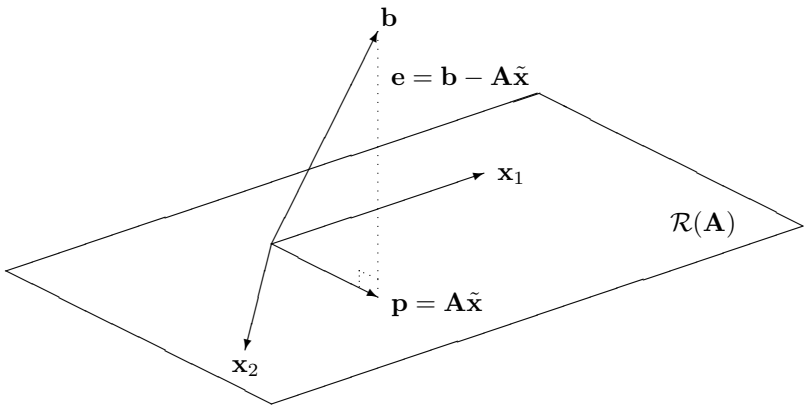
An Introduction to Linear Algebra



Andreas Jakobsson
Lund University

Version 080515

An Introduction to Linear Algebra

These notes were written as a part of a graduate level course on transform theory offered at King's College London during 2002 and 2003. The material is heavily indebt to the excellent textbook by Gilbert Strang [1], which the reader is referred to for a more complete description of the material; for a more in-depth coverage, the reader is referred to [2–6].

Andreas Jakobsson
andreas.jakobsson@ieee.org

CONTENTS

Chapter 1

LINEAR ALGEBRA

Dimidium facti, qui coepit, habet
Horace

1.1 Solving a set of linear equations

We introduce the topic of linear algebra by initially examining how to go about solving a set of linear equations; say

$$\left\{\begin{array}{rcrcrcr} 2u & + & v & + & w & = & 5 \\ 4u & - & 6v & & & = & -2 \\ -2u & + & 7v & + & 2w & = & 9 \end{array}\right. \tag{1.1.1}$$

By subtracting 2 times the first equation from the second equation, we obtain

$$\left\{\begin{array}{rcrcrcr} 2u & + & v & + & w & = & 5 \\ & - & 8v & - & 2w & = & -12 \\ -2u & + & 7v & + & 2w & = & 9 \end{array}\right. \tag{1.1.2}$$

Term this step (i). We then add the first equation to the third equation in (1.1.2),

$$\left\{\begin{array}{rcrcrcr} 2u & + & v & + & w & = & 5 \\ & - & 8v & - & 2w & = & -12 \\ & & 8v & + & 3w & = & 14 \end{array}\right. \tag{1.1.3}$$

We term this step (ii). Finally, in step (iii), we add the second and third equation in (1.1.3), obtaining

$$\left\{\begin{array}{rcrcrcr} 2u & + & v & + & w & = & 5 \\ & - & 8v & - & 2w & = & -12 \\ & & & & w & = & 2 \end{array}\right. \tag{1.1.4}$$

This procedure is known as Gaussian[1] elimination, and from (1.1.4) we can easily solve the equations via insertion, obtaining the solution $(u, v, w) = (1, 1, 2)$. An easy way to ensure that the found solution is correct is to simply insert (u, v, w) in (1.1.1). However, we will normally prefer to work with matrices instead of the above set of equations. The matrix approach offers many benefits, as well as opening up a new way to view the equations, and we will here focus on matrices and the geometries that can be associated with matrices.

1.2 Vectors and matrices

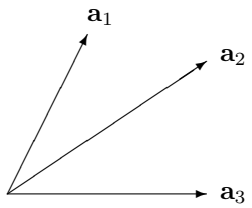
We proceed to write (1.1.1) on matrix form

$$\underbrace{\begin{bmatrix} 2 \\ 4 \\ -2 \end{bmatrix} u + \begin{bmatrix} 1 \\ -6 \\ 7 \end{bmatrix} v + \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} w}_{\begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}} = \begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix} \tag{1.2.1}$$

Note that we can view the columns of the matrix as vectors in a three-dimensional (real-valued) space, often written $\mathbb{R}^3$. This is only to say that the elements of the vectors can be viewed as the coordinates in a space, e.g., we view the third column vector, say $\mathbf{a}_3$,

$$\mathbf{a}_3 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \tag{1.2.2}$$

as the vector starting at origin and ending at $x = 1$, $y = 0$ and $z = 2$. Using a simplified geometric figure, we can depict the three column vectors as



where the vectors are seen as lying in a three-dimensional space. The solution to the linear set of equations in (1.1.1) is thus the scalings (u, v, w) such that if we

[1] Johann Carl Friedrich Gauss (1777-1855) was born in Brunswick, the Duchy of Brunswick (now Germany). In 1801, he received his doctorate in mathematics, proving the fundamental theorem of algebra [7]. He made remarkable contributions to all fields of mathematics, and is generally considered to be the greatest mathematician of all times.

combine the vectors as

$$u\mathbf{a}_1 + v\mathbf{a}_2 + w\mathbf{a}_3 = \mathbf{b}, \tag{1.2.3}$$

we obtain the vector $\mathbf{b}$, in this case $\mathbf{b} = \begin{bmatrix} 5 & -2 & 9 \end{bmatrix}^T$, where $(\cdot)^T$ denotes the transpose operator. This way of viewing the linear equations also tells us when a solution exists. We say that the matrix formed from the three column vectors,

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \end{bmatrix}, \tag{1.2.4}$$

span the *column space*, i.e., the vector space containing all points reachable by *any* linear combination of the column vectors of $\mathbf{A}$. For example, consider

$$\mathbf{B} = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{1.2.5}$$

In this case, all the points in the plane $\mathbf{b}_1$-$\mathbf{b}_2$ belongs to the column space of $\mathbf{B}$; this as by combining $\mathbf{b}_1$ and $\mathbf{b}_2$ all points in the plane might be reached. However, the point $\mathbf{b}_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$ does not lie in the column space since there is no way to combine $\mathbf{b}_1$ and $\mathbf{b}_2$ to reach $\mathbf{b}_3$. Thus, it holds that

> A linear set of equations is solvable *if and only if* the vector $\mathbf{b}$ lies in the column space of $\mathbf{A}$.

Viewing the linear equations as vectors will be central to our treatment, and we will shortly return to this way of viewing the problem. Before doing so, we will examine how we can view the above Gaussian elimination while working on the matrix form.

We can view the elimination in step (i) above as a multiplication by an elementary matrix, $\mathbf{E}$,

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{E}} \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \mathbf{E} \begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix} \tag{1.2.6}$$

which yields (cf. (1.1.2))

$$\begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ -2 & 7 & 2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 5 \\ -12 \\ 9 \end{bmatrix} \tag{1.2.7}$$

Similarly, we can view step (ii) and (iii) as multiplications with elementary matrices

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}}_{\mathbf{F}} \begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ -2 & 7 & 2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \mathbf{GF} \begin{bmatrix} 5 \\ -12 \\ 9 \end{bmatrix} \tag{1.2.8}$$

yielding (cf. 1.1.4))

$$\begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 5 \\ -12 \\ 2 \end{bmatrix} \tag{1.2.9}$$

The resulting matrix, which we will term $\mathbf{U}$, will be an *upper-triangular* matrix. We denote the sought solution $\mathbf{x} = \begin{bmatrix} u & v & w \end{bmatrix}^T$, and note that we have by the matrix multiplications rewritten

$$\mathbf{Ax} = \mathbf{b} \quad \Rightarrow \quad \mathbf{Ux} = \mathbf{c}, \tag{1.2.10}$$

where $\mathbf{U} = (\mathbf{GFE})\,\mathbf{A}$ and $\mathbf{c} = (\mathbf{GFE})\,\mathbf{b}$. As $\mathbf{U}$ is upper-triangular, the new matrix equation is obviously trivial to solve via insertion.

We proceed by applying the inverses of the above elementary matrices (in reverse order!), and note that

$$\underbrace{\mathbf{E}^{-1}\mathbf{F}^{-1}\mathbf{G}^{-1}}_{\mathbf{L}}\underbrace{\mathbf{GFEA}}_{\mathbf{U}} = \mathbf{A}. \tag{1.2.11}$$

This is the so-called LU-decomposition, stating that

> A square matrix that can be reduced to $\mathbf{U}$ (without row interchanges[2]) can be written as $\mathbf{A} = \mathbf{LU}$.

As the inverse of an elementary matrix can be obtain by simply changing the sign of the off-diagonal element[3],

$$\mathbf{E}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1.2.12}$$

we can immediately obtain the LU-decomposition after reducing $\mathbf{A}$ to $\mathbf{U}$,

$$\mathbf{A} = \mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ 0 & 0 & 1 \end{bmatrix} \tag{1.2.13}$$

Note that $\mathbf{L}$ will be a *lower-triangular* matrix. Often, one divides out the diagonal elements of $\mathbf{U}$, the so-called *pivots*[4], denoting the diagonal matrix with the pivots on the diagonal $\mathbf{D}$, i.e.,

$$\mathbf{A} = \mathbf{LU} = \mathbf{LD}\tilde{\mathbf{U}}. \tag{1.2.14}$$

[2] Note that in some cases, row exchanges are required to reduce $\mathbf{A}$ to $\mathbf{U}$. In such case, $\mathbf{A}$ can not be written as $\mathbf{LU}$. See [1] for further details on row interchanges.

