Lecture Notes for the Linear Algebra for Statisticians (Math 530)

Vladislav Kargin

December 4, 2020

Contents

Comments:

The first part of the Lecture Notes borrows material from Trefethen-Bau "Numerical Linear Algebra". The section about Markov Chains uses material from J. Norris "Markov Chains". The section about the multivariate Gaussian distribution based on the material in books by Anderson and Muirhead.

In retrospect, I would perhaps add some additional material on (a) the rank-nullity theorem, (b) determinants, and (c) the matrix resolvent and its relation to Markov Chains.

Chapter 1

Matrix Products, Range, Nullspace, Rank, Inverse

1.1 Matrix products

Let x be an n-dimensional column vector with entries x_i, $i = 1, \ldots, n$ and A be a an $m \times n$ matrix with entries $A_{i,j}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$. That is, matrix A has m rows and n columns. 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 the dimension of the matrix objects as a subscript and then we have

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

This formula can be interpreted in several ways. First, one can think about x as an element of an n-dimensional vector space V written in a specific coordinate basis. Then the formula $b = Ax$ represents an action of a linear transformation A on vector x. This linear transformation sends vector space V to an m-dimensional subspace W, and the formula (1.1) explains how to calculate the coordinates of the image b in a basis of W from the coordinates of vector x in a basis of V.

Another interpretation of formula (1.1) looks at A as 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

$$b = \sum_{i=1}^{n} x_i a_i, \tag{1.2}$$

where we should remember that each x_i is a number and each a_i is an m-dimensional vector.

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. If A and C are matrices representing linear transformations, the $B = AC$ is a matrix that represents a composition of these linear transformations when C acts first, and A second.

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 with the coefficients in c_1, which is the first column of C, the second column of B is the linear combination with the coefficients in c_2 and so on. As a result we get n linear combinations and each of them is an l-column vector.

1.2 Range, Nullspace, Rank

The *range* of a matrix A, denoted Range(A) is the set of vectors that can be expressed as Ax for some x. It is easy to check that it is a linear space and that it is spanned by columns of matrix A. It is also called the *column space* of A.

The null-space of A is the set of vectors x such that $Ax = 0$. It is denoted Null(A) or ker(A).

The *column rank* of a matrix A is the dimension of its column space. One can also define the row space and row rank similarly. One of fundamental theorems in linear algebra is that

$$\text{column rank} \;=\; \text{row rank}.$$

In particular one can simply talk about the rank of a matrix A, denoted rank(A).[1]

The dimension of the nullspace is called *nullity*. Another fundamental theorem of linear algebra is that for an $m \times n$ matrix A,

$$\text{nullity} + \text{rank} \;= n,$$

that is the sum of dimensions of the range and the nullspace are equal to the number of columns.

It is clear that $\text{rank}(A) \leq \min\{n, m\}$. If $\text{rank}(A) = \min\{n, m\}$, we say that matrix A is of full rank.

What is the meaning of full rank? If $m \geq n$ then the matrix of full rank A has $\text{rank}(A) = n$ so $nullity(A) = 0$ and so A have a trivial null-space. The meaning of this is that 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 on the column space Range(A). In other words two different vectors must go to two different vectors.

In particular if $m = n$, that is, the matrix A is square, and the matrix A has full rank, then map A is a bijection of R^n on $R^m = R^n$ and we have an inverse transformation. The matrix of this transformation is called the *inverse* of matrix A and denoted A^{-1}.

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

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

For numerical application, 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 which you studied in the first linear algebra course[2] 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.

[2]Chapter 1, Section I of Hefferon

Chapter 2

Adjoint matrices, scalar product, and orthogonality

2.1 Transposition and adjoint operation

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.

In the situation when matrix A has complex entries, it is typically more useful to define an adjoint matrix $A^\star$, with the entry $(A^t)_{ij}$ equal to $\overline{A_{ji}}$ where the overline denotes complex conjugation. For real matrices $A^* = A^t$, so we will use notation A^* for both real and complex matrix.

For various problem, an especially important class of matrices is *self-adjoint* matrices $A = A^*$. For the real case they are usually called *symmetric* and for the complex case, – *hermitian*.

2.2 Scalar product and vector norms

The *scalar product* (also called *dot product* or *inner product*) of two m-column vectors x and y is the matrix product of x^* and y.

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

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

From the course of linear algebra we know that the length of the vector u, – which we call its *norm* and denote $\|u\|$, – that it can be computed in terms of its inner product with itself:

$$\|u\| = \sqrt{u^* u}.$$

(I will sometime write $|u|$ instead of $\|u\|$ if no confusion can arise.)

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 $u = 0$.

There are other norms besides the usual norm that we described above. For example, a p-norm is defined for every $p \geq 1$. If $x \in \mathbb{R}^n$, then

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

This is an exercise that this function is indeed a norm. (One can check that if $p < 1$, then this function 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 $B = \{x : \|x\| \leq 1\}$ must be convex. And the second is to check that if $p < 1$, then the unit ball is not convex.)

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

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

In this notation our usual 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 and not any other p-norm and so we will skip this subscript.

The great advantage of the 2-norm is that it comes from the scalar product operation. 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.

2.3 Orthogonality

Two vectors are called *orthogonal* if they are both non-zero and their scalar product is zero.

A set of vectors $u_1, \ldots, u_n$ is called orthonormal if each of these vectors have length 1 and the vectors are orthogonal to each other (that is, $u_i^* u_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta symbol: $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.)

The useful thing about the systems of orthogonal vectors is that we can use them to decompose an arbitrary vector in orthogonal components.

Theorem 2.3.1. *The vectors an orthogonal set 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. □

In addition, we have the following result.

Theorem 2.3.2. *Let $\{u_1, \ldots u_n\}$ 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 $r - r', u_1, \ldots, u_n$ is linearly dependent. Both are not possible by Theorem 2.3.1.

□

A matrix is called *orthogonal* if:
(i) it is square, 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.

Theorem 2.3.2 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 $c = Q^* v$.

Exercise 2.3.3. A matrix Q is orthogonal (unitary) if and only if

$$Q^* Q = I,$$

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

Chapter 3

Matrix norms

We talked about the vector norm. We also want to measure norms of matrices. This is a big topic. Even in the case of vectors, there are many ways to define the norm besides the standard one. In the case of matrices the choice is even larger.

There are two most popular ways to define a norm of a matrix. The first one is called the *Frobenius norm* and it is defines 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 operation on square matrices: $\mathrm{Tr}M = \sum_{i=1} nM_{ii}$.

The Frobenius norm of A is 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: $\langle A, B\rangle = \mathrm{Tr}(A^*B)$. It is easy to calculate the Frobenius norm but, unfortunately, it is not as meaningful as another norm which is called the *operator norm* and which is define by the following formula:

$$\|A\| := \sup_{v\neq 0}\frac{|Av|}{|v|} = \sup_{v:|v|=1}|Av|. \tag{3.1}$$

This norm shows what is the maximum increase in the length of a vector that can be achieved by the transformation coded by matrix A. This is an obviously useful quantity but it is more difficult to calculate.

For matrix norms, sometimes some additional requirements are imposed on norms besides the usual properties of norms in vector spaces. These requirement are related to additional operations on matrices such taking the adjoint 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 text for derivation.)

Another important property of these two norms is that they are invariant relative to unitary transformations.

