix $(x, Ax) \geq 0$ for all x then it is called non-negative definite. (Similar concepts can be defined more generally for hermitian matrices.)

Theorem 11.1.1. *If v is a real random vector, then its covariance matrix C is non-negative definite.*

Proof. It is clear that the covariance matrix is symmetric. Let x be a non-

random vector. Then,

$$\begin{aligned}\mathbb{V}\mathrm{ar}(x^*v) = \mathbb{V}\mathrm{ar}\Big(\sum_{i=1}^m x_i v_i\Big) &= \sum_{i,j} x_i \mathrm{Cov}(v_i, v_j) x_j \\ &= x^* C x = (x, Cx)\end{aligned}$$

However, $\mathbb{V}\mathrm{ar}(x^*v) \geq 0$ by properties of variance. Hence, $(x, Cx) \geq 0$ for all x and therefore the matrix C is non-negative definite. □

The proof also shows that the matrix C is positive definite unless there is a linear combination of components of vector v that has zero variance.

Theorem 11.1.2. *Let x be a random m-vector with covariance matrix C, and suppose $y = Ax$, where A is an $n \times m$ non-random matrix. Then, the covariance matrix of vector y is ACA^*.*

Proof. We calculate:

$$\begin{aligned}\mathrm{Cov}(y_i, y_j) &= \mathrm{Cov}\Big(\sum_{k=1}^m A_{ik} x_k, \sum_{l=1}^m A_{jl} x_l\Big) \\ &= \sum_{k=1}^m \sum_{l=1}^m A_{ik} A_{jl} \mathrm{Cov}(x_k, x_l) \\ &= \sum_{k=1}^m \sum_{l=1}^m A_{ik} C_{kl} A_{jl} \\ &= (ACA^*)_{ij}\end{aligned}$$

□

Example 11.1.3 (Linear regression). Consider the linear statistical model

$$y = X\beta + \varepsilon, \tag{11.1}$$

where y is an m-vector, X is a non-random $m \times n$ matrix, β is a non-random n-vector, and ε is a random m-vectors. In the statistical setting y are m observations of a dependent variable, the columns of X are m observations of n independent (or explanatory) variables, β are unknown coefficients and ε are unknown error terms.

Assume that ε_i are i.i.d. with zero mean and variance σ^2, which we assume known for simplicity. The linear regression method gives the following estimator of β:

$$\widehat{\beta} = (X^*X)^{-1} X^* y. \tag{11.2}$$

This estimator is a random vector since y is a random vector. What is its covariance matrix?

Let us plugin equation (11.1) into (11.2):

$$\begin{aligned}\widehat{\beta} &= (X^*X)^{-1}X^*(X\beta + \varepsilon) \\ &= \beta + (X^*X)^{-1}X^*\varepsilon.\end{aligned}$$

The first term is non-random so it does not affect any of the covariances. So it is enough to calculate the covariance matrix of the second term. By applying Theorem 11.1.2 and using the fact that $\mathbb{V}\mathrm{ar}(\varepsilon) = \sigma^2 I_{m\times m}$, we get

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\widehat{\beta}) &= (X^*X)^{-1}X^*X(X^*X)^{-1} \\ &= \sigma^2(X^*X)^{-1}.\end{aligned}$$

What about the variance of the *fitted* values $\widehat{y}$?

For fitted values we have the formula:

$$\begin{aligned}\widehat{y} &= X(X^*X)^{-1}X^*y \\ &= X(X^*X)^{-1}X^*(X\beta + \varepsilon) \\ &= X\beta + X(X^*X)^{-1}X^*\varepsilon.\end{aligned}$$

So by applying Theorem 11.1.2, we find:

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\widehat{y}) &= X(X^*X)^{-1}X^*(\sigma^2 I)X(X^*X)^{-1}X^* \\ &= \sigma^2 X(X^*X)^{-1}X^*\end{aligned}$$

This formula can be used to write the variance of individual terms of $\mathbb{V}\mathrm{ar}(\widehat{y})$.

11.2 Eigenvalue and Cholesky factorizations of a covariance matrix

Theorem 11.1.2 implies that if a random vector x has an identity covariance matrix $C = I$, then the covariance matrix of Ax is $C = AA^*$.