[3] Please note that this *only holds* for elementary matrices!

[4] We note that pivots are *always* nonzero.

A bit confusing, $\tilde{\mathbf{U}}$ is normally just denoted $\mathbf{U}$, forcing the reader to determine which of the matrices are referred to from the context. For our example,

$$\mathbf{A} = \mathbf{L}\mathbf{D}\tilde{\mathbf{U}} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & -8 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0 & 1 & 0.25 \\ 0 & 0 & 1 \end{bmatrix} \tag{1.2.15}$$

The LU-decomposition suggests an efficient approach to solve a linear set of equations by first factoring the matrix as $\mathbf{A} = \mathbf{L}\mathbf{U}$, and then solve $\mathbf{L}\mathbf{c} = \mathbf{b}$ and $\mathbf{U}\mathbf{x} = \mathbf{c}$. It can be shown that the LU-decomposition of a $n \times n$ matrix requires $n^3/3$ operations, whereas the solving of the two triangular systems requires $n^2/2$ operations each. Thus, if we wish to solve the same set of equations (same $\mathbf{A}$ matrix) with a new $\mathbf{b}$ vector, this can be done in only n^2 operations (which is significantly less than $n^3/3 + n^2$). Often, the inherent structure of the $\mathbf{A}$ matrix can further simplify calculations. For example, we will often consider *symmetric* matrices, i.e., matrices satisfying

$$\mathbf{A} = \mathbf{A}^T, \tag{1.2.16}$$

or, more generally, *Hermitian*[5] matrices, i.e., matrices satisfying

$$\mathbf{A} = \mathbf{A}^H, \tag{1.2.17}$$

where $(\cdot)^H$ denotes the Hermitian, or conjugate transpose, operator. Obviously, a Hermitian matrix is also symmetric, whereas the converse is in general not true. If $\mathbf{A}$ is Hermitian, and if it can be factored as $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{U}$ (without row exchanges to destroy the symmetry), then $\mathbf{U} = \mathbf{L}^H$, i.e.,

> A square Hermitian matrix that (without row interchanges) can be reduced to $\mathbf{U}$, can be written as $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^H$, where $\mathbf{D}$ is diagonal, and $\mathbf{L}$ lower-triangular.

This symmetry will simplify the computation of the LU-decomposition, reducing the complexity to $n^3/6$ operations. Further, in the case when the diagonal elements of $\mathbf{D}$ are all positive, the LU-decomposition can be written as

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^H = \left(\mathbf{L}\mathbf{D}^{1/2}\right)\left(\mathbf{L}\mathbf{D}^{1/2}\right)^H = \mathbf{C}\mathbf{C}^H, \tag{1.2.18}$$

where $\mathbf{C}$ is the (matrix) square root of $\mathbf{A}$. However, if $\mathbf{C}$ is a square root of $\mathbf{A}$, then so is $\mathbf{C}\mathbf{B}$, i.e.,

$$\mathbf{A} = (\mathbf{C}\mathbf{B})(\mathbf{C}\mathbf{B})^H, \tag{1.2.19}$$

[5] Charles Hermite (1822-1901) was born in Dieuze, France, with a defect in his right foot, making it hard for him to move around. After studying one year at the Ecole Polytechnique, Hermite was refused the right to continue his studies because of this disability. However, this did not prevent his work, and in 1856 he was elected to the Académie des Sciences. After making important contributions to number theory and algebra, orthogonal polynomials, and elliptic functions, he was in 1869 appointed professor of analysis both at the Ecole Polytechnique and at the Sorbonne.

for any matrix $\mathbf{B}$ such that

$$\mathbf{B}\mathbf{B}^H = \mathbf{B}^H\mathbf{B} = \mathbf{I}. \tag{1.2.20}$$

A matrix satisfying (1.2.20) is said to be a *unitary matrix* [6]. Hence, there are an infinite number of square roots of a given matrix satisfying (1.2.18). Two often used choices for square roots are

(i) The Hermitian square root: $\mathbf{C} = \mathbf{C}^H$. The Hermitian square root is unique.

(ii) The Cholesky[7] factor. If $\mathbf{C}$ is lower-triangular with non-negative diagonal elements, then $\mathbf{C}$ is called the Cholesky factor of $\mathbf{A}$. This is the famous Cholesky factorization of $\mathbf{A}$.

We will now return to the discussion on vectors and vector spaces.

1.3 Fundamental subspaces

Recall that $\mathbf{A}\mathbf{x} = \mathbf{b}$ is solvable *if and only if* the vector $\mathbf{b}$ lies in the column space of $\mathbf{A}$, i.e., if it can be expressed as a combination of the column vectors of $\mathbf{A}$. We will now further discuss vector spaces[8], as well as different ways to represent such a space. For instance, how can we represent this vector space using as few vectors as possible? Can the matrix $\mathbf{A}$ be used to span any other vector spaces? To answer these questions, we will discuss what it means to span a space, and some matters that concern vector spaces. We begin with the central definition of linear dependence. If

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = 0 \quad \text{only happens when} \quad c_1 = c_2 = \cdots = c_k = 0 \tag{1.3.1}$$

then the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are *linearly independent*. Otherwise, at least one of the vectors is a linear combination of the others. The number of linearly independent columns of a matrix is termed the *column rank* of the matrix. If the matrix consists of only linearly independent column vectors, one say that it has *full column rank*. For example, consider

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} \tag{1.3.2}$$

[6] In the case of real-valued data, (1.2.20) can be written as $\mathbf{B}\mathbf{B}^T = \mathbf{B}^T\mathbf{B} = \mathbf{I}$. Such a matrix is said to be an *orthogonal matrix*. Here, the term orthonormal matrix would be more appropriate, but this term is not commonly used.

[7] Andre-Louis Cholesky (1875-1918) was born in Montguyon, France. He was a French military officer involved in geodesy and surveying in Crete and North Africa just before World War I. He developed the method now named after him to compute solutions to the normal equations for some least squares data fitting problems arising in geodesy. He died in battle during the war, his work published posthumously.

[8] See appendix A for a further discussion on vector spaces.

The matrix $\mathbf{A}$ clearly does not have full column rank, as the column vectors are linearly dependent: note that

$$\begin{aligned} \mathbf{a}_2 &= 3\mathbf{a}_1 \\ \mathbf{a}_4 &= \mathbf{a}_1 + \frac{1}{3}\mathbf{a}_3 \end{aligned}$$

Only the first and the third column vectors are linearly independent, and the column rank of $\mathbf{A}$ is thus two. We are now ready to clarify what we mean by the spanning of a space: if a vector space $\mathbf{X}$ consists of all linear combinations of the particular vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, then these vectors *span* the space. In other words, this means that *every* vector $\mathbf{v} \in \mathbf{X}$ can be expressed as some combination of these vectors:

$$\mathbf{v} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \ldots + c_n\mathbf{x}_n, \tag{1.3.3}$$

for some coefficients $\{c_k\}$. Note that a set of vectors that span a space does not need to be linearly independent. We can for instance use the column vectors of $\mathbf{A}$ in (1.3.2) to span the column space of $\mathbf{A}$, normally denoted $\mathcal{R}(\mathbf{A})$. Obviously, there is an infinite number of vectors that we could use to span $\mathcal{R}(\mathbf{A})$, but how many are really needed? This question leads to the idea of a *basis*.

A basis for a vector space is a set of vectors satisfying:

(i) It is linearly independent.

(ii) It spans the space.

The number of vectors spanning a basis is called the *dimension* of the space. For the matrix in (1.3.2), the first and the third column vectors are sufficient to span the space, and these two constitute a basis for $\mathcal{R}(\mathbf{A})$. The dimension of $\mathcal{R}(\mathbf{A})$, which is (always) the same as the column rank of the matrix $\mathbf{A}$, is thus two. As a result, the column space will be a *subspace* of $\mathbb{R}^3$, this as the column vectors lie in a three-dimensional (real-valued) space. Note that $\mathcal{R}(\mathbf{A})$, being two-dimensional, will not fill the entire $\mathbb{R}^3$, immediately raising the question what fills the remaining part of this space. We will return to this point shortly.

There is clearly an infinite number of bases spanning any given space, for instance, any scaled version of $\mathbf{a}_1$ and $\mathbf{a}_3$ would do. We will eventually examine how to proceed to pick an (in some sense) appropriate basis, but will first spend a moment on the row vectors. Considering (1.3.2), we can obviously view it as constructed from a set of *row vectors*,

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \end{bmatrix} \tag{1.3.4}$$

where

$$\mathbf{b}_1 = \begin{bmatrix} 1 & 3 & 3 & 2 \end{bmatrix} \tag{1.3.5}$$
$$\mathbf{b}_2 = \begin{bmatrix} 2 & 6 & 9 & 5 \end{bmatrix} \tag{1.3.6}$$
$$\mathbf{b}_3 = \begin{bmatrix} -1 & -3 & 3 & 0 \end{bmatrix} \tag{1.3.7}$$

Note that the row vectors lie not in $\mathbb{R}^3$, but in $\mathbb{R}^4$. This as they contain four, not three, (real-valued) indices. These vectors must for this reason reside in a *different* vector space than the column vectors; we will get back to this point shortly. Similarly to our definition of column rank, we define the row rank as the number of linearly independent row vectors, and speak of *full row rank* if all the row vectors are linearly independent. What is the row rank of the matrix in (1.3.4)? We can easily answer this using the following remarkable fact:

> The *number* of independent column vectors is always equal to the *number* of independent row vectors!