Theorem 3.0.1. *For every $m \times n$ matrix A and every unitary $m \times m$ matrix Q, we have*

$$\|QA\|_2 = \|A\|_2, \text{ and}$$
$$\|QA\|_F = \|A\|_F.$$

Chapter 4

Projectors

4.1 Definition and properties

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

Exercise 4.1.1. The range and nullspaces of P are invariant under P.

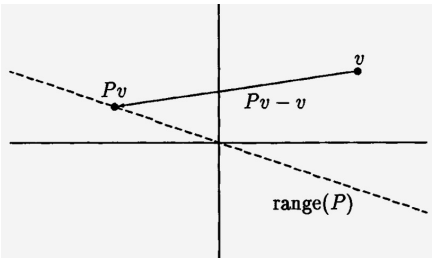

Figure 4.1: Oblique Projector

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

It is easy to see that the range of the complementary projector equals Null(P). For (i) $P(I - P)v = 0$, so Range$(I-P) \subset$ Null(P) and (ii) 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 sum of two vectors. One of them is from the range of P and another one is from the nullspace of P. This is useful since in many applications we want to separate a vector into two components, one of which is relevant and another is irrelevant. So if the relevant components form a linear space V and if we know how to write this V as a range of a projector P, then we can simply use the formula $v = Pv + (I - P)v$ to obtain the desired

decomposition.

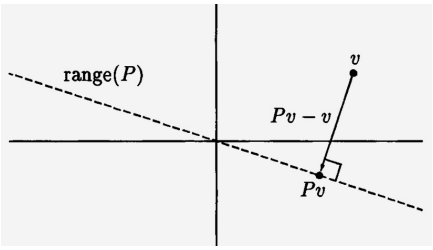

Figure 4.2: Ortogonal Projector

The most useful projectors are orthogonal projectors. An *orthogonal projector* P is a projector that has hermitian (or symmetric in the real case) matrix $P^* = P$.

It is called orthogonal because in this case its range and nullspace are orthogonal to each other. Indeed if $u_1 = Pv \in \text{Range}(P)$ and $u_2 \in \text{Null}(P)$, then

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

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 4.1.2. Suppose v is a column vector that has unit length. Then matrix $P = vv^*$ 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 $P^* = P$.

This projector is called a rank-one projector because its range is one-dimensional: it is spanned by the vector v.

Example 4.1.3. The previous example can be generalized. Suppose that 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 useful to write $P = QQ^*$ somewhat differently, as a sum of rank-one projectors.

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

In order to see this, the easiest way is to think about how P acts on vectors q_i. It is clear that Q^*q_i is the vector e_i that has 1 as its i-the component and 0 as all other components. So $QQ^*q_i = Qe_i = q_i$ for every i. This is exactly the same outcome as we obtain if we apply $\sum_{j=1}^n q_j q_j^*$ to q_i. Hence both operators act the same on the span of vectors $q_1, \dots q_n$. On all vectors orthogonal to the span of $q_1, \dots q_n$, both operators act by sending them to 0. This proves the desired equality.

But then it is easy to check that

$$P^2 = \Big(\sum_{i=1}^n q_i q_i^*\Big)^2 = \sum_{i=1}^n q_i q_i^* = P.$$

It is also clear that $P^* = P$. So, we proved that P is an orthogonal projector. It is also clear that this is a projector on the span of vectors $q_1, \dots q_n$.

Example 4.1.4. Now consider even more general case, when we want to project on a vector space V spanned by vectors $a_1, a_2, \dots, a_n$ which are not-necessarily orthogonal. So let matrix A has 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 is equivalent to requirement that columns a_i are linearly independent.)

First, by direct checking, $P^2 = P$ and $P^* = P$, so P is an orthogonal projection and we only need to check that it have correct range which should be V and 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

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

4.2 Relation to Least Squares Regression

In statistics we often need to solve the following problem:

$$y_i = \beta_1 x_i^{(1)} + \ldots \beta_n x_i^{(n)} + \varepsilon_i, \tag{4.1}$$

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.) The numbers ε_i are "error terms".

In statistics, 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 nature of ε_i. We simply assume that we observed y_i and $x_i^{(k)}$ 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 (4.1) 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

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

which leads to equations:

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

or

$$X^*X\beta = X^*y. \tag{4.2}$$

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 (4.2) as

$$\beta = (X^*X)^{-1}X^*y \tag{4.3}$$

Exercise 4.2.1. We have used in this derivation the differentiation of a function with respect to a vector β, which should be understood as that we differentiate with respect to each component of the vector and then put results of all differentiations in a vector. By writing the differentiations and the matrix product $(y - X\beta)^*(y - X\beta)$ in components, check that the first order condition equations are indeed as we wrote them.

The equations in (4.2) 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 predicted values of y as $\widehat{y} = X\beta$. Then

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

These are those linear combinations of explanatory random variables which minimize 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 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 operations. Since the matrix is symmetric and positive definite, this can be solved also by Cholesky factorization twice as fast. We will discuss the Cholesky factorization later.

Recently, there was a lot of interest when $m < n$, so that the number of explanatory variables exceeds the number of observations. In this case 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 $\|y-X\beta\|$ what is suggested is minimization of the "regularized problem",

$$\|y - X\beta\|^2 + \lambda\|\beta\|_1,$$

where λ is a regularization parameter and $\|\beta\|_1 = \sum_{i=1}^{n} |\beta_i|$ is the ℓ_1-norm of the parameter vector $\|\beta\|$. This is called the *lasso regression*. The main idea is to find a vector β which not only minimizes the error of the regression but that also has a lot of zeros as its components. We will not discuss the estimation details here.

Chapter 5

Singular Value Decomposition

5.1 Definition and the existence/uniqueness

A singular value decomposition (SVD) of an $m \times n$ matrix A is a way to write down this matrix as a product of two unitary (or orthogonal in the real case) matrices U and V^* and one diagonal matrix Σ.

Formally,

$$A = U\Sigma V^*, \tag{5.1}$$

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 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 the singular values. For a real matrix A all elements in these matrices can be chosen to be real.

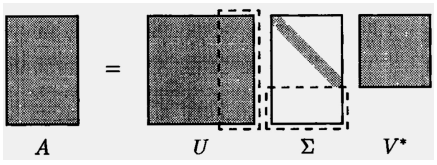

Figure 5.1: Full SVD decomposition, $m > n$

Intuitively, for $m \geq n$, if A represent a linear transformation, then we can write it as a rotation in $\mathbb{R}^n$, then a map that stretches the result and imbeds it isometrically to $\mathbb{R}^m$, and finally do another rotation in $\mathbb{R}^m$.

In particular this means 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

$$A = \widehat{U}\widehat{\Sigma}V^*,$$

where $\widehat{U}$ is an $m \times n$ matrix that has 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, it is an $n \times n$ unitary matrix.

In the reduced SVD, $\widehat{U}$ is not unitary since it is not square. 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.

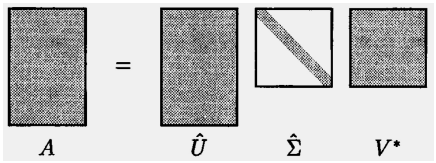

Figure 5.2: Reduced SVD decomposition, $m > n$

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 usually not happens in practice.

What is more essential is the question about the existence of the SVD decomposition.

Theorem 5.1.1. *Every matrix A has a singular value decomposition (??). 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 Trefethen's book. Here is a sketch of the proof of the existence claim.