Sometimes we are given a matrix C and want to find such A that $C = AA^*$. For example, one of the ways to generate a multivariate random Gaussian variable with m components and covariance matrix C is to generate m independent Gaussian variables with unit variance and multiply a vector of these variables by A. It is known that the resulting variable is Gaussian and Theorem 11.1.2 will ensure that it has the correct covariance matrix.

There are many factorizations $C = AA^*$. One is the eigenvalue factorization. Since C is symmetric, it has an eigenvalue decomposition:

$$C = U\Lambda U^*,$$

where U is an orthogonal matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues. Note that all eigenvalues of a non-negative definite matrix must be non-negative. Indeed, if $\lambda < 0$ is a negative eigenvalue of C with eigenvector u, then $u^*Cu = -\lambda\|u\|^2 < 0$, which contradicts the assumption that C is non-negative.

So, in particular we can take a square root of Λ. The result is the matrix $\Lambda^{1/2}$ that has $\sqrt{\lambda^i}$ on its diagonal. Then we can use matrix $A = U\Lambda^{1/2}$ to factorize C as $C = AA^*$.

Another factorization is particularly popular in practice because it is very simple to calculate.

Definition 11.2.1. The *Cholesky factorization* of a self-adjoint matrix C is a decomposition

$$C = RR^*,$$

where R is a lower-triangular matrix.

We have already considered this factorization previously in Section 8.1 and know that for positive definite matrices the Cholesky factorization always exists. (It is also unique, see Theorem 23.1 in Bao - Trefethen.)

11.3 Multivariate Gaussian distribution

Definition 11.3.1. Let μ be an m-vector and Σ a positive definite $m \times m$ real symmetric matrix. The multivariate normal random variable with parameters μ and Σ is a random m-vector X with the following density function:

$$f_X(x) = \frac{1}{(2\pi)^{m/2}(\det\Sigma)^{1/2}} \exp\Big[-\frac{1}{2}(x-\mu)^*\Sigma^{-1}(x-\mu)\Big] \qquad (11.3)$$

The density is called the *Gaussian density* and it is ubiquitous in statistics and in statistical physics.

Remark 1: here and in the following we use the convention that random variables are denoted by upper case roman letters, while their realizations by lower case letters. This is in some conflict with our previous practice

when we used uppercase letters to denote matrices and lowecase letters to denote vectors.

Remark 2: One can define a multivariate normal distribution in a more general sense, when Σ may have a non-trivial null-space. Then one defines $K = \Sigma^+$, the pseudo-inverse of matrix Σ and the density is

$$f_X(x) = \frac{(\det K)^{1/2}}{(2\pi)^{m/2}} \exp\Big[-\frac{1}{2}(x-\mu)^* K (x-\mu)\Big], \tag{11.4}$$

if $x - \mu \in \text{Range}(K)$ and $f_X(x) = 0$ if $x - \mu \in \text{Null}(K)$. This is useful for describing singular normal random vectors, for which the variances of some linear combinations of the components of X are zero.

The matrix $K = \Sigma^+$ is often called the concentration matrix. It useful even if Σ is invertible and $\Sigma^+ = \Sigma^{-1}$.

Theorem 11.3.2. *The function $f_X(x)$ in (11.3) is a valid probability density function and the expectation and variance of the random vector X are μ and Σ, respectively.*

Proof. Let V be a random m-vector whose components are independent standard normal random variables. By independence, its density is the product of the densities of the components:

$$f_V(v) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2} v_i^2 \Big) = \frac{1}{(2\pi)^{m/2}} \exp\Big(-\frac{1}{2} v^\star v \Big)$$

Now let $\Sigma = RR^*$ be the Cholesky factorization of Σ, and let $X = \mu + RV$. Then $\mathbb{E}X = \mu$ and by Theorem 11.1.2, $\mathbb{V}\text{ar}(X) = RI_{m\times m}R^* = \Sigma$.

