Refresher in Mathematics

M2 Data Science - Centrale Lille

Andrea Natale

September 19, 2022

This course is divided into three chapters. The first collects some basic tools from linear algebra. The second is devoted to the Singular Value Decomposition and its application to the resolution of inverse problems and to the Principal Component Analysis. The third chapter deals with matrix norms which are used to define low rank projections and study the conditioning of linear systems. The content of these notes is mostly based on:

1. Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013
2. Gilbert Strang. *Linear algebra and its applications.* Belmont, CA: Thomson, Brooks/Cole, 2006

plus additional references in the notes.

Throughout the course we will use the following notation:

- $\mathbb{K}$ denotes the field $\mathbb{R}$ or $\mathbb{C}$;
- $E, F, \dots$ denote vector spaces over $\mathbb{K}$;
- $\mathcal{L}(E, F)$ denotes the space of linear maps from E to F;
- $\mathcal{L}(E)$ denotes the space of linear maps from E to E;
- $\mathcal{M}_{m,n}(\mathbb{K})$ denotes the spaces of matrices over $\mathbb{K}$ with m rows and n coloumns;
- $\mathcal{M}_n(\mathbb{K})$ denotes the spaces of matrices over $\mathbb{K}$ with n rows and n coloumns;

1 Basic facts from Linear Algebra

1.1 Linearly independent vectors and bases

Given a vector space E over $\mathbb{K}$, a set of vectors $\{e_1, \dots, e_n\} \subset E$ is *linearly dependent* if there exist n scalars $\alpha_1, \dots, \alpha_n \in \mathbb{K}$ not all zero, such that $\sum_i \alpha_i e_i = 0$. The set is *linearly independent* if no such collection of scalars exists. A set of n linearly independent vectors $\{e_1, \dots, e_n\}$ is a *basis* of E if for any $x \in E$, $\{x, e_1, \dots, e_n\}$ is linearly dependent, or equivalently if any vector x of E can be expressed as a linear combination of vectors in the set $\mathbf{e} = \{e_1, \dots, e_n\}$:

$$x = \sum_{i=1}^{n} \alpha_i e_i$$

with $\alpha_i \in \mathbb{K}$. In this case, the collection of scalars $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{K}^n$ is uniquely defined. Such vector is the vector of *coordinates* of x with respect to the basis $\mathbf{e}$ and we will also denote it as follows:

$$[x]_{\mathbf{e}} = \alpha\,.$$

Given a vector $x \in \mathbb{K}^n$, we will usually denote by x_i the ith coordinate of x with respect to the canonical basis $\{e_1, \dots, e_n\}$ defined by $e_1 = (1, 0, \dots, 0)$, $\dots$, $e_n = (0, 0, \dots, 1)$.

The *dimension* of the vector space E is the maximal number of linearly independent vectors in E, or equivalently the number of vectors in any basis of E.

1.2 Norms, inner products and orthogonality

Definition 1.2.1 (Norm)**.** A norm on a vector space E over $\mathbb{K}$ is a function $\|\cdot\| : E \to \mathbb{R}^+$ verifying the following properties:

- (Positive definiteness) $\forall\, x \in E\,, \quad \|x\| \geq 0$ and $\|x\| = 0 \iff x = 0$;
- (1-Homogeneity) $\forall\, x \in E\,, \forall\, \lambda \in \mathbb{K}\,, \quad \|\lambda x\| = |\lambda| \|x\|$;
- (Triangular inequality) $\forall\, x, y \in E\,, \quad \|x + y\| \leq \|x\| + \|y\|$.

Example (p-norms on $\mathbb{K}^n$)*.* For $p \geq 1$, the p-norm of $x = (x_1, \dots, x_d) \in \mathbb{K}^d$ is defined by:

$$\|x\|_p \coloneqq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

We also define the maximum norm by: $\|x\|_\infty \coloneqq \max\{|x_1|, \dots, |x_n|\}$.