Proof of the existence claim. For concreteness, let us work with real matrices.

By a compactness argument, the supremum in the definition of the matrix norm (3.1) 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 $\sigma_1 = \|A\|$ and 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 the 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. However, 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 $w = 0$.

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. □

5.2 Relation to eigenvalue decomposition

A (right) eigenvector of a square $n \times n$ matrix A is a unit vector v such that $Av = \lambda v$. The number λ is called an eigenvalue. Similarly, a left eigenvector u satisfies $uA = \lambda u$.

For arbitrary (square) matrices A right eigenvectors do not coincide with left eigenvectors, but the eigenvalues can be written as roots of a characteristic polynomial $\det(X - \lambda I)$ and so their set is defined unambiguously. There are always n eigenvalues if we count them as the roots of this polynomial counted with multiplicities.

Unfortunately, even if the matrix is real, the eigenvalues are typically complex.

In general, the number of eigenvectors can be smaller than the dimension of the space n. One simple example of this is the matrix

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

(However, this never happens if the eigenvalues are distinct.)

For arbitrary matrices, eigenvectors are not orthonormal. Still, if they form a basis, the matrix A is "diagonalizable", that is, it is possible to write the decomposition:

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

where X is the matrix whose columns are (right) eigenvectors and Λ is a diagonal matrix with eigenvalues on the main diagonal. This should be thought of as another form of the eigenvector equation:

$$AX = X\Lambda.$$

You can read more about the conditions that ensure that the matrix is diagonalizable in Chapter 5, Section II of Hefferon's book. For example the matrix is always diagonalizable if all eigenvalues are distinct (the characteristic polynomial does not have multiple roots).

Geometrical meaning of the eigenvalue decomposition is that we can think about a diagonalizable matrix A as a matrix that in a certain basis acts by multiplication of each component of a vector by the corresponding eigenvalue. If a vector $v = a_1x_1 + \ldots + a_nx_n$, where $x_1, \ldots, x_n$ are eigenvectors of A, then $Av = \lambda_1a_1x_1 + \ldots + \lambda_na_nx_2$.

The bad thing is that for general matrices both eigenvectors and eigenvalues are complex and the eigenvectors do not form an orthonormal basis. For this reason, the intuitive meaning of this multiplication by complex scalars is not very clear.

For example, a rotation matrix

$$\begin{bmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{bmatrix}.$$

is diagonalizable and also can be thought as a matrix of stretching by some complex numbers $e^{i\varphi}$.

However, the eigenvalue decomposition do have one very important benefit. Namely, it helps us define the functions of a diagonalizable matrix A. For example, we can see that

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

and for any polynomial $P(t)$,

$$P(A) = XP(\Lambda)X^{-1}.$$

Since every continuous function can be approximated by polynomials, this suggests that we can define

$$f(A) = Xf(\Lambda)X^{-1},$$

where $f(\Lambda)$ is obtained by applying function f to each diagonal element of Λ.

For examples, this allows define non-integer powers of matrices.

In general, this cannot be done with singular value decomposition since matrices U and V are not related to each other.

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 decomposition becomes very important when the 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 and matrix X becomes orthogonal. This is very close to the SVD decomposition and the difference is that some eigenvalues may be negative and the singular values must be non-negative.

Theorem 5.2.1. *If A is a self-adjoint $n \times n$ 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 the self-adjoint matrices, we have the eigenvalue decomposition:

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

where Λ and Q are diagonal and orthogonal matrices, respectively. We can easily convert it to the SVD decompositions by multiplying some of the columns by -1,

$$A = Q|\Lambda|\text{sign}(\Lambda)Q^*,$$

where $\text{sign}(\Lambda)$ is the diagonal matrix with the diagonal entries $\text{sign}(\lambda_i)$. This shows that $\sigma_i = \lambda_i$. $\square$

So in the SVD decomposition, $U = X$ and V equals to X with some of the columns multiplied by -1.

We will talk a bit more about the eigenvalue decomposition after we finish the SVD decomposition.

5.3 Properties of the SVD and singular values

Theorem 5.3.1. *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. $\square$

The operator and Frobenius norms of a matrix can be written in terms of its singular values.

Theorem 5.3.2. *Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ be non-zero singular values of matrix A. Then,*

$$\begin{aligned} \|A\|_2 &= \sigma_1, \\ \|A\|_F &= \sqrt{\sigma_1^2 + \ldots + \sigma_r^2}. \end{aligned}$$

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:

$$\begin{aligned}\|A\|_F^2 = \mathrm{Tr}(A^*A) &= \mathrm{Tr}\Big((V\Sigma^*U^*)(U\Sigma V^*)\Big) \\ &= \mathrm{Tr}\Big(V\Sigma^*\Sigma V^*\Big) = \mathrm{Tr}(\Sigma^*\Sigma),\end{aligned}$$

where the last step is by the property of the trace: $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

And the last quantity is easy to calculate:

$$\mathrm{Tr}(\Sigma^*\Sigma) = \sigma_1^2 + \ldots + \sigma_r^2.$$

□

Theorem 5.3.3. *The non-zero singular values of A are the square roots of the non-zero eigenvalues of A^*A or AA^*. (These matrices have the same non-zero eigenvalues.)*

Proof. Let the (full) singular value decomposition for A be

$$A = U\Sigma V^*.$$

Then,

$$A^*A = V(\Sigma^*\Sigma)V^*.$$

Since V is orthogonal and $\Sigma^*\Sigma$ is diagonal, therefore, we found an eigenvalue decomposition of A^*A, so, in particular, non-zero eigenvalues of matrix A^*A are non-zero elements of Σ, that is, the singular values of A. For A^*A, the proof is similar. □

Note that this gives another proof of Theorem 5.2.1, since for self-adjoint 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 5.3.3, singular values of A are equal to $\sqrt{\lambda_i^2} = |\lambda_i|$, absolute values of eigenvalues of A.

Now let us consider the relation of eigenvalues and singular values to the determinant. For eigenvalues, it is possible to prove that $\det(A) = \prod_{i=1}^n \lambda_i$. This is true for every matrix A, and it is very simple to prove in the case

when matrix A has an eigenvalue decomposition. In this case we can use the multiplicative property of the determinant and write:

$$\begin{aligned}\det(A) &= \det(X\Lambda X^{-1}) = \det(X)\det(\Lambda)\det(X)^{-1} \\ &= \det(\Lambda) = \prod_{i=1}^{n} \lambda_i.\end{aligned}$$

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 5.3.4. *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:

$$\begin{aligned}|\det(A)| &= |\det(U\Sigma V^*)| = |\det(U)||\det(\Sigma)||\det(V^*)| \\ &= |\det(\Sigma)| = \prod_{i=1}^{m} \sigma_i.\end{aligned}$$

In order to go to the second line, we used the fact that the determinant of a unitary matrix have absolute value 1. This fact holds because (i) $\det(U)\det(U^*) = \det(UU^*) = 1$, and (ii) $\det(U^*) = \det(U)^*$. Hence $|\det(U)|^2 = 1$, and therefore $|\det(U)| = 1$. □

5.4 Low-rank approximation via SVD

The SVD is useful because it allows us to construct low-rank approximations a matrix which are optimal in the Frobenius or operator norm.

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 5.4.1. *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 $\leq \nu$ 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, for the linear subspace V spanned by the first $\nu + 1$ singular vectors of A, 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.