In order to calculate the density function for X, we note that $V = (R)^{-1}(X-\mu)$, and therefore,

$$\begin{aligned} -\frac{1}{2} v^* v &= -\frac{1}{2}(x-\mu)^* (R^*)^{-1} R^{-1} (x-\mu) \\ &= -\frac{1}{2}(x-\mu)^* (RR^*)^{-1} (x-\mu) \\ &= -\frac{1}{2}(x-\mu)^* \Sigma^{-1} (x-\mu). \end{aligned}$$

Next we note that the transformation $v = R^{-1}(x-\mu)$ is one-to-one and linear, and that the matrix of derivatives for this transformation is

$$\frac{\partial v(x)}{\partial x} := \left[\frac{\partial v_i(x)}{\partial x_j} \right]_{i,j=1,\ldots,m} = R^{-1}.$$

Hence the Jacobian of this transformation is $|\det R^{-1}| = |\det R|^{-1}$. On the other hand $\det \Sigma = \det R^* \det R = |\det R|^2$. It follows that the Jacobian of the transformation $v = R^{-1}(x - \mu)$ is $(\det \Sigma)^{-1/2}$.

The by the general theorem about the density function for transformed random variables, we find that the density function of the random vector X is

$$f_X(x) = \frac{1}{(2\pi)^{m/2}(\det \Sigma)^{1/2}} \exp\Big[-\frac{1}{2}(x-\mu)^* \Sigma^{-1}(x-\mu)\Big]$$

and this completes the proof of the theorem. □

Theorem 11.3.3. *Let X be a multivariate normal m-vector with zero mean and variance Σ. Then, for every non-random m-vector v:*

$$\mathbb{E}\exp(v^*X) = \exp\Big(\frac{1}{2}v^*\Sigma v\Big)$$

Before doing the general proof, let us look at the one-dimensional case when X is a usual zero mean normal random variable with variance σ^2. In this case, v is a scalar and we can calculate:

$$\begin{aligned}
\mathbb{E}\exp(vX) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\Big(vx - \frac{x^2}{2\sigma^2}\Big)\, dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\Big(-\frac{(x-\sigma^2 v)^2 - \sigma^4 v^2}{2\sigma^2}\Big)\, dx \\
&= \exp\Big(\frac{\sigma^2 v^2}{2}\Big) \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\Big(-\frac{(x-\sigma^2 v)^2}{2\sigma^2}\Big)\, dx \\
&= \exp\Big(\frac{\sigma^2 v^2}{2}\Big),
\end{aligned}$$

where the last integral is computed by the change of variable $y = x - \sigma^2 v$.

Proof. We need to calculate the multiple integral

$$\mathbb{E}\exp(v^*X) = c \int_{\mathbb{R}^m} dx\, \exp\Big[-\frac{1}{2}x^*\Sigma^{-1}x + v^*x\Big], \tag{11.5}$$

where

$$c = \frac{1}{(2\pi)^{m/2}(\det \Sigma)^{1/2}}.$$

Let $\Sigma^{-1} = Q\Lambda Q^*$ be the eigenvalue decomposition of Σ^{-1} with an orthogonal matrix Q and a diagonal positive definite matrix Λ with diagonal entries $(\lambda_1, \dots \lambda_m)$. Define $y = Q^*x$. Since $|\det Q| = 1$, the Jacobian of the transformation is 1 and the integral (11.5) can be written as:

$$c\int_{\mathbb{R}^m} dy\, \exp\Big[-\frac{1}{2}y^*Q^*\Sigma^{-1}Qy + v^*Qy\Big]$$
$$= c\int_{\mathbb{R}^m} dy\, \exp\Big[-\frac{1}{2}y^*\Lambda y + u^*y\Big],$$

where $u = Q^*v$.