This means that one (and only one) of the row vectors in (1.3.4) must be linearly dependent. Can you see this? This also means that the row rank and column ranks are always equal. For this reason, we normally only refer to the row and the column rank as the *rank* of the matrix[9],

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T). \tag{1.3.8}$$

Consequently, we say that the matrix is *full rank* when all the column vectors *and* the row vectors are linearly independent. For obvious reasons, only square matrices can be full rank. Let us study some examples:

$$\begin{bmatrix} 1 & 1 & -1 \\ -1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & -1 \end{bmatrix} \quad \begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}$$

What are the dimensions of the column spaces spanned by these matrices? What are the ranks of the matrices? Which of the matrices are bases?

We are now ready to examine the question if $\mathbf{A}$ will span other spaces than $\mathcal{R}(\mathbf{A})$. From the previous discussion, it might seem as spanning one space is plenty: in fact $\mathbf{A}$ will span *four* different spaces, the so-called *fundamental subspaces*. Consider a matrix $\mathbf{A}$ containing m row vectors and n column vectors, normally referred to as a $m \times n$ matrix, with rank r. Which are then these fundamental subspaces? The two first are easily grasped:

(i) The column space, $\mathcal{R}(\mathbf{A})$, spanned by the column vectors of $\mathbf{A}$.

[9] We note that the rank of a matrix is equal to the number of pivots.

(ii) The row space, $\mathcal{R}(\mathbf{A}^T)$, spanned by the row vectors of $\mathbf{A}$.

Both these spaces will have dimension r, as the number of linearly independent column vectors equal the number of linearly independent row vectors. The two remaining subspaces consist of the so-called *nullspaces*. Consider a vector $\mathbf{x}$ such that

$$\mathbf{A}\mathbf{x} = \mathbf{0}. \tag{1.3.9}$$

Obviously the null vector, $\mathbf{x} = \mathbf{0}$, will satisfy this, but other vectors may do so as well. Any vector, other than the null vector, satisfying (1.3.9) is said to lie in the *nullspace* of $\mathbf{A}$. Similarly, we denote the space spanned by the vectors $\mathbf{y}$, such that

$$\mathbf{y}^T\mathbf{A} = \mathbf{0}, \tag{1.3.10}$$

the *left nullspace* of $\mathbf{A}$, this to denote that the vector $\mathbf{y}$ multiply $\mathbf{A}$ *from the left*[10]. The two final fundamental subspaces are thus:

(iii) The nullspace, $\mathcal{N}(\mathbf{A})$, is spanned by the vectors $\mathbf{x}$ satisfying (1.3.9).

(iv) The left nullspace, $\mathcal{N}(\mathbf{A}^T)$, is spanned by the vectors $\mathbf{y}$ satisfying (1.3.10).

For example, for the matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{1.3.11}$$

the columns space, $\mathcal{R}(\mathbf{B})$, and row space, $\mathcal{R}(\mathbf{B}^T)$, are the lines (both having dimension $r = 1$) through

$$\begin{bmatrix} 1 & 0 \end{bmatrix}^T \quad \text{respectively} \quad \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$$

Further, the nullspace, $\mathcal{N}(\mathbf{B})$, is a plane containing

$$\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T \quad \text{and} \quad \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$$

and the left nullspace, $\mathcal{N}(\mathbf{B}^T)$, is the line through $\begin{bmatrix} 0 & 1 \end{bmatrix}^T$. We note that the nullspaces have different dimensions; how are these dimensions related to the matrix $\mathbf{B}$? To answer this, we will need examine the notion of orthogonality.

1.4 Orthogonality

We begin by introducing the *inner product* between two vectors $\mathbf{x}$ and $\mathbf{y}$ as

$$\langle \mathbf{x}, \mathbf{y} \rangle \stackrel{\triangle}{=} \mathbf{x}^H\mathbf{y}. \tag{1.4.1}$$

The inner product is a *measure of similarity* between the vectors, and is proportional to the degree of closeness between $\mathbf{x}$ and $\mathbf{y}$; the more similar the vectors are,

[10] Thus, the ordinary nullspace can be thought of as the "right" nullspace.

the bigger the inner product. If the vectors are parallel (either direct or reverse), the magnitude of the inner product reaches its maximum. If the vectors are perpendicular, the inner product is zero (more on this in the next section). For example, the inner product between $\mathbf{a}_1$ and $\mathbf{a}_2$ in (1.3.2) is

$$\langle \mathbf{a}_1, \mathbf{a}_2 \rangle = \mathbf{a}_1^T \mathbf{a}_2 = 3\, \mathbf{a}_1^T \mathbf{a}_1 = 18.$$

From geometry, it is well known that one can compute the angle between two vectors using the inner product between the vectors and their lengths as

$$\cos\theta = \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|}{||\mathbf{x}||_2 \, ||\mathbf{y}||_2}. \tag{1.4.2}$$

where the length of a vector is (in Euclidean space) the 2-norm[11] of the vector, i.e., for a vector in $\mathbf{l}_2(\mathbb{C})$,

$$||\mathbf{x}||_2 \stackrel{\triangle}{=} \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\mathbf{x}^H \mathbf{x}}. \tag{1.4.3}$$

From (1.4.3), we conclude that no vector (excluding the null vector) can have zero length. Considering (1.4.2), we see that two vectors are *orthogonal* if and only if their inner product $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. For example, the vectors

$$\begin{aligned} \mathbf{x} &= \begin{bmatrix} 2 & 2 & -1 \end{bmatrix}^T \\ \mathbf{y} &= \begin{bmatrix} -1 & 2 & 2 \end{bmatrix}^T \end{aligned}$$

are orthogonal, as their inner product is zero. Also note that the vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly independent! In general, it holds that:

Orthogonal vectors are linearly independent.

Further, we say that two subspaces are orthogonal to each other if *every* vector in the first subspace $\mathbf{V}$ is orthogonal to *every* vector in the second subspace $\mathbf{W}$. The subspace $\mathbf{W}$ is then said to lie in the orthogonal complement of $\mathbf{V}$, denoted $\mathbf{V}^\perp$ (pronounced "V perp"). Similarly, $\mathbf{V}$ will lie in $\mathbf{W}^\perp$. Recalling the definitions of the fundamental subspaces, we see that

(i) The nullspace is the orthogonal complement of the row space.

(ii) The left nullspace is the orthogonal complement of the column space.

This explains the dimensions of the subspaces spanned by $\mathbf{B}$ in (1.3.11): every vector that does not lie in the row space will belong to the nullspace, every vector not in the column space will belong to the left nullspace! Thus, the dimensions of the four fundamental subspaces for a $m \times n$ matrix $\mathbf{A}$ are

[11] See appendix A for a further discussion on norms and spaces. To make the discussion more general, we hereafter consider only complex-valued vectors unless otherwise specified, with real-valued vectors being a special case. Note that according to the Cauchy-Schwartz inequality, as stated in (C.1.2), (1.4.2) is limited to $0 \leq \theta \leq \frac{\pi}{2}$.

(i) $\dim\{\mathcal{R}(\mathbf{A})\} = \text{rank}\{\mathbf{A}\} = r$

(ii) $\dim\{\mathcal{R}(\mathbf{A}^T)\} = r$

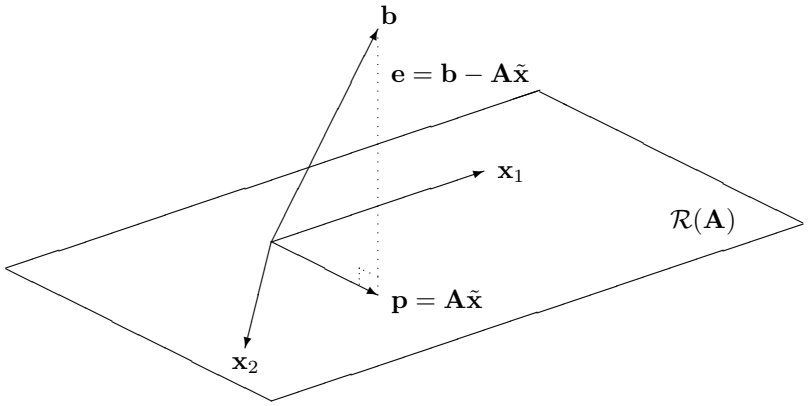
(iii) $\dim\{\mathcal{N}(\mathbf{A})\} = n - r.$

(iv) $\dim\{\mathcal{N}(\mathbf{A}^T)\} = m - r.$

The notion of orthogonality can be of further use; recall that $\mathbf{Ax} = \mathbf{b}$ is only solvable if $\mathbf{b} \in \mathcal{R}(\mathbf{A})$. What happens if $\mathbf{b} \notin \mathcal{R}(\mathbf{A})$? The obvious answer is that there exists no vector $\mathbf{x}$ such that $\mathbf{Ax} = \mathbf{b}$. However, we can still find a vector $\tilde{\mathbf{x}}$ such that $\mathbf{A}\tilde{\mathbf{x}}$ will (in some sense) lie as close as possible to $\mathbf{b}$. Often, one choose the vector $\tilde{\mathbf{x}}$ such that the Euclidean distance $\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|_2^2$ is minimized, in which case $\tilde{\mathbf{x}}$ is referred to as the *least squares solution*[12], i.e., the vector minimizing the squared error. Put differently,

$$\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2 \tag{1.4.4}$$

Geometrically, this is the same as finding the vector $\mathbf{p} = \mathbf{A}\tilde{\mathbf{x}}$, lying in $\mathcal{R}(\mathbf{A})$, such that the error, $\mathbf{e}$, between $\mathbf{b}$ and $\mathbf{A}\tilde{\mathbf{x}}$, is as short as possible. Thus, if $\mathbf{b}$ lies in $\mathcal{R}(\mathbf{A})$, the error is zero and the exact solution $\tilde{\mathbf{x}} = \mathbf{x}$ will exist. Otherwise, the error will be minimized if chosen such that it is *perpendicular* to the column space[13]. This is a very important point; spend a moment examining the figure below and make sure that you understand it!