Definition 1.2.2 (Equivalent norms)**.** Two norms $\|\cdot\|$ and $\|\cdot\|'$ on a $\mathbb{K}$-vector space E are equivalent if and only if there exists a constant $C > 0$ such that for all $x \in E$, $\|x\| \leq C\|x\|'$ and $\|x\|' \leq C\|x\|$.

Theorem 1.2.3. *All norms on a finite-dimensional vector space over $\mathbb{K}$ are equivalent.*

Norms allow us to quantify the intensity of vectors and to introduce a notion of convergence. Given a normed vector space $(E, \|\cdot\|)$, a sequence $(x_k)_{k\in\mathbb{N}} \subset E$ converges to x, if and only if

$$\lim_{n\to\infty} \|x_k - x\| = 0\,.$$

If E is finite-dimensional, convergence of a sequence on E with respect to a norm implies convergence with respect to any other norm.

Definition 1.2.4 (Inner product)**.** An inner product on a real vector space E is a map $\langle\cdot,\cdot\rangle : E \times E \to \mathbb{R}$ satisfying the following properties:

- (Bilinearity) $\forall\, x, y, z \in E$, $\forall\, \lambda \in \mathbb{R}$,

$$\langle x + \lambda y, z\rangle = \langle x, z\rangle + \lambda\langle y, z\rangle,$$

$$\langle x, y + \lambda z\rangle = \langle x, y\rangle + \lambda\langle x, z\rangle;$$

- (Symmetry) $\forall\, x, y \in E, \quad \langle x, y\rangle = \langle y, x\rangle$;
- (Positive definiteness) $\forall x \in E, \quad \langle x, x\rangle \geq 0$ and $\langle x, x\rangle = 0 \iff x = 0$.

If E is a complex vector space an inner product is a map $\langle\cdot,\cdot\rangle : E \times E \to \mathbb{C}$, and the first two properties are replaced by:

- (Sesquilinearity) $\forall\, x, y, z \in E$, $\forall\, \lambda \in \mathbb{C}$,

$$\langle x + \lambda y, z\rangle = \langle x, z\rangle + \overline{\lambda}\langle y, z\rangle,$$

$$\langle x, y + \lambda z\rangle = \langle x, y\rangle + \lambda\langle x, z\rangle;$$

- (Hermiticity) $\forall\, x, y \in E, \quad \langle x, y\rangle = \overline{\langle y, x\rangle}$.

Example. Here are some important examples of inner products:

- Canonical inner product on $\mathbb{R}^n$: $x = (x_i)_i, y = (y_i)_i \in \mathbb{R}^n\,, \langle x, y\rangle \coloneqq \sum_i x_i y_i$
- Canonical inner product on $\mathbb{C}^n$: $x = (x_i)_i, y = (y_i)_i \in \mathbb{R}^n\,, \langle x, y\rangle \coloneqq \sum_i \overline{x_i} y_i$
- L^2 inner product on $C([0,1])$: $f, g \in C([0,1]), \quad \langle f, g\rangle \coloneqq \int_0^1 f(t)g(t)\,\mathrm{d}t$.

Proposition 1.2.5. *Let E be a vector space over $\mathbb{K}$ equipped with an inner product $\langle\cdot,\cdot\rangle_E$, then the map $\|\cdot\|_E : E \to \mathbb{R}^+$ defined by $\|x\|_E = \sqrt{\langle x, x\rangle_E}$, for all $x \in E$, is a norm.*

Example. The norm $\|\cdot\|_2$ is the norm induced by the canonical inner product on $\mathbb{K}^n$.

Note that the fact that $\|\cdot\|_E$ verifies the triangular inequality can be easily deduced from the Cauchy-Schwarz inequality:

Proposition 1.2.6 (Cauchy-Schwarz inequality)**.** *For any $x, y \in E$,*

$$|\langle x, y\rangle_E| \leq \|x\|_E \|y\|_E$$

with equality if an only if x and y are linearly dependent.