5.5 Applications

5.5.1 Relation to Linear Regression

If the data matrix X has the reduced SVD

$$X = U\Sigma V^*,$$

then it is full rank if and only if all its singular values are positive. Then,

$$X^*X = V\Sigma U^*U\Sigma V^* = V\Sigma^2 V^*,$$

where V is the orthogonal matrix, with $V^{-1} = V^*$. Hence, the pseudo-inverse, which is an $n \times m$ matrix, can be written as follows.

$$\begin{aligned} X^+ &= (X^*X)^{-1}X^* = V\Sigma^{-2}V^*V\Sigma U^* \\ &= V\Sigma^{-1}U^*. \end{aligned}$$

This gives a method for solving the normal equations of the linear regression problem. Namely, calculate the singular value decomposition of X, and then calculate the pseudo-inverse by the previous formula. Then the solution of normal equations is given by

$$\beta = X^+ y.$$

It is easy to calculate Σ^{-1} because this is simply an $n \times n$ diagonal matrix with diagonal entries σ_1^{-1}, σ_2^{-1}, ..., σ_n^{-1}.

The most work 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.

5.5.2 Principal Component Analysis

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 a matrix X is the matrix of data. The rows of this matrix are observations and the columns are various variables or features of the observation. 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 is to analyze the empirical covariance matrix of the data: X^*X (we assume that the columns of X have zero mean.

The principal component analysis is the eigenvalue decomposition of the empirical covariance matrix:

$$X^*X = V\Lambda V^*$$

The eigenvectors (the columns of V) are interpreted as the linear combination of characteristics that have some significance in terms of the covariance matrix. For example, the eigenvector with the largest eigenvalue correspond to the linear combination of characteristics with the largest variation across individuals.

The eigenvectors with the largest eigenvalues are called the principal components. An interpretation of them is that these are factors that contain the most variability in the system.

From the linear algebra viewpoint this calculation to finding the singular values and right singular vectors in the SVD decomposition of the matrix X.

One example where this method is used is the data of financial stock returns. It turns out that the empirical covariance matrix exhibit three important factors (principal components with large corresponding singular value).

5.5.3 Factor analysis

The factor analysis takes into account not only the matrix V in the SVD decomposition of the data matrix X but also the matrix U.

Namely, we can write:

$$X = U\Sigma V^* = \sum_{k=1}^{r} \sigma_k u^{(k)} \left(v^{(k)}\right)^*,$$

where $u^{(k)}$ and $v^{(k)}$ are k-th columns of the matrices U and V, respectively.

In components, this can be written as

$$x_{tj} = \sum_{k=1}^{r} \sigma_k u_t^{(k)} v_j^{(k)}$$

where $t = 1, \ldots, m$ and $j = 1, \ldots, n$. Then vector $\sigma_k u_t^{(k)}$ is called the k-th factor (or the observations of the k-th factor) and the vector $v^{(k)}$ is the vector of factor loadings.

The interpretation is that fluctuation in factors capture the main sources of variations in the data and the factor loadings show how this variation is reflected in individual characteristics (say, in the returns of individual stocks.)

5.5.4 Face recognition

The SVD is used for face and writing recognition. This is a variant of factor analysis. 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.

The vectors u_i^k are eigenfaces, where "eigenface" means a singular vector corresponding to a sufficiently large singular value.

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.

5.5.5 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 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.

5.5.6 Other applications

– The SVD has some application in continuous mechanics and in robotics since it decomposes a matrix as a product of two rotations, which can be accomplished without stress and a stretching matrix.

– Eigenvalue decomposition is used in the spectral clustering algorithm.

Chapter 6

Eigenvalues and eigenvectors

6.1 Definition and properties

We already discussed some facts about eigenvalues and eigenvectors. In particular, in one of the assignments, you were asked to prove that the eigenvalues of a symmetric matrix are real and that if two eigenvalues of this matrix are different then the corresponding eigenvectors are orthogonal.

Here we give some more detail. First, recall that 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.

The characteristic polynomial of A is the polynomial $p_A(z)$ defined by $\det(zI - A)$.

A very important theorem connects eigenvalues and the characteristic polynomial.

Theorem 6.1.1. *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 vector x (its corresponding eigenvector), such that $(\lambda I - A)x = 0$. This happens if and only if the matrix $\lambda I - A$ is singular (that is, if its rank is smaller than its dimension). And here we can use a property of the determinant that the singularity of matrix $\lambda I - A$ is equivalent to $\det(\lambda I - A) = 0$. (Intuitively the determinant of a matrix equals to zero if and only if there is a linear dependence among the column vectors of this matrix.) □

From this theorem we immediately obtain the consequence 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. (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?

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

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 $X^{-1}AX$, and conversely. □

Theorem 6.1.3. *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 $A\widehat{V} = \lambda\widehat{V}$.

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 its geometric multiplicity. □

Example 6.1.4. 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.

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.

Theorem 6.1.5. *An $m \times m$ matrix A is non-defective if and only if it has an eigenvalue decomposition $A = X\Lambda X^{-1}$.*

Proof. If matrix A has an eigenvalue decomposition than it is similar to matrix Λ and hence has same eigenvalues with same multiplicities. Since it is easy to check that a diagonal matrix is non-defective, hence Λ is non-defective and the same holds for A.

In the converse direction, we can check that eigenvectors corresponding to different eigenvalues are linearly independent (exercise). If a 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, then we obtain the set of m linearly independent eigenvectors. If these m independent eigenvectors are formed into the columns of a matrix X, then X is nonsingular and we have $A = X\Lambda X^{-1}$. □

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

Theorem 6.1.6. *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.

□

One of the most important properties of every self-adjoint matrix is that it is possible to form a basis that consists of its eigenvectors and as a consequence they admit a unitary diagonalization:

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

where Q is an orthogonal matrix.

This is true if all eigenvalues are different since then all eigenvectors are orthogonal. In general, we will prove it 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 6.1.7. *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 6.1.8. *If $A^* = A$, then A admits unitary diagonalization:*

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

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.

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

Proof. For the case when all eigenvalues of matrices A and B are distinct, the proof proceeds by showing that if x is an eigenvector of A then Bx is also an eigenvector of A therefore 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*. We omit the proof. Intuitively, the eigenvalue matrices for A^*A and AA^* are matrices of left and right singular vectors for A. If they coincide we can conclude that the matrices U and V in the singular value decomposition are the same and the singular value decomposition becomes the eigenvalue decomposition.

Two examples of normal matrices are projection matrices and unitary matrices. For projection matrices all eigenvalues are either 0 or 1 and the corresponding eigenspaces are the nullspace and the range of the matrix. For the unitary matrices all eigenvalues must have unit absolute value since unitary matrices preserve the length of the vectors.

6.2 Applications

6.2.1 Difference equations

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 the initial conditions $x_{k-1}, x_{k-2}, \ldots, x_0$ 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 $x^{(i)} = [x_{k+i-1}, x_{k+i-2}, \ldots, x_i]^*$ 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

$$x^{(i+1)} = Ax^{(i)}. \tag{6.1}$$

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

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

If we know both Λ and the matrix of eigenvectors X we can write an explicit formula for $x^{(n)}$. Even if we don't know X, a typical situation is that the matrix Λ has a single eigenvalue λ_1 with the largest absolute value. If in addition we assume that the first component of $X^{-1}x^{(0)}$ is not zero, then the growth of $|x_n|$ is approximately $c|\lambda_1|^n$. In particular, if $|\lambda_1| < 1$ then the sequence declines to zero, and if $|\lambda_1| > 1$ then the sequence grows unboundedly.