As the error, $\mathbf{e} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$, will be orthogonal to $\mathcal{R}(\mathbf{A})$, the vector $\mathbf{p} = \mathbf{A}\tilde{\mathbf{x}}$ can be viewed as the *projection* of $\mathbf{b}$ onto $\mathcal{R}(\mathbf{A})$. To see this, note that as $\mathbf{e} \perp \mathcal{R}(\mathbf{A})$,

$$\mathbf{A}^H\mathbf{e} = \mathbf{A}^H(\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}) = \mathbf{0} \quad \Rightarrow \quad \mathbf{A}^H\mathbf{A}\tilde{\mathbf{x}} = \mathbf{A}^H\mathbf{b}. \tag{1.4.5}$$

[12]The least squares solution was introduced by C. F. Gauss, who used it to determine the orbit of the asteroid Ceres in 1821 by formulating the estimation problem as an optimization problem.

[13]This fact is often exploited in estimation theory, where it is termed the *orthogonality principle*.

Using the fact that

> If the $\mathbf{A}$ has linearly independent columns, then $\mathbf{A}^H\mathbf{A}$ is *square, Hermitian* and *invertible*.

we obtain

$$\tilde{\mathbf{x}} = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{b} \triangleq \mathbf{A}^\dagger\mathbf{b}, \tag{1.4.6}$$

where $\mathbf{A}^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{A}$. The projection of $\mathbf{b}$ onto $\mathcal{R}(\mathbf{A})$ is thus

$$\mathbf{p} = \mathbf{A}\tilde{\mathbf{x}} = \mathbf{A}\left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{b} = \mathbf{A}\mathbf{A}^\dagger\mathbf{b} \triangleq \mathbf{\Pi}_\mathbf{A}\mathbf{b}, \tag{1.4.7}$$

where $\mathbf{\Pi}_\mathbf{A}$ is the projection matrix onto $\mathcal{R}(\mathbf{A})$. What would happen if you would project $\mathbf{p}$ onto $\mathcal{R}(\mathbf{A})$? In other words, what is $\mathbf{\Pi}_\mathbf{A}^2\mathbf{b}$? How can you explain this geometrically? Another point worth noting is that

$$\mathbf{e} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} - \mathbf{\Pi}_\mathbf{A}\mathbf{b} = \left(\mathbf{I} - \mathbf{\Pi}_\mathbf{A}\right)\mathbf{b} \triangleq \mathbf{\Pi}_\mathbf{A}^\perp\mathbf{b}, \tag{1.4.8}$$

where $\mathbf{\Pi}_\mathbf{A}^\perp$ is the projection matrix onto the space orthogonal to $\mathcal{R}(\mathbf{A})$, i.e., $\mathbf{\Pi}_\mathbf{A}^\perp$ is the projection matrix onto the left nullspace. Minimizing the error vector is thus the same thing as minimizing the projection of $\mathbf{b}$ onto the *left nullspace*, i.e., we write $\mathbf{b} = \mathbf{\Pi}_\mathbf{A}\mathbf{b} + \mathbf{\Pi}_\mathbf{A}^\perp\mathbf{b}$ and find the vector $\tilde{\mathbf{x}}$ such that

$$\tilde{\mathbf{x}} = \arg\min_\mathbf{x} \|\mathbf{\Pi}_\mathbf{A}^\perp\mathbf{b}\|_2^2, \tag{1.4.9}$$

which is equivalent to the expression in (1.4.4). As an example, lets examine an overdetermined ($m > n$) system

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}}_{\mathbf{A}}\mathbf{x} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{\mathbf{b}} \tag{1.4.10}$$

As $\mathbf{b}$ does not lie in $\mathcal{R}(\mathbf{A})$ (show this!), no solution will exist. The closest solution, in the least squares sense, can be found as

$$\tilde{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{b} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \tag{1.4.11}$$

which corresponds to

$$\mathbf{p} = \mathbf{A}\tilde{\mathbf{x}} = \mathbf{\Pi}_\mathbf{A}\mathbf{b} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T \tag{1.4.12}$$

Verify that the error vector is orthogonal to the column space by computing the angle between $\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ and the columns of $\mathbf{A}$!

We now proceed to discuss how to find an orthogonal basis for a given space, say $\mathcal{R}(\mathbf{A})$; as commented earlier, a basis is formed by a set of linearly independent vector spanning the space. By necessity, this set must contain $r = \text{rank}\,\{\mathbf{A}\}$ vectors (why?), but obviously there exist an infinite number of such vector sets. We are here interested in finding an *orthogonal* set of vectors using $\mathbf{A}$. This can be done in a series of projection termed the Gram-Schmidt orthogonalization which will now be described briefly; consider the matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \dots & \mathbf{a}_n \end{bmatrix} \tag{1.4.13}$$

and choose the first basis vector, $\mathbf{q}_1$, as

$$\mathbf{q}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2} \tag{1.4.14}$$

Note that due to the normalization, $\mathbf{q}_1$ has unit length, we say it is a *unit vector*. The second basis vector should be *orthogonal* to $\mathbf{q}_1$, and for this reason we form

$$\mathbf{q}_2' = \mathbf{a}_2 - \langle \mathbf{q}_1, \mathbf{a}_2 \rangle \mathbf{q}_1 \tag{1.4.15}$$

and choose the second basis vector as

$$\mathbf{q}_2 = \frac{\mathbf{q}_2'}{\|\mathbf{q}_2'\|_2} \tag{1.4.16}$$

The second basis vector is thus formed from $\mathbf{a}_2$, but with the component in the direction of $\mathbf{q}_1$ subtracted out. Note that (1.4.15) can be written as

$$\mathbf{q}_2' \;=\; \mathbf{a}_2 - \mathbf{q}_1^H \mathbf{a}_2 \mathbf{q}_1 = \mathbf{a}_2 - \underbrace{\mathbf{q}_1 \left(\mathbf{q}_1^H \mathbf{q}_1\right)^{-1} \mathbf{q}_1^H}_{\mathbf{\Pi}_{\mathbf{q}_1}} \mathbf{a}_2 = \mathbf{\Pi}_{\mathbf{q}_1}^{\perp} \mathbf{a}_2 \tag{1.4.17}$$

where we used the fact that $\|\mathbf{q}_1\| = 1$. Thus, the second vector is nothing but the projection of $\mathbf{a}_2$ onto the space orthogonal to $\mathcal{R}(\mathbf{q}_1)$; as a result it must be orthogonal to $\mathbf{q}_1$. Due to the normalization in (1.4.14) and (1.4.16), the two vectors are *orthonormal*, i.e., they are orthogonal unit vectors. We proceed to construct

$$\mathbf{Q} = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{bmatrix} \tag{1.4.18}$$

and form $\mathbf{q}_3$ to be orthonormal to the vectors in $\mathbf{Q}$, i.e.,

$$\mathbf{q}_3 = \frac{\mathbf{\Pi}_{\mathbf{Q}}^{\perp} \mathbf{a}_3}{\left\| \mathbf{\Pi}_{\mathbf{Q}}^{\perp} \mathbf{a}_3 \right\|_2} \tag{1.4.19}$$

By adding $\mathbf{q}_3$ to $\mathbf{Q}$ and repeating the projections, we can in this manner construct the orthonormal basis. Written in matrix form, this yields

$$\mathbf{A} = \mathbf{Q}\,\mathbf{R} = \begin{bmatrix} \mathbf{q}_1 & \dots & \mathbf{q}_n \end{bmatrix} \begin{bmatrix} \langle \mathbf{q}_1, \mathbf{a}_1 \rangle & \langle \mathbf{q}_1, \mathbf{a}_2 \rangle & \dots & \langle \mathbf{q}_1, \mathbf{a}_n \rangle \\ & \langle \mathbf{q}_2, \mathbf{a}_2 \rangle & \dots & \langle \mathbf{q}_2, \mathbf{a}_n \rangle \\ & & \ddots & \vdots \\ \mathbf{0} & & & \langle \mathbf{q}_n, \mathbf{a}_n \rangle \end{bmatrix} \tag{1.4.20}$$

where the $m \times n$ matrix $\mathbf{Q}$ forms the orthonormal basis, and $\mathbf{R}$ is upper triangular and invertible. This is the famous Q-R factorization. Note that $\mathbf{A}$ does not need to be a square matrix to be written as in (1.4.20); however, if $\mathbf{A}$ is a square matrix, then $\mathbf{Q}$ is a unitary matrix. In general:

> If $\mathbf{A} \in \mathbb{C}^{m\times n}$, with $m \geq n$, there is a matrix $\mathbf{Q} \in \mathbb{C}^{m\times n}$ with orthonormal columns and an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{n\times n}$ such that $\mathbf{A} = \mathbf{QR}$. If $m = n$, $\mathbf{Q}$ is unitary; if in addition $\mathbf{A}$ is nonsingular, then $\mathbf{R}$ may be chosen so that all its diagonal entries are positive. In this case, the factors $\mathbf{Q}$ and $\mathbf{R}$ are unique.