Inner products allow us to quantify how different two vectors are in terms of their relative orientation. Two vectors $x, y \in E$ are *orthogonal* with respect a given inner product $\langle\cdot,\cdot\rangle$, if $\langle x, y\rangle = 0$. Note that this notion is not preserved in general if one chooses a different inner product.

Definition 1.2.7 (Orthonormal basis)**.** Let E be an n-dimensional $\mathbb{K}$-vector space equipped with an inner product $\langle\cdot,\cdot\rangle_E$. An orthonormal basis of E is a set of vectors $\{e_1, \ldots, e_n\}$, with $e_i \in E$, which form a basis of E and that are mutually orthogonal, i.e., such that $\langle e_i, e_j\rangle_E = \delta_{i,j}$ for all $1 \leq i, j \leq n$, where $\delta_{i,j}$ denotes the Kronecker delta ($\delta_{i,j} = 1$ if $i = j$, and $\delta_{i,j} = 0$ otherwise).

Given any basis of E one can construct an orthonormal basis using the Gram-Schimdt orthogonalization process. Specifically, given a basis $\{u_1, \ldots, u_n\}$ of E, this consists in defining $e_1 = u_1/\|u_1\|_E$ and then iteratively

$$\tilde{e}_{k+1} = u_{k+1} - \sum_{i=1}^{k} \langle e_k, u_{k+1}\rangle_E e_k\,, \quad e_{k+1} = \frac{\tilde{e}_{k+1}}{\|\tilde{e}_{k+1}\|_E}\,.$$

Consider any vector $x \in E$ and an orthonormal basis $\{e_1, \ldots, e_n\}$ of E. Then since $\mathbf{e}$ is a basis, $x = \sum_{i=1}^n \alpha_i e_i$ where the scalars α_i are uniquely defined. Taking the inner product with e_j from the left we get $\alpha_j = \langle e_j, x\rangle_E$ for al $1 \leq j \leq n$, or equivalently

$$[x]_{\mathbf{e}} = (\langle e_1, x\rangle_E, \ldots, \langle e_n, x\rangle_E) \quad \text{or} \quad x = \sum_{i=1}^{n} \langle e_i, x\rangle_E e_i\,. \tag{1.2.1}$$

Example. The canonical basis $\mathbf{e} = \{e_1, \ldots, e_n\}$ of $\mathbb{K}^n$, defined by $e_1 = (1, 0, \ldots, 0)$, $e_2 = (0, 1, \ldots, 0)$, $\ldots$, $e_n = (0, 0, \ldots, 1)$, is orthonormal with respect to the canonical inner product.

Example (**Discrete Fourier Transform**). The discrete Fourier transform (DFT) can be interpreted as a change of basis on $\mathbb{C}^n$ from the canonical basis $\mathbf{e}$ to new orthormal basis $\mathbf{u} = \{u_1, \ldots, u_n\}$ where for $0 \leq k \leq n-1$,

$$u_{k+1} = \frac{1}{\sqrt{n}}\left(1, \exp\left(i\frac{2\pi k}{n}\right), \ldots, \exp\left(i\frac{2\pi k}{n}(n-1)\right)\right) = \sum_{j=0}^{n-1} \exp\left(i\frac{2\pi k}{n}j\right) e_{j+1}. \tag{1.2.2}$$

The orthonormality can be verified using the explicit expression for the geometric sum. Note also that for all $0 \leq j \leq n-1$ and $1 \leq k \leq n-1$,

$$\exp\left(-i\frac{2\pi k}{n}j\right) = \exp\left(i2\pi j - i\frac{2\pi k}{n}j\right) = \exp\left(i\frac{2\pi(n-k)}{n}j\right) \implies \overline{u_{k+1}} = u_{n-k+1} \tag{1.2.3}$$

where the complex conjugate is intended component-wise. A graphical representation of this basis for $n = 6$ is represented in Figure 1.1.