Many other dynamic problems in biology, engineering and physics can be cast in the form (6.1) 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 6.2.1 (Fibonacci numbers). A classic example for this concept is the Fibonacci numbers, which are defined by the relation:

$$f_n = f_{n-1} + f_{n-2}.$$

Then we have $f^0 = 1, 0$ and

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

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

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

After some calculation one can get from this formula:

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

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 $832,040 + 2.4063 \times 10^{-7}$.

6.2.2 Power iteration as a tool to find the largest eigenvalue

In fact we can use the previous result in the converse direction to find the eigenvalue with the largest absolute value.

Suppose that a matrix A has simple eigenvalues $\lambda_1, \ldots \lambda_m$ with corresponding eigenvectors $x_1, \ldots, x_m$. If we start with an arbitrary vector v, which in the basis of $x_1, \ldots, x_n$ has the representation $v = c_1 x_1 + \ldots c_m x_m$, then

$$A^s v = c_1(\lambda_1)^s x_1 + \ldots c_m(\lambda_m)^s x_m$$

If λ_1 is the eigenvalue with the largest absolute value, then for large s this sum is dominated by the term $c_1(\lambda_1)^s x_1$. So for large s the vector A^s becomes very close to being proportional to the eigenvector x_1.

The algorithm works as follows. Start with a random unit vector v, and repeat the following steps.

1) calculate $u = Av$.

2) normalize u by computing unit vector $v' = u/|u|$.

3) if $|v' - v|$ is sufficiently small, calculate $\lambda = (Av, v)$. This is the desired eigenvalue with eigenvector equal to v. Otherwise, set $v = v'$ and repeat.

The other eigenvalues are found on the basis of this algorithm. One method is to calculate $B = (\alpha - A)^{-1}$. This maps an eigenvalue in the vicinity of α to the largest eigenvalue of B. Other methods and ideas can be found in the book by Trefethen and Bau.

6.2.3 Markov Chains

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

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

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

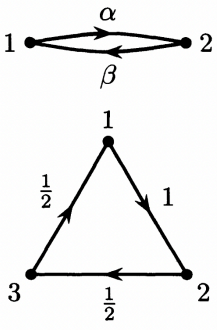

Figure 6.1

This is illustrated by diagrams in Figure 6.1.

The transition matrix P is a square $m \times m$ matrix 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.

One property of the stochastic matrices is that they always have an eigenvalue λ_1, which corresponds to the (right) eigenvector $v = [1, 1, \dots 1]$. However, the left eigenvectors are also interesting because the multiplication of a vector of probabilities on the right by matrix P corresponds to the evolution of the probability distributions.

Indeed, the definition of the transition probabilities allows us to calculate the joint distributions. For every sequence of states, $(x_0, \dots, x_n)$

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

In particular if we sum over all $x_0, \dots x_{n-1}$, 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.

It follows that

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

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

Example 6.2.2 (Random walk on a graph).

Recall that a *graph* $G = (V, E)$ is a set of vertices V and a set of edges E, which are simply a pair of vertices $E \subset V \times V$. We will usually assume that the graph is simple, that is, that there are no multiple edges (edges with the same endpoints) and that there are no loops, edges that have the same vertex as both endpoints. The edges (v_1, v_2) and (v_2, v_1) are not distinguished, so the graph is undirected. 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 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.

A distribution μ on a (countable) state S space is a non-zero vector with non-negative entries. We call it a probability distribution if the sum of the entries is 1.

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

$$\pi P = \pi$$

The terms *equilibrium* or *stationary* measure are also used to mean the same.

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

Note that $\sum_j P_{ij} = 1$ for all i, which means that matrix P has a *right* eigenvector with eigenvalue 1 that has all its entries equal to 1. From an algebraic viewpoint an invariant measure is a *left* eigenvector of the matrix P with eigenvalue 1. This gives us a practical method for computation of the invariant distribution if the state space is finite (and not too large).

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 there is always a left eigenvector with non-negative entries. One of the proves 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.

We formulate this theorem for a class of non-negative matrices called primitive. It can be generalized to a wider class of non-negative matrices, called irreducible. (I follow the book "Non-negative matrices" by Seneta here.)

Definition 6.2.3. A square non-negative matrix P is called primitive if for some positive integer k, all entries of the matrix P^k are positive.

Theorem 6.2.4. *Suppose P 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.

The other eigenvalues are also important. Indeed, if $\mu^(0)$ is an initial distribution on the state space than we can expand it in the basis of left eigenvectors of matrix P:

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

where π is the invariant distribution and v_k are other left eigenvectors. Then, 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 (1) $c_1 = 1$, and

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

that is, the distribution converges to the stationary distribution. and (2) if λ_2 is the eigenvalue that has the second-largest absolute value, and if $c_2 \neq 0$, then

$$|\mu^{(t)} - \pi| \sim |\lambda_2|^t |c_k v_k|.$$

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.

6.2.4 Reversible Markov Chains

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

$$\mu_j P_{ji} = \mu_i P_{ij} \tag{6.2}$$

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 $X_n, \ldots, X_0$. It turns out that reversible Markov chains are easier to understand than non-reversible chains.

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

$$DP = P^* D, \tag{6.3}$$

where D is the diagonal matrix with the entries $D_{ii} = \mu_i$.

The solution μ of the equation (6.2) is the invariant distribution.

Lemma 6.2.5. *If the probability distribution μ satisfy (6.2) , then μ is invariant.*

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

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

□

Often this property gives us a convenient tool for finding the invariant distribution of a chain.

Example 6.2.6 (Random walk on a graph).

Consider a graph G with vertices $v \in V$. A *degree* (or *valence*) 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. Here d_u denotes the degree of the vertex u. 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. It is easy to see from the detailed balance equation that the uniform distribution is invariant for this chain.

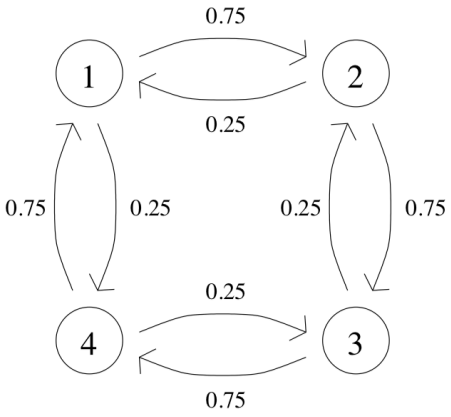

Figure 6.2: An example of a non-reversible chain: a random walk with a bias.

In Figure 6.2, a non-reversible chain is presented. It is clear from symmetry that the invariant distribution is uniform, but then the detailed balance equation is not satisfied: $P_{ji} \neq P_{ij}$.