Note that in coordinates

$$-\frac{1}{2}y^*\Lambda y + u^*y = \sum_{i=1}^{m}\Big[-\frac{1}{2}y_i^2\lambda_i + u_iy_i\Big],$$

so the multiple integral splits into a product of one-dimensional integrals, which we have already done. (We need only to set $\sigma_i^2 = 1/\lambda_i$.) So we calculate the integral as

$$c\prod_{i=1}^{m}\sqrt{2\pi}\lambda_i^{-1/2}\exp\Big[\frac{u_i^2}{2\lambda_i}\Big]$$

Product of λ_i equals $\det\Lambda = \det\Sigma^{-1}$. Hence,

$$c\prod_{i=1}^{m}\sqrt{2\pi}\lambda_i^{-1/2} = c(2\pi)^{m/2}(\det\Sigma)^{1/2} = 1.$$

And

$$\prod_{i=1}^{m}\exp\Big[\frac{u_i^2}{2\lambda_i}\Big] = \exp\Big[\frac{1}{2}(Q^*v)^*\Lambda^{-1}Q^*v\Big]$$
$$= \exp\Big[\frac{1}{2}v^*Q\Lambda^{-1}Q^*v\Big] = \exp\Big[\frac{1}{2}v^*\Sigma v\Big]$$

□

Essentially this result gives the moment-generating and characteristic functions of the multivariate normal distribution.

Corollary 11.3.4. *Let X be a multivariate normal m-vector with zero mean and variance Σ, and let $t = [t_1, \ldots, t_m]^*$ be a vector in $\mathbb{R}^m$. Then the moment generating function of X is*

$$m_X(t) := \mathbb{E}e^{t^*X} = \exp\Big(\frac{1}{2}t^*\Sigma t\Big),$$

and the characteristic function of X is

$$\varphi_X(t) := \mathbb{E}e^{i(t^*X)} = \exp\Big(-\frac{1}{2}t^*\Sigma t\Big)$$

By using the moment-generating function, we can calculate the moments of the multivariate normal distribution. The following result was proved by Leon Isserlis in 1918. Recently, it was made popular by particle physicists under the name Wick's theorem. The physicists used it in the perturbative Quantum Field Theory and Statistical Field Theory.

Theorem 11.3.5 (Wick's theorem)**.** *Let $X = (x_i)$ be a multivariate normal m-vector with zero mean and variance Σ. Then,*

$$\mathbb{E}\big(x_{i_1}x_{i_2}\ldots x_{i_k}\big) = \sum \Sigma_{ab}\ldots\Sigma_{yz},$$

where the sum is over all different pairings $(ab), \ldots, (yz)$ of the set of indices $\{i_1, i_2, \ldots, i_k\}$.

An example should make this statement more clear. For two indices, we simply have $\mathbb{E}(x_ix_j) = \Sigma_{ij}$. For four indices, we have:

$$\mathbb{E}(x_ix_jx_kx_l) = \Sigma_{ij}\Sigma_{kl} + \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}.$$

Proof of Theorem 11.3.5. By a well-known result, we can write the moment as the multiple derivative of the moment generating function evaluated at zero:

$$\begin{aligned}\mathbb{E}\big(x_{i_1}x_{i_2}\ldots x_{i_k}\big) &= \frac{\partial^k}{\partial t_{i_1}\ldots\partial t_{i_k}} m_X(t)\Big|_{t=0} \\ &= \frac{\partial^k}{\partial t_{i_1}\ldots\partial t_{i_k}} \exp\Big(\frac{1}{2}t^*\Sigma t\Big)\Big|_{t=0}\end{aligned}$$

Consider first the derivative with respect to t_{i_1}. By the chain rule it gives

$$\Big(\sum_{j=1}^{m}\Sigma_{i_1j}t_j\Big)\exp\Big(\frac{1}{2}t^*\Sigma t\Big)$$

Further differentiations will act either on the sum or on the exponential. If they act on the exponential they generate new sums of the similar form as a factor. If they act on the sum, they generate as scalar factor.

Note, however, that one of the further differentiations must act on the sum. Otherwise, the evaluation $t = 0$ will set the result to zero. Let it be differentiation with respect to t_{i_s}. Then we have a pairing of i_1 with i_s and this pairing results in a factor $\Sigma_{i_1 i_s}$.

Quite similar we see that every differentiation either generate a new sum or is paired with a previous differentiation to reduce one of these sums to a scalar. □

Let us accept without proof two facts. First, that a linear transformation of a multivariate normal random vector is a multivariate normal, although perhaps in the generalized sense with the density as in (11.4). The second is that a multivariate normal distribution is completely determined by its mean and variance (even if the distribution is singular, in which case one should use $A = \Sigma^+$, the pseudo-inverse of Σ). Then, we have the following theorem.