The DFT of a signal $x \in \mathbb{C}^n$, denoted $\mathrm{DFT}(x) = \hat{x} \in \mathbb{C}^n$, is the vector of coordinates of x with respect to the basis $\mathbf{u}$, i.e.

$$\mathrm{DFT}(x) = \hat{x} \coloneqq [x]_{\mathbf{u}}.$$

By the orthonormality of the basis the coordinates $\hat{x}_k$ can be easily computed:

$$x = \sum_{j=1}^{n} \hat{x}_j u_j \quad \Longrightarrow \quad [\mathrm{DFT}(x)]_{k+1} = \hat{x}_{k+1} = \langle u_{k+1}, x \rangle = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} \exp\left(-i \frac{2\pi k}{n} j\right) x_{j+1}.$$

If x is real, then $\hat{x}_1 = n^{-1/2} \sum_i x_i$ is real and by (1.2.3), for all $1 \leq k \leq n-1$,

$$\overline{\hat{x}_{k+1}} = \langle \overline{u_{k+1}}, \overline{x} \rangle = \langle u_{n-k+1}, x \rangle = \hat{x}_{n-k+1}\,.$$

The inverse DFT maps the vector of coordinates $\hat{x}$ with respect to the basis $\mathbf{u}$ to the vector of coordinates of x with respect to the canonical basis, i.e. $\mathrm{DFT}^{-1}(\hat{x}) = x \in \mathbb{C}^n$. More explicitly, this is computed as follows

$$[\mathrm{DFT}^{-1}(\hat{x})]_{j+1} = x_{j+1} = \langle e_{j+1}, x \rangle = \langle e_{j+1}, \sum_{k=0}^{n-1} \hat{x}_{k+1} u_{k+1} \rangle = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \exp\left(i \frac{2\pi k}{n} j\right) \hat{x}_{k+1}.$$

1.3 Orthogonal subspaces and projections

Let E be a vector space over $\mathbb{K}$ equipped with an inner product $\langle \cdot, \cdot \rangle_E$. Given a subspace $G \subset E$, we denote by $G^\perp$ the subspace of E whose vectors are orthogonal to those in G:

$$G^\perp \coloneqq \{x \in E \,:\, \langle x, y \rangle_E = 0 \quad \forall\, y \in G\}.$$

Note that $G^\perp \cap G = \{0\}$. If E is finite-dimensional, we can pick an orthonormal basis $\{e_1, \ldots, e_k\}$ of G and for any vector of $x \in E$, we define the *orthogonal projection* onto G by

$$\mathrm{Proj}_G(x) \coloneqq \sum_{i=1}^{k} \langle e_i, x \rangle_E e_i$$

and one can verify easily that this definition does not depend on the choice of the basis. In fact, given a different orthonormal basis $\{\tilde{e}_1, \ldots, \tilde{e}_n\}$, we can write $e_i = \sum_j \langle \tilde{e}_j, e_i \rangle_E \tilde{e}_j$, and therefore

$$\sum_{i=1}^{k} \langle e_i, x \rangle_E e_i = \sum_{i=1}^{k} \langle \sum_{j=1}^{k} \langle \tilde{e}_j, e_i \rangle_E \tilde{e}_j, x \rangle_E e_i = \sum_{j=1}^{k} \langle \tilde{e}_j, x \rangle_E \Big(\sum_{i=1}^{k} \overline{\langle \tilde{e}_j, e_i \rangle_E} e_i \Big) = \sum_{j=1}^{k} \langle \tilde{e}_j, x \rangle_E \tilde{e}_j$$