Theorem 6.2.7. *The eigenvalues of the transition matrix P of a reversible Markov chain are real and P has the following eigenvalue decomposition:*

$$P = D^{-1/2} U \Lambda U^* D^{1/2}.$$

where D is a diagonal matrix whose diagonal entries are the elements of the invariant distribution, and U 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 (6.3) can be written as

$$\begin{aligned} D^{1/2}PD^{-1/2} &= D^{-1/2}P^*D^{1/2} \\ &= \left(D^{1/2}PD^{-1/2}\right)^*. \end{aligned}$$

In other words, the matrix

$$\widehat{P} = D^{1/2}PD^{-1/2}$$

is symmetric. Therefore, it has an orthogonal decomposition $U\Lambda U^*$ and its eigenvalues are real. Since P is similar to $\widehat{P}$ its eigenvalues are also real and it has decomposition

$$P = D^{-1/2}U\Lambda U^*D^{1/2}.$$

□

Chapter 7

Covariances and Multivariate Gaussian Distribution

7.1 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 7.1.1. *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 7.1.2. *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 7.1.3 (Linear regression). Consider the linear statistical model

$$y = X\beta + \varepsilon, \tag{7.1}$$

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:

$$\widehat{\beta} = (X^*X)^{-1}X^*y. \tag{7.2}$$

This estimator is a random vector since y is a random vector. What is its covariance matrix?

Let us plugin equation (7.1) into (7.2):

$$\begin{aligned}\widehat{\beta} &= (X^*X)^{-1}X^*(X\beta + \varepsilon)\\ &= \beta + (X^*X)^{-1}X^*\varepsilon.\end{aligned}$$

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 7.1.2 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 7.1.2, 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 $\mathbb{V}\mathrm{ar}(\widehat{y})$.

7.2 Eigenvalue and Cholesky factorizations of a covariance matrix

Theorem 7.1.2 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 7.1.2 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 7.2.1. The *Cholesky factorization* of a self-adjoint matrix C is a decomposition

$$C = RR^*,$$

where R is a lower-triangular matrix.

When does this factorization exist?

Theorem 7.2.2. *Every hermitian positive definite matrix has a unique Cholesky factorization.*

Proof. The proof is by induction. Let A be an hermitian positive definite matrix. First note that the fact that a matrix $A = (a_{ij})$ is positive definite implies that all diagonal elements are positive. Let $\alpha = \sqrt{a_{11}}$ and write the first step of the factorization:

$$\begin{aligned} A &= \begin{bmatrix} a_{11} & w^* \\ w & K \end{bmatrix} \\ &= \begin{bmatrix} \alpha & 0 \\ w/\alpha & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & K - ww^*/a_{11} \end{bmatrix} \begin{bmatrix} \alpha & w^*/\alpha \\ 0 & I \end{bmatrix} \end{aligned} \tag{7.3}$$

Indeed,

$$\begin{bmatrix} \alpha & 0 \\ w/\alpha & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & K - ww^*/a_{11} \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ w/\alpha & K - ww^*/a_{11} \end{bmatrix}$$

and one more multiplication verifies the equality in (7.3).

We can write this equality as $A = R_1 A_1 R_1^*$, where R_1 is lower-triangular. It is easy to check that A_1 is positive definite (exercise). This implies that its principal sub-matrix $K - ww^*/a_{11}$ is also positive definite. In particular, it is possible to apply the induction assumption that this submatrix has a Cholesky factorization $R_2 R_2^*$. Then if we define

$$\widehat{R}_2 = \begin{bmatrix} 1 & 0 \\ 0 & R_2 \end{bmatrix},$$

then we obtain the factorization $A = R_1 \widehat{R}_2 \widehat{R}_2^* R_1^*$. This is the desired factorization with the lower-triangular $R = R_2 R_1$.

For the uniqueness, see the proof of Theorem 23.1 in Bao - Trefethen. □

7.3 Multivariate Gaussian distribution

Definition 7.3.1. 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:

$$f_X(x) = \frac{1}{(2\pi)^{m/2}(\det \Sigma)^{1/2}} \exp\Big[-\frac{1}{2}(x-\mu)^* \Sigma^{-1}(x-\mu)\Big] \tag{7.4}$$

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

$$f_X(x) = \frac{(\det K)^{1/2}}{(2\pi)^{m/2}} \exp\Big[-\frac{1}{2}(x-\mu)^* K(x-\mu)\Big], \tag{7.5}$$

if $x - \mu \in \mathrm{Range}(K)$ and $f_X(x) = 0$ if $x - \mu \in \mathrm{Null}(K)$. 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 7.3.2. *The function $f_X(x)$ in (7.4) 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 7.1.2, $\mathbb{V}\mathrm{ar}(X) = RI_{m\times m}R^* = \Sigma$.

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 7.3.3. *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:

$$\begin{aligned}\mathbb{E}\exp(vX) &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\Big(vx-\frac{x^2}{2\sigma^2}\Big)\,dx\\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\Big(-\frac{(x-\sigma^2v)^2-\sigma^4v^2}{2\sigma^2}\Big)\,dx\\ &= \exp\Big(\frac{\sigma^2v^2}{2}\Big)\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\Big(-\frac{(x-\sigma^2v)^2}{2\sigma^2}\Big)\,dx\\ &= \exp\Big(\frac{\sigma^2v^2}{2}\Big),\end{aligned}$$

where the last integral is computed by the change of variable $y = x - \sigma^2 v$.

Proof. We need to calculate the multiple integral

$$\mathbb{E}\exp(v^*X) = c\int_{\mathbb{R}^m} dx\,\exp\Big[-\frac{1}{2}x^*\Sigma^{-1}x + v^*x\Big], \tag{7.6}$$

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 (7.6) can be written as:

$$\begin{aligned}&c\int_{\mathbb{R}^m} dy\,\exp\Big[-\frac{1}{2}y^*Q^*\Sigma^{-1}Qy + v^*Qy\Big]\\ &= c\int_{\mathbb{R}^m} dy\,\exp\Big[-\frac{1}{2}y^*\Lambda y + u^*y\Big],\end{aligned}$$

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

$$\begin{aligned}\prod_{i=1}^{m}\exp\Big[\frac{u_i^2}{2\lambda_i}\Big] &= \exp\Big[\frac{1}{2}(Q^*v)^*\Lambda^{-1}Q^*v\Big]\\ &= \exp\Big[\frac{1}{2}v^*Q\Lambda^{-1}Q^*v\Big] = \exp\Big[\frac{1}{2}v^*\Sigma v\Big]\end{aligned}$$

□

Essentially this result gives the moment-generating and characteristic functions of the multivariate normal distribution.

Corollary 7.3.4. *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 7.3.5 (Wick's theorem)**.** *Let $X = (x_i)$ be a multivariate normal m-vector with zero mean and variance Σ. Then,*

$$\mathbb{E}\big(x_{i_1} x_{i_2} \dots x_{i_k}\big) = \sum \Sigma_{ab} \dots \Sigma_{yz},$$

where the sum is over all different pairings $(ab), \dots, (yz)$ of the set of indices $\{i_1, i_2, \dots, i_k\}$.

An example should make this statement more clear. For two indices, we simply have $\mathbb{E}(x_i x_j) = \Sigma_{ij}$. For four indices, we have:

$$\mathbb{E}(x_i x_j x_k x_l) = \Sigma_{ij}\Sigma_{kl} + \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}.$$

Proof of Theorem 7.3.5. By a well-known result, we can write the moment as the multiple derivative of the moment generating function evaluated at zero:

$$\begin{aligned}\mathbb{E}\big(x_{i_1} x_{i_2} \dots x_{i_k}\big) &= \frac{\partial^k}{\partial t_{i_1} \dots \partial t_{i_k}} m_X(t)\Big|_{t=0} \\ &= \frac{\partial^k}{\partial t_{i_1} \dots \partial t_{i_k}} \exp\Big(\frac{1}{2} t^* \Sigma t\Big)\Big|_{t=0}\end{aligned}$$

Consider first the derivative with respect to t_{i_1}. By the chain rule it gives

$$\Big(\sum_{j=1}^{m} \Sigma_{i_1 j} t_j\Big) \exp\Big(\frac{1}{2} t^* \Sigma t\Big)$$

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 (7.5). 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 7.3.6. *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 7.1.2. □

A consequence of this theorem is that the marginal distributions of the multivariate normal vector are normal.