Theorem 11.3.6. *Let X be a random m-vector with the normal distribution and let $\mathbb{E}X = \mu$, $\mathbb{V}\mathrm{ar}(X) = \Sigma$. Suppose that B is an $k \times m$ matrix and b is a (non-random) k-vector. Then $Y = BX + b$ has the normal distribution, and*

$$\mathbb{E}Y = b + B\mu,$$
$$\mathbb{V}\mathrm{ar}Y = B\Sigma B^*.$$

Proof. This result follows from the two facts that we stated before the theorem, and the calculation of the expectation and variance. In particular, variance can be computed by formula in Theorem 11.1.2. □

A consequence of this theorem is that the marginal distributions of the multivariate normal vector are normal.

Theorem 11.3.7. *Let X be a random m-vector with the normal distribution and let $\mathbb{E}X = \mu$, $\mathbb{V}\mathrm{ar}(X) = \Sigma$. Suppose $X = (X_1, X_2)^*$, where X_1 is a k-vector with $k < m$, and suppose $\mu = (\mu_1, \mu_2)$, where μ_1 is a k-vector, and*

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}, \end{bmatrix}$$

where Σ_{11} is a $k \times k$ matrix. Then X_1 is normally distributed k-vector with mean μ_1 and covariance matrix Σ_{11}.

Proof. This result follows from Theorem if we take $b = 0$ and

$$B = [I_{k\times k}, 0_{k\times(m-k)}],$$

that is B is a $k \times m$ matrix that consists of the $k \times k$ identity matrix followed by $m-k$ columns of zeros. Then $X_1 = BX$ and a calculation gives the expectation and variance of X_1 stated in the corollary. □

We can also derive a formula for conditional distributions. Recall that if X_1 and X_2 are two random variables with the joint density $f_{X_1,X_2}(x_1, x_2)$, then the conditional density of X_1 given $X_2 = x_2$ is defined as

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)},$$

where $f_{X_2}(x_2)$ is the marginal density of X_2. The conditional mean and variance of X_1 given $X_2 = x_2$ are calculated as mean and variance with respect to the conditional density $f_{X_1|X_2}(x_1|x_2)$.

It turns out that the conditional density of a normal multivariate distribution is also normal and there are nice formulas for the conditional expectation and variance.

Theorem 11.3.8. *Assume the notation of theorem 11.3.7 and let Σ_{22} be non-singular. Then the conditional distribution of X_1 given X_2 is normal with mean*

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2),$$

and variance

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Remark 1: The theorem is actually also valid for singular Σ_{22} if one uses the pseudo-inverse Σ_{22}^{+} instead of Σ_{22}^{-1}.

Proof. In principle, the calculation of the conditional density is straightforward from the definition. If random vector $X = [X_1, X_2]^*$ and its value is $x = [x_1, x_2]^*$, then

$$\begin{aligned} f_{X_1|X_2}(x_1|x_2) &\propto \exp\left((x-\mu)^*\Sigma^{-1}(x-\mu) - (x_2-\mu_2)^*\Sigma_{22}^{-1}(x_2-\mu_2)\right) \\ &\propto \exp\left((x-\mu)^*\Sigma^{-1}(x-\mu)\right). \end{aligned}$$

where symbol $\propto$ means "proportional to" and the coefficient of proportionality does not depend on x_1. Then it remains to invert the block matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}, \end{bmatrix}$$

and complete the square so that the result has the form

$$f_{X_1|X_2}(x_1|x_2) \propto \exp\left((x_1 - \mu_{1|2})^* \Sigma_{1|2}^{-1} (x_1 - \mu_{1|2}) \right)$$

This is possible to do and there are formulas for the inversion of the 2×2 block matrix Σ, which are called Schur's complement formulas. However, we will use only the fact that the resulting conditional density is normal and calculate the conditional expectation $\mu_{1|2}$ and variance $\Sigma_{1|2}$ in a different way.

First, X_1 and X_2 are two random vectors, define the covariance of these vectors as a matrix $C = \mathrm{Cov}(X_1, X_2)$ with entries

$$C_{ij} = \mathrm{Cov}\big((X_1)_i, (X_2)_j\big),$$

where $(X_1)_i$ and $(X_2)_j$ are the i-th and j-th components of the vectors X_1 and X_2, respectively.