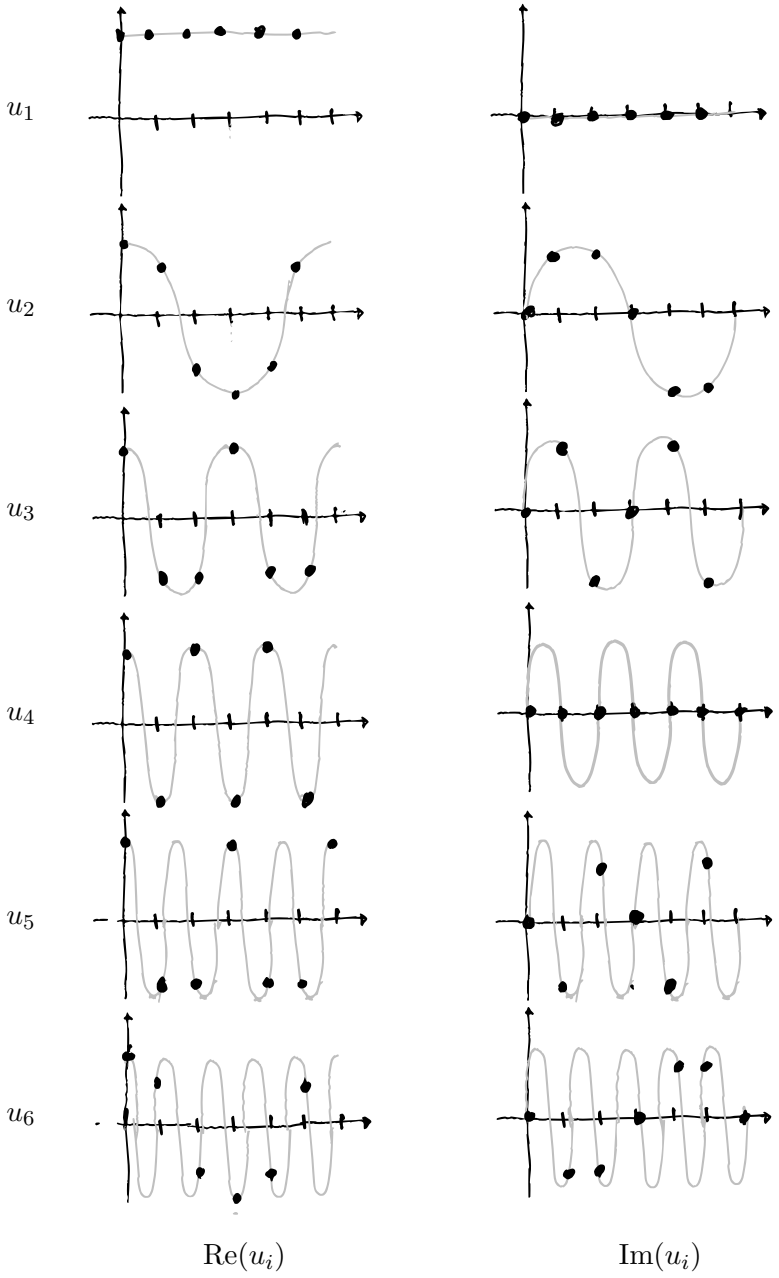

Figure 1.1: Graphical representation of the DFT basis for $n = 6$. The y axis coordinates of the black dots correspond to the components of the vectors u_i (ordered from left to right). Note that $(u_2)_j = \overline{(u_6)_j}$, $(u_3)_j = \overline{(u_5)_j}$ and $(u_4)_j = \overline{(u_4)_j}$ for all $1 \leq j \leq 6$.

Moreover one can show that the orthogonal projection of x onto G minimizes is the vector in G minimizing its distance from x as measured by the norm $\|\cdot\|_E$, i.e.,

$$\mathrm{Proj}_G(x) = \arg\min_{u\in G} \|u - x\|_E \,.$$

Defining $v \coloneqq x - \mathrm{Proj}_G(x)$, then $v \in G^\perp$. This means that any vector $x \in E$ can be decomposed as $x = u + v$ where $u \in G$ and $v \in G^\perp$, and

$$\|x\|_E^2 = \|u\|_E^2 + \|v\|_E^2 \,.$$

The decomposition is unique, since $x = u + v$ implies that $u = \mathrm{Proj}_G(x)$ by computing the coordinates of u with respect to any orthonormal basis. In other words we have $G \oplus G^\perp = E$, and therefore $(G^\perp)^\perp = G$.