The Q-R factorization concludes our discussion on orthogonality, and we proceed to dwell on the notion of determinants.

1.5 Determinants

As we have seen above, we often wish to know when a *square* matrix is invertible[14]. The simple answer is:

> A matrix $\mathbf{A}$ is invertible, when the determinant of the matrix is non-zero, i.e., $\det\{\mathbf{A}\} \neq 0$.

That is a good and short answer, but one that immediately raises another question. What is then the determinant? This is less clear cut to answer, being somewhat elusive. You might say that the determinant is a number that capture the soul of a matrix, but for most this is a bit vague. Let us instead look at the definition of the determinant; for a 2 by 2 matrix it is simple:

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \tag{1.5.1}$$

For a larger matrix, it is more of a mess. Then,

$$\det \mathbf{A} = a_{i1}\mathbf{A}_{i1} + \cdots + a_{in}\mathbf{A}_{in}, \tag{1.5.2}$$

where the cofactor $\mathbf{A}_{ij}$ is

$$\mathbf{A}_{ij} = (-1)^{i+j} \det \mathbf{M}_{ij} \tag{1.5.3}$$

and the matrix $\mathbf{M}_{ij}$ is formed by deleting row i and column j of $\mathbf{A}$ (see [1] for further examples of the use of (1.5.2) and (1.5.3)). Often, one can use the properties of the determinant to significantly simplify the calculations. The main properties are

(i) The determinant depends linearly on the first row.

$$\begin{vmatrix} a + a' & b + b' \\ c & d \end{vmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} + \begin{vmatrix} a' & b' \\ c & d \end{vmatrix} \tag{1.5.4}$$

[14] We note in passing that the inverse does not exist for non-square matrices; however, such matrices might still have a left or a right inverse.

(ii) The determinant change sign when two rows are exchanged.

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = - \begin{vmatrix} c & d \\ a & b \end{vmatrix} \tag{1.5.5}$$

(iii) $\det \mathbf{I} = 1$.

Given these three basic properties, one can derive numerous determinant rules. We will here only mention a few convenient rules, leaving the proofs to the interested reader (see, e.g., [1, 6]).

(iv) If two rows are equal, then $\det \mathbf{A} = 0$.

(v) If $\mathbf{A}$ has a zero row, then $\det \mathbf{A} = 0$.

(vi) Subtracting a multiple of one row from another leaves the $\det \mathbf{A}$ unchanged.

(vii) If $\mathbf{A}$ is triangular, then $\det \mathbf{A}$ is equal to the product of the main diagonal.

(viii) If $\mathbf{A}$ is singular, $\det \mathbf{A} = 0$.

(ix) $\det \mathbf{AB} = (\det \mathbf{A})(\det \mathbf{B})$

(x) $\det \mathbf{A} = \det \mathbf{A}^T$.

(xi) $|c\mathbf{A}| = c^n |\mathbf{A}|$, for $c \in \mathbb{C}$.

(xii) $\det \left(\mathbf{A}^{-1}\right) = (\det \mathbf{A})^{-1}$.

(xiii) $|\mathbf{I} - \mathbf{AB}| = |\mathbf{I} - \mathbf{BA}|$.

where $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n \times n}$. Noteworthy is that the determinant of a singular matrix is zero. Thus, as long as the determinant of $\mathbf{A}$ is non-zero, the matrix $\mathbf{A}$ will be non-singular and invertible. We will make immediate use of this fact.

1.6 Eigendecomposition

We now move to study eigenvalues and eigenvectors. Consider a square matrix $\mathbf{A}$. By definition, $\mathbf{x}$ is an *eigenvector* of $\mathbf{A}$ if it satisfies

$$\mathbf{Ax} = \lambda \mathbf{x}. \tag{1.6.1}$$

That is, if multiplied with $\mathbf{A}$, it will only yield a scaled version of itself. Note that (1.6.1) implies that

$$(\mathbf{A} - \lambda \mathbf{I})\, \mathbf{x} = 0, \tag{1.6.2}$$

requiring $\mathbf{x}$ to be in $\mathcal{N}(\mathbf{A} - \lambda \mathbf{I})$. We require that the *eigenvector* $\mathbf{x}$ must be non-zero, and as a result, $\mathbf{A} - \lambda \mathbf{I}$ must be singular[15], i.e., the scaling λ is an *eigenvalue* of $\mathbf{A}$ if and only if

$$\det (\mathbf{A} - \lambda \mathbf{I}) = 0. \tag{1.6.3}$$

[15] Otherwise, the nullspace would not contain any vectors, leaving only $\mathbf{x} = \mathbf{0}$ as a solution.

The above equation is termed the characteristic equation, and also points to how one should go about computing the eigenvalues of a matrix. For example, let

$$\mathbf{A} = \begin{bmatrix} 4 & -5 \\ 2 & -3 \end{bmatrix} \tag{1.6.4}$$

and thus,

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 4-\lambda & -5 \\ 2 & -3-\lambda \end{vmatrix} = (4-\lambda)(-3-\lambda) + 10 = 0, \tag{1.6.5}$$

or $\lambda = -1$ or 2. By inserting the eigenvalues into (1.6.1), we obtain

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{bmatrix} 5 \\ 2 \end{bmatrix} \tag{1.6.6}$$

The eigendecomposition is a very powerful tool in data processing, and is frequently used in a variety of algorithms. One reason for this is that:

> If $\mathbf{A}$ has n distinct eigenvalues, then $\mathbf{A}$ has n linearly independent eigenvectors, and is *diagonalizable*.

If $\mathbf{A}$ has n linearly independent eigenvectors, then if placed as columns of a matrix $\mathbf{S}$, we can diagonalize $\mathbf{A}$ as

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \Lambda \stackrel{\triangle}{=} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{A} = \mathbf{S}\Lambda\mathbf{S}^{-1}. \tag{1.6.7}$$

That is, $\mathbf{A}$ is *similar* to a diagonal matrix. In general, we say a matrix $\mathbf{B}$ is similar to a matrix $\mathbf{A}$ is there exists a nonsingular matrix $\mathbf{S}$ such that

$$\mathbf{B} = \mathbf{S}\mathbf{A}\mathbf{S}^{-1}. \tag{1.6.8}$$

The transformation $\mathbf{A} \rightarrow \mathbf{S}\mathbf{A}\mathbf{S}^{-1}$ is often called a *similarity transformation* by the *similarity matrix* $\mathbf{S}$. We note that:

> If $\mathbf{A}$ and $\mathbf{B}$ are *similar*, then they have the same eigenvalues.

To illustrate the importance of (1.6.7), we note as an example that it immediately implies that

$$\mathbf{A}^k = \mathbf{S}\Lambda^k\mathbf{S}^{-1}. \tag{1.6.9}$$

Check this for $\mathbf{A}^2$ using (1.6.4). Similarly,

$$\mathbf{A}^{-1} = \mathbf{S}\Lambda^{-1}\mathbf{S}^{-1}, \tag{1.6.10}$$

with Λ^{-1} having $\left\{\lambda_k^{-1}\right\}$ as diagonal elements. Check! We proceed with some observations; the eigenvalues of $\mathbf{A}^T$ are the same as those of $\mathbf{A}$, whereas the eigenvalues of $\mathbf{A}^H$ are the complex conjugate of those of $\mathbf{A}$. Further,

$$\prod_{k=1}^{n} \lambda_k \;=\; \det \mathbf{A} \tag{1.6.11}$$

$$\sum_{k=1}^{n} \lambda_k \;=\; \operatorname{tr} \mathbf{A} \tag{1.6.12}$$

where $\operatorname{tr} \mathbf{A}$ denotes the *trace*[16] of $\mathbf{A}$, i.e, the sum of the diagonal elements of $\mathbf{A}$. Check this for the example in (1.6.4). Note that eigenvalues can be complex-valued, and can be zero. As mentioned, we are often encountering Hermitian matrices. In this case:

> A Hermitian matrix has *real-valued* eigenvalues and eigenvectors forming an *orthonormal* set.

Thus, for a Hermitian matrix, (1.6.7) can be written as

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^H = \sum_{k=1}^{n} \lambda_k \mathbf{x}_k \mathbf{x}_k^H \tag{1.6.13}$$

where $\mathbf{U}$ is a *unitary* matrix whose columns are the eigenvectors of $\mathbf{A}$. Equation (1.6.13) is often referred to as the *spectral theorem*. Note that $\mathbf{x}_k \mathbf{x}_k^H$ is nothing but the projection onto the kth eigenvector, i.e.,

$$\mathbf{\Pi}_{\mathbf{x}_k} = \mathbf{x}_k \mathbf{x}_k^H . \tag{1.6.14}$$

Using (1.6.14), one can therefore view the spectral theorem as

$$\mathbf{A} = \sum_{k=1}^{n} \lambda_k \mathbf{\Pi}_{\mathbf{x}_k} \tag{1.6.15}$$

The eigenvectors form a basis spanning $\mathcal{R}(\mathbf{A})$. The spectral theorem is closely related to the Karhunen-Loève transform, briefly described in Appendix B.

We proceed by briefly discussing quadratic minimization; this problem is so frequently reoccurring (recall for instance the least squares problem in (1.4.4)) that it deserves a moment of our attention. As an example, consider the minimization of the function

$$f = 2x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + 2x_3^2 \tag{1.6.16}$$

[16] An often used result is the fact that $\operatorname{tr}\{\mathbf{AB}\} = \operatorname{tr}\{\mathbf{BA}\}$.