Theorem 7.3.7. *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. $\square$

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 7.3.8. *Assume the notation of theorem 7.3.7 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 7.3.9. 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, $\mathbb{V}\text{ar}(X_1|X_2, X_3) = 1/K_{11} = 1/3$ and so on.

7.4 An application

Theorem 7.4.1. *Let $X_1, X_2, \ldots, X_n$ be independent random variables distributed according to $N(\mu, \sigma^2)$. Then,*

$$\frac{(n-1)S^2}{\sigma^2} = \sum (X_i - \overline{X})^2$$

has a χ^2 distribution with $(n-1)$ df. Also, $\overline{X}$ and S^2 are independent random variables.

For the proof see Exercise 13.93 in Wackerly "Mathematical Statistics with Applications"

Chapter 8

QR factorization

8.1 Gram-Schmidt orthogonalization

In some cases we are given a basis $(a_1, a_2, \ldots)$ of a linear space V and we want to construct a orthogonal 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}. These is especially 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 had the unit length.

So,

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

where $r_{kk} = \|v_k\|$.

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 $i < j$ and $r_{ii} > 0$ is as defined above.

In a matrix form it can be written as

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

where A is an $m \times n$ matrix, Q is an $m \times n$ matrix with orthonormal columns and R is an upper-diagonal $n \times n$ matrix with positive diagonal elements.

This factorization is called the *reduced QR factorization* and 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 another, a faster method to calculate this factorization based on so-called Householder reflections. For details, see the textbook by Trefethen and Bau.

8.2 Relation to least squares problem

Since the columns of $\widehat{Q}$ give an orthonormal basis of the columns space of matrix A, it is very easy to project on this column space if $\widehat{Q}$ is computed. So QR factorization gives an algorithm for solving a linear squares problem.

Formally, we have a problem

$$Ax = y + \varepsilon,$$

where we assume that A is a full-rank matrix.

Its least-squares solution is the solution of the normal equations:

$$A^*Ax = A^*y,$$

which we can rewrite as

$$\begin{aligned} \widehat{R}^*\widehat{Q}^*\widehat{Q}\widehat{R}b &= \widehat{R}^*\widehat{Q}^*y \\ \widehat{R}^*\widehat{R}b &= \widehat{R}^*\widehat{Q}^*y \end{aligned}$$

Since the matrix $\widehat{R}^*$ is lower-diagonal and all its diagonal elements are positive, hence it is invertible, and we have:

$$\widehat{R}b = \widehat{Q}^*y$$

This gives the desired algorithm.

1. Compute the reduced QR factorization of A.
2. Calculate $z = \widehat{Q}^*y$
3. Solve the upper-triangular system $\widehat{R}x = z$ for x.

Apparently, this is one of the fastest methods to solve the least squares problem.

8.3 Relation to eigenvalue calculation

Theorem 8.3.1. *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 $A = A_0$) . If $A_{k-1} = Q_{k-1}R_{k-1}$, then we define

$$A_k = R_{k-1}Q_{k-1}.$$

By the previous theorem all A_k are real symmetric and have the same eigenvalues as A.

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 by the proof of the previous theorem, we have

$$A_k = (Q^{(k)})^*AQ^{(k)},$$

where $Q^{(k)} = Q_1Q_2 \dots Q_k$. Define also

$$R^{(k)} = R_kR_{k-1} \dots R_1$$

Theorem 8.3.2. *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 note that

$$\begin{aligned} AQ^{(k-1)} &= AQ_1Q_2 \dots Q_{k-1} = Q_1A_1Q_2 \dots Q_{k-1} \\ &= Q_1Q_2A_2 \dots Q_{k-1} \\ &\dots \\ &= Q_1Q_2Q_{k-1}A_k = Q^{(k-1)}Q_kR_k \end{aligned}$$

This implies that $A^k = Q^{(k)}R^{(k)}$. □

That is $Q_{(k)}$ is 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,

$$b_1\lambda_1^k u_1 + \ldots + b_n\lambda_1^k u_n.$$

if we assume $\lambda_1 > \lambda_2 > \ldots > \lambda_n$, then all columns of A^k, including the first one, are dominated 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. (Since 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 since $A = U\Lambda U^*$, we have $A_k = (Q^{(k)})^* A Q^{(k)} = (Q^{(k)})^* U \Lambda U^* Q^{(k)}$, and therefore A_k is close to Λ.

The detailed implementation of this plan is omitted.

8.4 Rayleigh quotient

Recall that we defined the norm of a matrix A as the maximum of the quotient

$$\frac{\|Ax\|}{\|x\|}$$

over all possible non-zero x. So the square of the norm of matrix A maximizes

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

For symmetric matrices, the eigenvalues and the norm are closely related. Indeed, in this case we have $A = U\Lambda U^*$, where U is unitary and Λ is diagonal, and since the unitary transformation does not change the norm, we know that $\|A\| = \|\Lambda\|$. For the diagonal matrix Λ we can directly solve the maximization problem and find that the norm of Λ (and hence A) is $|\lambda_1\|$, the largest absolute value of an eigenvalue of A.

In particular, this argument shows that for a *symmetric* matrix the square of the largest absolute value of an eigenvalue equals to the maximum of the ratio

$$\frac{(x, A^*Ax)}{(x, x)}.$$

The Rayleigh quotient is a modification of this idea, which focuses directly on eigenvalues, not their squares. By definition the Rayleigh quotient is the ratio:

$$R(x) = \frac{(x, Ax)}{(x, x)}.$$

Theorem 8.4.1 (Rayleigh-Ritz)**.** *If 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

$$\lambda_n(x, x) \leq (x, Ax) \leq \lambda_1(x, x) \tag{8.1}$$

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 (8.1) are achieved if we set x equal to the eigenvectors corresponding to eigenvalues λ_1 and λ_n. $\square$

This theorem 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.

An interesting extension to this 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 min max problem. Namely,

$$\lambda_k = \min_{V_{k-1}} \max_{x \neq 0, x \perp V_k} R(x),$$

where the minimization is over all $k-1$ dimensional subspaces V_{k-1}. The benefit is that one does not need to assume knowledge of the eigenvectors $u_1, \ldots, u_{k-1}$.

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 8.4.2. *Let A and B be two Hermitian $n \times n$ matrices. For each $k = 1, 2, \ldots, n$, we have*

$$\lambda_k(A) + \lambda_n(B) \leq \lambda_k(A + B) \leq \lambda_k(A) + \lambda_1(B)$$

In particular, if matrix B is non-negative definite, then all eigenvalues of A increase when we add B.

$$\lambda_k(A + B) \geq \lambda_k(A)$$

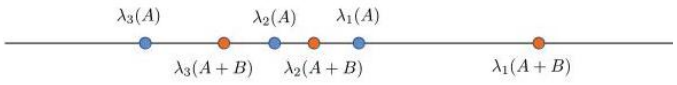

Figure 8.1

Another important and surprising result is the *interlacing theorem.* If the matrix B is nonnegative 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 8.1

It is 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.