1.4 Linear maps, range and kernel

Given two vector spaces E and F over $\mathbb{K}$, a linear map is a function $L : E \to F$ such that

$$L(\alpha x + \beta y) = \alpha L(x) + \beta L(y)$$

for all $\alpha, \beta \in \mathbb{K}$ and $x, y \in E$.

Example. Here are some examples of linear maps:

- let $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ a permutation, the following map is linear

$$L : (\alpha_1, \dots, \alpha_n) \in \mathbb{K}^n \to (\alpha_{\sigma(1)}, \dots, \alpha_{\sigma(n)}) \in \mathbb{K}^n \,;$$

- for any $k \geq 1$, let $C^k(\mathbb{R})$ the space of real-valued functions with continuous k^{th} derivatives, then the following map is linear

$$L : f \in C^k(\mathbb{R}) \to f' \in C^{k-1}(\mathbb{R});$$

- for any $k \geq 0$, the following map is linear

$$L : f \in C^k(\mathbb{R}) \to g \in C^{k+1}(\mathbb{R}) \quad \text{where} \quad g(x) = \int_0^x f(s)\mathrm{d}s \,.$$

The range and kernel of a linear operator from E to F are particular subspaces of E and F, respectively, denoted by $\mathrm{Im}(L)$ and $\mathrm{Ker}(L)$, and defined as follows:

Definition 1.4.1. Let $L \in \mathcal{L}(E, F)$. We define:

$$\mathrm{Ker}(L) \coloneqq \{x \in E \,:\, L(x) = 0\} \subset E \,,$$

$$\mathrm{Im}(L) \coloneqq \{L(x) \in F \,:\, x \in E\} \subset F \,.$$

1.5 Linear maps on finite dimensional spaces

Linear maps on a finite-dimensional inner product space $(E, \langle\cdot,\cdot\rangle_E)$ with values in $\mathbb{K}$ can always be represented as the inner product with a fixed vector in E:

Theorem 1.5.1 (Finite dimensional Riesz representation theorem)**.** *Let E a finite dimensional $\mathbb{K}$-vector space equipped with an inner product $\langle\cdot,\cdot\rangle_E$, and $L \in \mathcal{L}(E,\mathbb{K})$ a linear map. Then, there exists a unique $y_L \in E$ such that*

$$L(x) = \langle y_L, x\rangle_E \quad \forall\, x \in E\,.$$

Proof (Sketch). One needs to find a $y_L \in E$ verifying this expression by computing its coordinates with respect to an orthonormal basis of E, and then prove it is unique. □

The above result allows us to introduce the notion of adjoint:

Theorem 1.5.2 (Adjoint of a linear operator)**.** *Let E and F be two finite dimensional vector spaces over $\mathbb{K}$ equipped with inner products $\langle\cdot,\cdot\rangle_E$ and $\langle\cdot,\cdot\rangle_F$, and let $L \in \mathcal{L}(E,F)$, then there exists a unique map $L^* \in \mathcal{L}(F,E)$ verifying:*

$$\langle L^*(y), x\rangle_E = \langle x, L(y)\rangle_F \quad \forall\, x \in E\,, \forall\, y \in F$$

Such a map is called the adjoint of L.

Proof. For a given $y \in F$, consider the linear operator $T_y \in \mathcal{L}(E,\mathbb{K})$ defined by $T_y(x) = \langle y, L(x)\rangle_F$. By the Riesz represenation theorem there exists a unique vector $L^*(y) \in E$ such that $\langle L^*(y), x\rangle_E = \langle x, L(y)\rangle_F$ for all $x \in E$. This defines the map $L^* : F \to E$, and one can easily verify that this map is linear. □

Note that by definition $(L^*)^* = L$. Furthermore, in the finite dimensional setting range and kernel of L and its adjoint L^* can be related explicitly:

Theorem 1.5.3. *In the setting of Theorem 1.5.2,*

$$\mathrm{Ker}(L) = \mathrm{Im}(L^*)^\perp\,, \quad \mathrm{Im}(L) = \mathrm{Ker}(L^*)^\perp\,.$$