Finding the minimum is easily done by setting the partial derivatives to zero, yielding $x_1 = x_2 = x_3 = 0$. But which solution different from the zero solution minimizes (1.6.16)? Rewriting (1.6.16) as

$$f = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{x}^T \mathbf{A} \mathbf{x} \tag{1.6.17}$$

we conclude that it must be $\mathbf{A}$ that determines the minimum of f. Let

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$$

and define the *Rayleigh's quotient*

$$R(\mathbf{x}) = \frac{\mathbf{x}^H \mathbf{A} \mathbf{x}}{\mathbf{x}^H \mathbf{x}} \tag{1.6.18}$$

Then, according to the Rayleigh-Ritz theorem, it holds that

$$\lambda_1 \geq \frac{\mathbf{x}^H \mathbf{A} \mathbf{x}}{\mathbf{x}^H \mathbf{x}} \geq \lambda_n \tag{1.6.19}$$

and that $R(\mathbf{x})$ is minimized by the eigenvector corresponding to the minimum eigenvalue, λ_n. Similarly, $R(\mathbf{x})$ is maximized by the dominant eigenvector, corresponding to the largest eigenvalue, λ_1. For our example, the eigenvalues of $\mathbf{A}$ are $\lambda_1 = 3.41$, $\lambda_2 = 2$ and $\lambda_3 = 0.58$. The minimum of f, excluding the zero solution, is thus 0.58, obtained for

$$\mathbf{x}_1 = \begin{bmatrix} 0.5 & 0.7071 & 0.5 \end{bmatrix}^T$$

Concluding this section, we introduce the notion of positive definiteness.

> The Hermitian matrix $\mathbf{A}$ is said to be *positive definite* if and only if $\mathbf{x}^H \mathbf{A} \mathbf{x}$ can be written as a sum of n independent squares, i.e.,
>
> $$\mathbf{x}^H \mathbf{A} \mathbf{x} > 0, \quad \forall \mathbf{x} \quad (\text{except } \mathbf{x} = \mathbf{0})$$

If we also allow for equality, $\mathbf{A}$ is said to *positive semi-definite*. Note that

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i \quad \Rightarrow \quad \mathbf{x}_i^H \mathbf{A} \mathbf{x}_i = \lambda_i \mathbf{x}_i^H \mathbf{x}_i = \lambda \tag{1.6.20}$$

as the eigenvectors are orthonormal. We conclude that a positive definite matrix has positive eigenvalues. The reverse is also true; if a matrix has only positive eigenvalues, it is positive definite (the proof is left as an exercise for the reader).

> The Hermitian matrix $\mathbf{A}$ is said to be *positive definite* if and only if $\mathbf{A}$ has positive (real-valued) eigenvalues.

In passing, we note that if $\mathbf{A}$ is positive semi-definite, it implies that there exists a matrix $\mathbf{C}$ (of rank n) such that $\mathbf{A} = \mathbf{C}\mathbf{C}^H$ (compare with the discussion at the end of Section 1.2). For this case, we note that $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{C})$.

1.7 Singular value decomposition

As mentioned in previous section, the eigendecomposition in (1.6.7) is a very powerful tool, finding numerous applications in all fields of signal processing. However, the decomposition is limited by the fact that it only exists for *square* matrices. We will now proceed to study the more general case.

> Every matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ can be factored as
>
> $$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$$
>
> where $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$ is diagonal, and $\mathbf{U} \in \mathbb{C}^{m\times m}$, $\mathbf{V} \in \mathbb{C}^{n\times n}$ are unitary matrices.

This is the famous *singular value decomposition*[17] (SVD), being one of the most powerful tools in signal processing. The diagonal elements of $\mathbf{\Sigma}$,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_p & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} \tag{1.7.1}$$

are denoted the *singular values* of $\mathbf{A}$. Here, $p = \text{rank}(\mathbf{A}) \leq \min(m, n)$, and $\sigma_k \geq 0$; in general

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$$

We note that

$$\mathbf{A}^H\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^H\mathbf{\Sigma}\mathbf{V}^H \quad \Rightarrow \quad \mathbf{A}^H\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{\Sigma}^H\mathbf{\Sigma} \tag{1.7.2}$$

which implies that σ_i^2 are the eigenvalues of $\mathbf{A}^H\mathbf{A}$ (similarly, it is easy to show that σ_i^2 are the eigenvalues of $\mathbf{A}\mathbf{A}^H$). In the special case when $\mathbf{A}$ is a square matrix, the

[17] The singular value decomposition has a rich history, dating back to the independent work by the Italian differential geometer E. Beltrami in 1873 and the French algebraist C. Jordan in 1874. Later, J. J. Sylvester in 1889, being unaware of the previous work by Beltrami and Jordan, gave a third proof for the factorization. Neither of these used the term singular values, and their work was repeatedly rediscovered up to 1939 when C. Eckart and G. Young gave a clear and complete statement of the decomposition for a rectangular matrix. In parallel to the algebraists work, researchers working with the theory of integral equations rediscovered the decomposition. In 1907, E. Schmidt published a general theory for symmetric and nonsymmetric kernels, making use of the decomposition. In 1919, É. Picard's generalized this theory, coining Schmidt's "eigenvalues" as singular values ("valeurs singulières"). It took many years of rediscovering the decomposition, under different names, until A. Horn in 1954 writing a paper on matrix theory used the term "singular values" in the context of matrices, a designation that have since become standard terminology (it should be noted that in the Russian literature, one also sees singular values referred to as *s-numbers*).

singular values coincide with the modulus of the eigenvalues[18]. Further, we note that as only the first $p = \text{rank}(\mathbf{A})$ singular values are non-zero, we can rewrite the SVD as

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^H \\ \mathbf{V}_2^H \end{bmatrix} && (1.7.3)\\
&= \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^H && (1.7.4)\\
&= \sum_{k=1}^{p} \sigma_k \mathbf{u}_k \mathbf{v}_k^H && (1.7.5)
\end{aligned}$$

where $\mathbf{\Sigma}_1 \in \mathbb{R}^{p \times p}$, $\mathbf{U}_1 \in \mathbb{C}^{m \times p}$, $\mathbf{U}_2 \in \mathbb{C}^{m \times (n-p)}$, $\mathbf{V}_1 \in \mathbb{C}^{n \times p}$, $\mathbf{V}_2 \in \mathbb{C}^{n \times (n-p)}$, and $\mathbf{u}_k$ and $\mathbf{v}_k$ denote the kth column of $\mathbf{U}_1$ and $\mathbf{V}_1$, respectively. As the rank of a matrix is equal to the number of nonzero singular values, $\mathbf{\Sigma}_2 = \mathbf{0}$. Using (1.7.4), we can conclude that

$$\begin{aligned}
\mathcal{R}(\mathbf{A}) &= \left\{ \mathbf{b} \in \mathbb{C}^m : \mathbf{b} = \mathbf{A}\mathbf{x} \right\} \\
&= \left\{ \mathbf{b} \in \mathbb{C}^m : \mathbf{b} = \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^H \mathbf{x} \right\} \\
&= \left\{ \mathbf{b} \in \mathbb{C}^m : \mathbf{b} = \mathbf{U}_1 \mathbf{y} \right\} \\
&= \text{span}\,(\mathbf{U}_1)
\end{aligned}$$

The dominant singular vectors forms a basis spanning $\mathcal{R}(\mathbf{A})$! Using a similar technique, we find that

$$\begin{aligned}
\mathcal{N}(\mathbf{A}) &= \text{span}\,(\mathbf{V}_2) \\
\mathcal{R}(\mathbf{A}^H) &= \text{span}\,(\mathbf{V}_1) \\
\mathcal{N}(\mathbf{A}^H) &= \text{span}\,(\mathbf{U}_2)
\end{aligned}$$

which also yields that

$$\begin{aligned}
\mathbf{\Pi}_{\mathbf{A}} &= \mathbf{U}_1 \mathbf{U}_1^H && (1.7.6)\\
\mathbf{\Pi}_{\mathbf{A}}^{\perp} &= \mathbf{I} - \mathbf{U}_1 \mathbf{U}_1^H = \mathbf{U}_2 \mathbf{U}_2^H && (1.7.7)
\end{aligned}$$

For a moment returning to the least squares minimization in (1.4.4),

$$\begin{aligned}
\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 &= \min_{\mathbf{x}} \|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{x} - \mathbf{b}\|_2^2 \\
&= \min_{\mathbf{x}} \|\mathbf{\Sigma}\mathbf{V}^H\mathbf{x} - \mathbf{U}^H\mathbf{b}\|_2^2 \\
&= \min_{\mathbf{x}} \|\mathbf{\Sigma}\mathbf{y} - \hat{\mathbf{b}}\|_2^2 && (1.7.8)
\end{aligned}$$

where we defined $\mathbf{y} = \mathbf{V}^H\mathbf{x}$ and $\hat{\mathbf{b}} = \mathbf{U}^H\mathbf{b}$. Minimizing (1.7.8) yields

$$\mathbf{y} = \mathbf{\Sigma}^{\dagger}\hat{\mathbf{b}} \quad \Rightarrow \quad \mathbf{x} = \mathbf{V}\mathbf{\Sigma}^{\dagger}\mathbf{U}^H\mathbf{b} \tag{1.7.9}$$

[18] Due to the numerical robustness of the SVD algorithm, it is in practice often preferable to use the SVD instead of the eigenvalue decomposition to compute the eigenvalues.