Let $Z = X_1 + AX_2$, where $A = -\Sigma_{12}\Sigma_{22}^{-1}$. Then,

$$\begin{aligned} \mathrm{Cov}(Z, X_2) &= \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(AX_2, X_2) \\ &= \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0. \end{aligned}$$

So Z and X_2 are uncorrelated. (In fact, A was chosen precisely to ensure this property.) Crucially, for normal random variables this implies that the variables are also independent. It follows that

$$\begin{aligned} \mathbb{E}(X_1|X_2) &= \mathbb{E}(Z - AX_2|X_2) = \mathbb{E}(Z|X_2) - AX_2 \\ &= \mathbb{E}Z - AX_2 = \mu_1 + A\mu_2 - AX_2, \end{aligned}$$

and this gives the desired formula for the conditional expectation.

For the conditional variance we calculate,

$$\begin{aligned} \mathbb{V}\mathrm{ar}(X_1|X_2) &= \mathbb{V}\mathrm{ar}(Z - AX_2|X_2) \\ &= \mathbb{V}\mathrm{ar}(Z|X_2) + \mathbb{V}\mathrm{ar}(AX_2|X_2) - \mathrm{Cov}(Z, X_2)A^* - A\mathrm{Cov}(X_2, Z). \end{aligned}$$

The second term is equal to zero because AX_2 is not random given X_2. The third and fourth term are equal to zero because Z and X_2 are independent.

Finally, the first term equals to the unconditional variance $\mathbb{V}\mathrm{ar}(Z)$ again because Z and X_2 are independent. Therefore,

$$\begin{aligned}\mathbb{V}\mathrm{ar}(X_1|X_2) &= \mathbb{V}\mathrm{ar}(Z) = \mathbb{V}\mathrm{ar}(X_1 + AX_2)\\ &= \mathbb{V}\mathrm{ar}(X_1) + A\mathbb{V}\mathrm{ar}(X_2)A^* + \mathrm{Cov}(X_1, X_2)A^* + A\mathrm{Cov}(X_2, X_1)\\ &= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\end{aligned}$$

□

These formulas is also possible to write in terms of the concentration matrix. Let

$$K = \Sigma^{-1} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}.$$

Then for the conditional distribution of X_1 given X_2, we have formulas:

$$\begin{aligned}\mu_{1|2} &= \mathbb{E}(X_1|X_2) = \mu_1 - K_{11}^{-1}K_{12}(X_2 - \mu_2)\\ K_{1|2} &= \mathbb{V}\mathrm{ar}(X_1|X_2)^{-1} = K_{11}.\end{aligned}$$

This formulas can be obtained by manipulating formulas that express K_{11}, K_{12}, and K_{22} in terms of Σ_{11}, Σ_{12}, and Σ_{22}.

Example 11.3.9. Consider a 3-dimensional normal random vector $X = [X_1, X_2, X_3]^*$ with zero mean and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

Then, we can calculate the concentration matrix

$$K = \Sigma^{-1} = \begin{bmatrix} 3 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

The marginal distribution of (X_2, X_3) has the covariance and concentration matrices

$$\Sigma^{(23)} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \text{ and } K^{(23)} = \left(\Sigma^{(23)}\right)^{-1} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

The conditional distribution of (X_1, X_2) given X_3 has the concentration and covariance matrices

$$K^{(12|3)} = \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix} \text{ and } \Sigma^{(12|3)} = \left(K^{(12|3)}\right)^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}.$$

Simlarly, $\mathbb{V}\mathrm{ar}(X_1|X_2, X_3) = 1/K_{11} = 1/3$ and so on.

11.4 Exercises

Exercise 11.4.1. Suppose (X, Y) is a bi-variate normal vector with $\mu_X = \mu_Y = 0$, standard deviations $\sigma_X = \sigma_Y = 1$, and correlation $\rho = 1/2$. (Recall that ρ is defined as $\rho = \sigma_{XY}/(\sigma_X \sigma_Y)$.)

Find $\mathbb{P}(Y > 0 | X = 1)$.