Proof. Since $(L^*)^* = L$, it suffices to prove just one of the two statements. We have

$$\begin{aligned} x \in \mathrm{Im}(L^*)^\perp &\iff \langle x, L^*y\rangle = 0 \quad \forall y \in F \\ &\iff \langle Lx, y\rangle = 0 \quad \forall y \in F \\ &\iff x \in \mathrm{Ker}(L)\,. \end{aligned}$$

□

Theorem 1.5.4 (Rank-nullity theorem)**.** *Let $L \in \mathcal{L}(E,F)$ with $\dim(E) = n$. Then,*

$$\dim(\mathrm{Ker}(L)) + \dim(\mathrm{Im}(L)) = n.$$

Definition 1.5.5 (Self-adjoint operator)**.** A linear operator $L \in \mathcal{L}(E)$ is self-adjoint if and only if $L = L^*$.

Example. The simplest example of self-adjoint operator is the identity operator $I_E : x \in E \to x \in E$.

1.6 Matrix representations of linear maps

Given two vector spaces E and F over $\mathbb{K}$ of dimensions n and m, respectively, two (ordered) basess $\mathbf{e} = \{e_1, \dots, e_n\}$ and $\mathbf{f} = \{f_1, \dots, f_m\}$ of E and F, respectively, and a linear operator $L \in \mathcal{L}(E, F)$, there exist unique scalars $a_{i,j} \in \mathbb{K}$ such that

$$L(e_j) = \sum_{i=1}^{m} a_{i,j} f_i \quad \text{for } 1 \leq j \leq n$$

The *matrix representation* of $L \in \mathcal{L}(E, F)$ with respect to the bases $\mathbf{e}$ and $\mathbf{f}$ is the matrix $[L]_{\mathbf{e},\mathbf{f}} \coloneqq A = (a_{i,j})_{i,j} \in \mathcal{M}_{m,n}(\mathbb{K})$, or more explicitly

$$[L]_{\mathbf{e},\mathbf{f}} \coloneqq \left[\begin{array}{c|c|c|c} & & & \\ [L(e_1)]_{\mathbf{f}} & [L(e_2)]_{\mathbf{f}} & \cdots & [L(e_n)]_{\mathbf{f}} \\ & & & \end{array}\right]$$

where $[L(e_i)]_{\mathbf{f}} \in \mathbb{K}^m$ is the vectors of coordinates of $L(e_i)$ with respect to the basis $\mathbf{f}$.

Then if $u = \sum_{i=1}^{n} \alpha_i e_i \in E$ and $L(u) = \sum_{i=1}^{m} \beta_i f_i \in F$, we have $\beta = A\alpha$, where $\alpha = (\alpha_1, \dots, \alpha_n)^T$ and $\beta = (\beta_1, \dots, \beta_n)^T$, or equivalently:

$$[v]_{\mathbf{f}} = [L]_{\mathbf{e},\mathbf{f}} [u]_{\mathbf{e}} \,.$$

Note that if $\mathbf{f}$ is an orthonormal basis, by equation (1.2.1) we simply have

$$a_{i,j} = \langle f_i, L(e_j) \rangle \,.$$

It is easy to verify that the following properties hold:

1. For all $L^1 \in \mathcal{L}(E, F)$ and $L^2 \in \mathcal{L}(F, G)$, and any bases $\mathbf{e}$, $\mathbf{f}$, $\mathbf{g}$ of E, F, and G, respectively,
$$[L^2 \circ L^1]_{\mathbf{e},\mathbf{g}} = [L^2]_{\mathbf{f},\mathbf{g}} [L^1]_{\mathbf{e},\mathbf{f}} \,.$$
2. If $L \in \mathcal{L}(E, F)$ is invertible then for any bases $\mathbf{e}$, $\mathbf{f}$ of E and F, respectively,
$$[L^{-1}]_{\mathbf{f},\mathbf{e}} = [L]^{-1}_{\mathbf{e},\mathbf{f}} \,.$$
3. For any $L \in \mathcal{L}(E, F)$ and any bases $\mathbf{e}$, $\mathbf{f}$,
$$\dim(\mathrm{Ker}(L)) = \dim(\mathrm{Ker}([L]_{\mathbf{e},\mathbf{f}})) \,, \quad \dim(\mathrm{Im}(L)) = \dim(\mathrm{Im}([L]_{\mathbf{e},\mathbf{f}})) \,.$$