Comparing with (1.4.6), yields

$$\mathbf{A}^{\dagger} = \mathbf{V}\mathbf{\Sigma}^{\dagger}\mathbf{U}^{H} \tag{1.7.10}$$

and the conclusion that the least squares solution can immediately be obtained using the SVD. We also note that the singular values can be used to compute the 2-norm of a matrix,

$$\|\mathbf{A}\|_2 = \sigma_1. \tag{1.7.11}$$

As is clear from the above brief discussion, the SVD is a very powerful tool finding use in numerous applications. We refer the interested reader to [3] and [4] for further details on the theoretical and representational aspects of the SVD.

1.8 Total least squares

As an example of the use of the SVD, we will briefly touch on total least squares (TLS). As mentioned in our earlier discussion on the least squares minimization, it can be viewed as finding the vector $\mathbf{x}$ such that

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \tag{1.8.1}$$

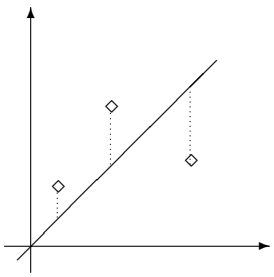
Here, we are implicitly assuming that the errors are confined to the "observation" vector $\mathbf{b}$, such that

$$\mathbf{A}\mathbf{x} = \mathbf{b} + \Delta\mathbf{b} \tag{1.8.2}$$

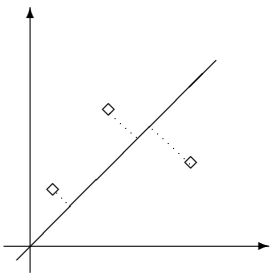
It is often relevant to also assume the existence of errors in the "data" matrix, i.e.,

$$(\mathbf{A} + \Delta\mathbf{A})\mathbf{x} = (\mathbf{b} + \Delta\mathbf{b}) \tag{1.8.3}$$

Thus, we seek the vector $\mathbf{x}$, such that $(\mathbf{b} + \Delta\mathbf{b}) \in \mathcal{R}(\mathbf{A} + \Delta\mathbf{A})$, which minimize the norm of the perturbations. Geometrically, this can be viewed as finding the vector $\mathbf{x}$ minimizing the (squared) *total distance*, as compared to finding the minimum (squared) *vertical distance*; this is illustrated in the figure below for a line.



(a) LS: Minimize vertical distance to line.



(b) TLS: Minimize total distance to line.

Let $[\mathbf{A}\,|\,\mathbf{B}]$ denote the matrix obtained by concatenating the matrix $\mathbf{A}$ with the matrix $\mathbf{B}$, stacking the columns of $\mathbf{B}$ to the right of the columns of $\mathbf{A}$. Note that,

$$\begin{aligned}\mathbf{0} &= \left[\begin{array}{c|c} \mathbf{A}+\Delta\mathbf{A} & \mathbf{b}+\Delta\mathbf{b} \end{array}\right]\left[\begin{array}{c} \mathbf{x} \\ -1 \end{array}\right] \\ &= \left[\begin{array}{c|c} \mathbf{A} & \mathbf{b} \end{array}\right]\left[\begin{array}{c} \mathbf{x} \\ -1 \end{array}\right] + \left[\begin{array}{c|c} \Delta\mathbf{A} & \Delta\mathbf{b} \end{array}\right]\left[\begin{array}{c} \mathbf{x} \\ -1 \end{array}\right] \end{aligned} \tag{1.8.4}$$

and let

$$\begin{aligned}\mathbf{C} &= \left[\begin{array}{c|c} \mathbf{A} & \mathbf{b} \end{array}\right] \\ \mathbf{D} &= \left[\begin{array}{c|c} \Delta\mathbf{A} & \Delta\mathbf{b} \end{array}\right] \end{aligned} \tag{1.8.5}$$

Thus, in order for (1.8.4) to have a solution, the augmented vector $[\mathbf{x}^T, -1]^T$ must lie in the nullspace of $\mathbf{C}+\mathbf{D}$, and in order for the solution to be non-trivial, the perturbation $\mathbf{D}$ must be such that $\mathbf{C}+\mathbf{D}$ is rank deficient[19]. The TLS solution finds the $\mathbf{D}$ with smallest norm that makes $\mathbf{C}+\mathbf{D}$ rank deficient; put differently, the TLS solution is the vector minimizing

$$\min_{\Delta\mathbf{A},\Delta\mathbf{b},\mathbf{x}} \|\,[\Delta\mathbf{A}\ \Delta\mathbf{b}]\,\|_F^2 \quad \text{such that} \quad (\mathbf{A}+\Delta\mathbf{A})\mathbf{x} = \mathbf{b}+\Delta\mathbf{b} \tag{1.8.6}$$

where

$$\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^H\mathbf{A}) = \sum_{k=1}^{\min(m,n)} \sigma_k^2 \tag{1.8.7}$$

denotes the *Frobenius* norm, with σ_k being the kth singular value of $\mathbf{A}$. Let $\mathbf{C} \in \mathbb{C}^{m\times(n+1)}$. Recalling (1.7.5), we write

$$\mathbf{C} = \sum_{k=1}^{n+1} \sigma_k \mathbf{u}_k \mathbf{v}_k^H \tag{1.8.8}$$

The matrix of rank n closest to $\mathbf{C}$ can be shown to be the matrix formed using only the n dominant singular values of $\mathbf{C}$, i.e.,

$$\tilde{\mathbf{C}} = \sum_{k=1}^{n} \sigma_k \mathbf{u}_k \mathbf{v}_k^H \tag{1.8.9}$$

Thus, to ensure that $\mathbf{C}+\mathbf{D}$ is rank deficient, we let

$$\mathbf{D} = -\sigma_{n+1}\mathbf{u}_{n+1}\mathbf{v}_{n+1}^H \tag{1.8.10}$$

[19] Recall that if $\mathbf{C}+\mathbf{D}$ is full rank, the nullspace only contains the null vector.

For this choice of $\mathbf{D}$, $\mathbf{C}+\mathbf{D}$ does not contain the vector[20] $\mathbf{v}_{n+1}$, and the solution to (1.8.4) must thus be a multiple α of $\mathbf{v}_{n+1}$, i.e.,

$$\begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \alpha \mathbf{v}_{n+1} \tag{1.8.11}$$

yielding

$$\mathbf{x}_{TLS} = -\frac{\mathbf{v}_{n+1}(1:n)}{\mathbf{v}_{n+1}(n+1)} \tag{1.8.12}$$

We note in passing that one can similarly write the solution to the TLS matrix equation $(\mathbf{A}+\Delta\mathbf{A})\mathbf{X} = (\mathbf{B}+\Delta\mathbf{B})$, where $\mathbf{B} \in \mathbb{C}^{m\times k}$, as

$$\mathbf{X}_{TLS} = -\mathbf{V}_{12}\mathbf{V}_{22}^{-1} \tag{1.8.13}$$

if $\mathbf{V}_{22}^{-1}$ exists, where

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \tag{1.8.14}$$

with $\mathbf{V}_{11} \in \mathbb{C}^{n\times n}$, $\mathbf{V}_{22} \in \mathbb{C}^{k\times k}$ and $\mathbf{V}_{12}, \mathbf{V}_{21}^T \in \mathbb{C}^{n\times k}$. The TLS estimate yields a strongly consistent estimate of the true solution when the errors in the concatenated matrix $[\mathbf{A}\ \mathbf{b}]$ are rowwise independently and identically distributed with zero mean and equal variance. If, in addition, the errors are normally distributed, the TLS solution has maximum likelihood properties. We refer the interested reader to [4] and [5] for further details on theoretical and implementational aspects of the TLS.

[20] The vector $\mathbf{v}_{n+1}$ now lies in $\mathcal{N}(\mathbf{C}+\mathbf{D})$.

Appendix A

BANACH AND HILBERT SPACES

Mister Data, there is a subspace communication for you
Star Trek, the Next Generation

In this appendix, we will provide some further details on the notion of vector spaces, and in particular on the definition of Banach and Hilbert spaces, the latter being the vector space of primary interest in signal processing.

A *vector space* is by definition a set of vectors $\mathbf{X}$ for which any vectors in the set, say $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, will satisfy that $\mathbf{x}_1 + \mathbf{x}_2 \in \mathbf{X}$ and $\alpha\mathbf{x}_1 \in \mathbf{X}$ for some (possibly complex-valued) constant α, as well as the vector space axioms:

(1) $\mathbf{x}_1 + \mathbf{x}_2 = \mathbf{x}_2 + \mathbf{x}_1$ (commutability of addition),

(2) $\mathbf{x}_1 + (\mathbf{x}_2 + \mathbf{x}_3) = (\mathbf{x}_1 + \mathbf{x}_2) + \mathbf{x}_3$ (associativity of addition),

(3) $\exists \mathbf{0} \in \mathbf{X} : \mathbf{0} + \mathbf{x}_1 = \mathbf{x}_1$ (existence of null element),

(4) $\forall \mathbf{x}_1 \in \mathbf{X} \Rightarrow \exists - \mathbf{x}_1 \in \mathbf{X} : \mathbf{x}_1 + (-\mathbf{x}_1) = \mathbf{0}$ (existence of the inverse element),

(5) $1 \cdot \mathbf{x}_1 = \mathbf{x}_1$ (unitarism),

(6) $\alpha(\beta\mathbf{x}_1) = (\alpha\beta)\mathbf{x}_1$ (associativity with respect to number multiplication),

(7) $\alpha(\mathbf{x}_1 + \mathbf{x}_2) = \alpha\mathbf{x}_1 + \alpha\mathbf{x}_2$ (distributivity with respect to vector additism),

(8) $(\alpha + \beta)\mathbf{x}_1 = \alpha\mathbf{x}_1 + \beta\mathbf{x}_1$ (distributivity with respect to number additism).