Chapter 9

Exercises

Exercise 9.0.1 (Exercise 1.1 in Trefethen-Bau). 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 9.0.2 (Exercise 1.4 in Trefethen-Bau). 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 this lecture that $d1, \dots, d_8$ 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 $c_l, \dots, c_8$. What is the i, j entry of A^{-1}?

Exercise 9.0.3 (Exercise 2.3 in Trefethen-Bau). Let A be a hermitian matrix. An eigenvector of A is a non-zero vector x such that $Ax = \lambda x$ for some λ which can potentially be a complex number and which is the eigenvalue corresponding to the eigenvector x.

(a) Prove that all eigenvalues of A are real.

(b) Prove that if x and y are eigenvectors corresponding to distinct eigenvalues, then x and y are orthogonal.

Exercise 9.0.4 (Exercise 2.6 in Trefethen-Bau). If u and v are vectors, the matrix $A = I + uv^*$ is known as a *rank-one perturbation of the identity*. Show that if A is nonsingular, 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)?

Exercise 9.0.5 (Exercise 3.4 in Trefethen-Bau). Let A be an $m \times n$ matrix and let B be a submatrix of A, that is, a $\mu \times \nu$ matrix obtained by selecting certain rows and columns of A.

(a) Explain how B can be obtained by multiplying A by certain row and column "deletion" matrices" as in step 7 of Exercise 1.1.

(b) Using this product, show that $\|B\|_p \leq \|A\|_p$ for any p with $1 \leq p \leq \infty$.

Exercise 9.0.6 (Exercise 3.5 in Trefethen-Bau). Example 3.6 shows that if E is an outer product $E = uv^*$, then $\|E\|_2 = \|u\|_2\|v\|_2$. Is the same true for the Frobenius norm, that is, is $\|E\|_F = \|u\|_F\|v\|_F$? Prove it or give a counterexample.

Exercise 9.0.7 (Exercise 6.1 in Trefethen-Bau). If P is an orthogonal projector, then $I - 2P$ is unitary. Prove this algebraically, and give a geometric interpretation.

Exercise 9.0.8 (Exercise 6.2 in Trefethen-Bau). 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 9.0.9 (Exercise 6.3 in Trefethen-Bau). 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.

Exercise 9.0.10 (Exercise 6.4 in Trefethen-Bau). Consider the matrices

$$A = \begin{bmatrix} 1, & 0 \\ 0, & 1 \\ 1, & 0 \end{bmatrix} \text{ and } B = \begin{bmatrix} 1, & 2 \\ 0, & 1 \\ 1, & 0 \end{bmatrix}$$

Answer the following questions by hand calculation.

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

(b) Same question for B.

Exercise 9.0.11 (Exercises 4.1 and 5.1 in Trefethen-Bau). Determine the SVDs of the following matrices (by hand calculation):

$$\text{(a)} \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}, \text{(b)} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}, \text{(c)} \begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \text{(d)} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \text{(e)} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

What are the singular values of the matrix

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

Exercise 9.0.12 (Exercise 4.2 in Trefethen-Bau). 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.

Exercise 9.0.13 (Exercise 4.4 in Trefethen - Bau). Two matrices $m \times m$ matrices A and B are unitarily equivalent if $A = QBQ^*$ for some unitary Q. Is it true or false that A and B are unitarily equivalent if and only if they have the same singular values?

Exercise 9.0.14 (Exercise 5.4 in Trefethen - Bau). Suppose an $m \times m$ matrix A has an SVD $A = U\Sigma V^\star$. Find an eigenvalue decomposition $X\Lambda X^{-1}$ of the $2m \times 2m$ hermitian matrix

$$B = \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}.$$

Exercise 9.0.15 (Exercise 24.1 in Trefethen-Bau). 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 and "ew" stands for eigenvalue. (This comes from the German "Eigenwert." The corresponding abbreviation for eigenvector is "ev," from "Eigenvektor.")

a. If λ is an ew of A and $\mu \in \mathbb{C}$, then $\lambda - \mu$ is an ew of $A - \mu I$.

b. If A is real and λ is an ew of A, then so is $-\lambda$.

c. If A is real and λ is an ew of A, then so is $\overline{\lambda}$.

d. If λ is an ew of A and A is non-singular, then λ^{-1} is an ew of A^{-1}.

e. If all the ew's of A are zero, then $A = 0$.

f. If A is hermitian and λ is an ew of A then $|\lambda|$ is a singular value of A.

g. If A is diagonalizable and all its ew's are equal, then A is diagonal.

2) Exercise 5.2.6 in Strang:

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 is $(3, -1)$, what is the second row?

Exercise 9.0.16 (Exercise 5.3.4 in Strang). 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 as $n \to \infty$ of the matrices A^n.

c. If $G_0 = 0$ and $G_1 = 1$, show that the Gibonacci numbers approach $\frac{2}{3}$.

Exercise 9.0.17 (Exercise 24.4 (a) in Trefethen-Bau). The *spectral radius* $\rho(A)$ of a square matrix A is the largest absolute value $|\lambda|$ 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 (Theorem 24.9):

$$\lim_{n\to\infty} \|A^n\| = 0 \text{ if and only if } \rho(A) < 1.$$

Exercise 9.0.18 (Ex 1.1.4 from Norris' "Markov Chains"). A flea hops about at random on the vertices of a triangle with all jumps equally likely. 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 9.0.19 (From Norris' "Markov Chains"). Let X_n, $n = 0, 1, \ldots$, be a Markov chain on $\{1, 2, 3\}$ with transition matrix

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

Calculate the invariant distribution for this chain in each of the following cases: (a) $p = 1/16$, (b) $p = 1/6$, (c) $p = 1/2$.

Exercise 9.0.20 (Ex. 1.9.1 From Norris). 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 9.0.21. 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 $\mathbb{P}(Y > 0 | X = 1)$.

Exercise 9.0.22 (From Boyd's "Applied Linear Algebra"). A disease is introduced into a population. In each period (say, days) we count the fraction of the population that is in four different infection states:

1. Susceptible. These individuals can acquire the disease the next day.
2. Infected. These individuals have the disease.
3. Recovered (and immune). These individuals had the disease and survived, and now have immunity.
4. Deceased. These individuals had the disease, and unfortunately died from it.

There are many mathematical models that predict how the disease state fractions x_t evolve over time. One simple model can be expressed as a linear dynamical system. The model assumes the following happens over each day.

- 5% of the susceptible population will acquire the disease. (The other 95% will remain susceptible.)
- 1% of the infected population will die from the disease, 10% will recover and acquire immunity, and 4% will recover and not acquire immunity (and therefore, become susceptible). The remaining 85% will remain infected.

Write down the transition matrix for this model and find the steady state. (You can use software to calculate eigenvectors.)

Write a Python program to simulate the evolution of this system. Start with the initial state vector in which 0.1% of population is infected and all others are succeptible. By using Matplotlib, draw a picture, where each line illustrates the evolution of the percentage of each type of population.

How many days will it take for system to converge to equilibrium (with the convergence defined as the time when the norm of the difference of the state vector from the invariant distribution is smaller than 0.001.)