Any element in a vector space is termed a vector. Often, we are primarily interested in vector spaces that allow some form of distance measure, normally termed the

norm. Such a space is termed a *Banach*[1] *space*. A Banach space is a *complete*[2] vector space that admits a norm, $\|\mathbf{x}\|$. A norm always satisfies

$$\|\mathbf{x}\| \geq 0 \tag{A.1.1}$$

$$\|\lambda\mathbf{x}\| = |\lambda| \, \|\mathbf{x}\| \tag{A.1.2}$$

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \tag{A.1.3}$$

for any $\mathbf{x}$, $\mathbf{y}$ in the Banach space and $\lambda \in \mathbb{C}$. The so-called p-norm of a (complex-valued) discrete sequence, $x(k)$, is defined as

$$\|\mathbf{x}\|_p \triangleq \left(\sum_{k=-\infty}^{\infty} |x(k)|^p \right)^{\frac{1}{p}}, \tag{A.1.4}$$

for any integer $p > 0$. Similarly, for a (complex-valued) continuous function, $f(t)$, the p-norm is defined as

$$\|f\|_p \triangleq \left(\int_{-\infty}^{\infty} |f(t)|^p \, dt \right)^{\frac{1}{p}} < \infty. \tag{A.1.5}$$

Here, we are primarily concerned with the 2-norm, also called the Euclidean norm as $\|\mathbf{x} - \mathbf{y}\|_2$ yields the Euclidean distance between any two vectors, $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$. For instance, for

$$\mathbf{x} = \begin{bmatrix} 1 & 3 & 2 \end{bmatrix}^T \tag{A.1.6}$$

$$\mathbf{y} = \begin{bmatrix} 0 & 0 & 2 \end{bmatrix}^T \tag{A.1.7}$$

the 2-norms are $\|\mathbf{x}\|_2 = \sqrt{1^2 + 3^2 + 2^2} = \sqrt{14}$ and $\|\mathbf{y}\|_2 = 2$. Similarly, the Euclidean distance is $\|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{10}$. If not specified, we will here assume that the used norm is the 2-norm. Note that the definitions allows for both *infinite-dimensional* vectors and vectors defined as *continuous functions*.

As is obvious from the above discussion, a Banach space is quite general, and it is often not enough to just require that a vector lies in a Banach space. For this reason, we will further restrict our attention to work only with Banach spaces that has an inner product (this to define angles and orthogonality). A *Hilbert space* is

[1] Stefan Banach (1892-1945) was born in Kraków, Austria-Hungary (now Poland). He received his doctorate in mathematics at Lvov Technical University in 1920, and in 1924, he was promoted to full professor. He used to spend his days in the cafés of Lvov, discussing mathematics with his colleagues. He suffered hardship during the Nazi occupation and died of lung cancer shortly after the liberation. His Ph.D. thesis is often said to mark the birth of modern functional analysis.

[2] A complete vector space is a vector space where every Cauchy sequence converges to an element in the space. A sequence is a Cauchy sequence if for any $\epsilon > 0$, $\|x(n) - x(p)\| < \epsilon$ for large enough n and p.

a Banach space with an inner product, $\langle \mathbf{x}, \mathbf{y} \rangle$. This inner product also defines the length of a vector in the space,

$$\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}. \tag{A.1.8}$$

An example of a Hilbert space is $\mathbf{l}_2(\mathbb{C})$, the space of "square summable sequences", i.e., the space of (complex-valued) sequences satisfying $\|\mathbf{x}\|_2 < \infty$, where the inner product is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^H \mathbf{y} = \sum_{k=-\infty}^{\infty} x^*(k) y(k), \tag{A.1.9}$$

and where $(\cdot)^*$ denotes conjugate. Note that although every Hilbert space is also a Banach space, the converse need not hold. Another example is $\mathbf{L}_2(\mathbb{C})$, the space of square integrable functions, i.e., (complex-valued) functions satisfying $\|f\|_2 < \infty$, where the inner product is

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f^*(x) g(x)\, dx. \tag{A.1.10}$$

In words, this means simply that the inner-product is finite, i.e., any vector in a Hilbert space (even if the vector is infinite-dimensional or is a continuous function) has finite length. Obviously, this is for most practical purposes not a very strong restriction.

Appendix B

THE KARHUNEN-LOÈVE TRANSFORM

E ne l'idolo suo si trasmutava
Dante

In this appendix, we will briefly describe the Karhunen-Loève transform (KLT). Among all linear transforms, the KLT is the one which best approximates a stochastic process in the least square sense. Furthermore, the coefficients of the KLT are *uncorrelated*. These properties make the KLT very interesting for many signal processing applications, such as coding and pattern recognition.

Let

$$\mathbf{x}_N = [x_1, \ldots, x_N]^T \tag{B.1.1}$$

be a zero-mean, stationary and ergodic random process with covariance matrix (see, e.g., [8]).

$$\mathbf{R}_{xx} = E\{\mathbf{x}_N \mathbf{x}_N^H\} \tag{B.1.2}$$

where $E\{\cdot\}$ denotes the expectation operator. The eigenvalue problem

$$\mathbf{R}_{xx}\mathbf{u}_j = \lambda_j \mathbf{u}_j, \quad \text{for} \quad j = 0, \ldots, N-1, \tag{B.1.3}$$

yields the j:th eigenvalue, λ_j, as well as the j:th eigenvector, $\mathbf{u}_j$. Given that $\mathbf{R}_{xx}$ is a covariance matrix[1], all eigenvalues will be real and non-negative (see also Section 1.6). Furthermore, N orthogonal eigenvectors can always be found, i.e.,

$$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_N] \tag{B.1.4}$$

will be a unitary matrix. The KLT of $\mathbf{x}_N$ is found as

$$\boldsymbol{\alpha} = \mathbf{U}^{\boldsymbol{H}}\mathbf{x}_{\boldsymbol{N}} \tag{B.1.5}$$

[1] A covariance matrix is always positive (semi) definite.

where the KLT coefficients,

$$\boldsymbol{\alpha} = [\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_N}]^{\boldsymbol{T}}, \tag{B.1.6}$$

will be uncorrelated, i.e.,

$$E\{\alpha_j \alpha_k\} = \lambda_j \delta_{jk} \quad \text{for} \quad j, k = 1, \ldots, N. \tag{B.1.7}$$

We note that this yields that

$$\mathbf{U}^{\boldsymbol{H}} \mathbf{R}_{\boldsymbol{xx}} \mathbf{U} = \boldsymbol{\Lambda} \quad \Leftrightarrow \quad \mathbf{R}_{\boldsymbol{xx}} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^{\boldsymbol{H}} \tag{B.1.8}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N)$.

Appendix C

USEFUL FORMULAE

This one's tricky. You have to use imaginary numbers, like eleventeen
Calvin & Hobbes

Euler's formula

$$e^{i\omega t} = \cos(\omega t) + i\sin(\omega t) \tag{C.1.1}$$

Cauchy-Schwartz inequality

For all $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$, it hold that

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}\sqrt{\langle \mathbf{y}, \mathbf{y} \rangle} \tag{C.1.2}$$

with equality if and only if $\mathbf{x}$ and $\mathbf{y}$ are dependent. Similarly, for two continuous functions, f_x and g_x, it holds that

$$\int_{-\infty}^{\infty} |f_x|^2\, dx \cdot \int_{-\infty}^{\infty} |g_x|^2\, dx \geq \left| \int_{-\infty}^{\infty} f_x^* g_x\, dx \right|^2 \tag{C.1.3}$$

Some standard integrals

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega t}\, d\omega = \delta_t \tag{C.1.4}$$

$$\int_{-\infty}^{\infty} e^{-ax^2}\, dx = \sqrt{\frac{\pi}{a}} \tag{C.1.5}$$

$$\int_{-\infty}^{\infty} x^2 e^{-ax^2}\, dx = \frac{1}{2a}\sqrt{\frac{\pi}{a}} \tag{C.1.6}$$

BIBLIOGRAPHY

[1] G. Strang, *Linear algebra and its applications, 3rd Ed.* Thomson Learning, 1988.

[2] R. A. Horn and C. A. Johnson, *Matrix Analysis*. Cambridge, England: Cambridge University Press, 1985.

[3] R. A. Horn and C. A. Johnson, *Topics in Matrix Analysis*. Cambridge, England: Cambridge University Press, 1991.

[4] G. H. Golub and C. F. V. Loan, *Matrix Computations (3rd edition)*. The John Hopkins University Press, 1996.

[5] S. V. Huffel and J. Vandevalle, *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM Publications, Philadelphia, 1991.

[6] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, N.J.: Prentice Hall, 1997.

[7] J. C. F. Gauss, *Demonstratio Nova Theorematis Omnem Functionem Algebraicam Rationalem Integram Unius Variabilis in Factores Reales Primi Vel Secundi Gradus Resolve Posse*. PhD thesis, University of Helmstedt, Germany, 1799.

[8] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. New York: John Wiley and Sons, Inc., 1996.