Example. Consider the linear operator $R_\alpha : \mathbb{R}^2 \to \mathbb{R}^2$ describing an anti-clockwise rotation of α rad. Then, denoting by $\mathbf{e}$ the canonical basis,

$$[R_\alpha]_{\mathbf{e},\mathbf{e}} = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \,.$$

Example. The matrix of the identity map $\mathrm{Id} \in \mathcal{L}(E)$ with respect to any basis $\mathbf{e}$ of E is the identity matrix, i.e. $[\mathrm{Id}]_{\mathbf{e},\mathbf{e}} = I = (\delta_{i,j})_{i,j}$, where $\delta_{i,j}$ denotes the Kronecker delta ($\delta_{i,j} = 1$ if $i = j$, and $\delta_{i,j} = 0$ otherwise).

Remark 1.6.1. *To any matrix $A \in \mathcal{M}_{m,n}(\mathbb{K})$ one can associate a linear operator*

$$L_A : x \in \mathbb{K}^n \to Ax \in \mathbb{K}^m \,.$$

Then A is the matrix of L_A with respect to the canonical bases on $\mathbb{K}^n$ and $\mathbb{K}^m$.

1.7 Special matrices and maps

Definition 1.7.1 (Transpose and adjoint of a matrix)**.** Let $A \in \mathcal{M}_{m,n}(\mathbb{K})$, then

- the *transpose* of A is the matrix $A^T \in \mathcal{M}_{n,m}(\mathbb{K})$ of coefficients $a^T_{i,j}$ verifying $a^T_{i,j} = a_{j,i}$ for all $1 \leq i, j \leq n$. A is *symmetric* if and only if $A^T = A$;
- the *adjoint* of A is the matrix $A^* \in \mathcal{M}_{n,m}(\mathbb{K})$ of coefficients $a^*_{i,j}$ verifying $a^*_{i,j} = \overline{a_{j,i}}$ for all $1 \leq i, j \leq n$. A is *hermitian* if and only if $A^* = A$.

Proposition 1.7.2. *For any $A \in \mathcal{M}_{m,n}(\mathbb{K})$, we have*

$$\langle x, Ay \rangle = \langle A^* x, y \rangle \,, \quad \forall\, x \in \mathbb{K}^m,\, y \in \mathbb{K}^n$$

where $\langle \cdot, \cdot \rangle$ is the canonical inner product on $\mathbb{K}^n$ or $\mathbb{K}^m$.

In other words, the adjoint of the linear operator $L_A : x \in \mathbb{K}^n \to Ax \in \mathbb{K}^m$ with respect to the canonical inner product is

$$L^*_A = L_{A^*} \,, \quad \text{where} \quad L_{A^*} : y \in \mathbb{K}^m \to A^* y \in \mathbb{K}^n \,.$$

Remark 1.7.3 (Self-adjoint maps vs hermitian matrices)**.** *Note that if $L \in \mathcal{L}(E)$ is a self-adjoint map, the matrix of L with respect to a given basis $\mathbf{e}$, $[L]_{\mathbf{e},\mathbf{e}}$, is not necessarily hermitian. Vice-versa, if $[L]_{\mathbf{e},\mathbf{e}}$ is hermitian, the map $L \in \mathcal{L}(E)$ may not be self-adjoint. However, if $\mathbf{e}$ is an orthonormal basis, $A = [L]_{\mathbf{e},\mathbf{e}}$ is hermitian if and only if L is self-adjoint, since in this case*

$$a_{i,j} = \langle e_i, L(e_j) \rangle = \langle L(e_i), e_j \rangle = \overline{a_{j,i}} \,.$$

Definition 1.7.4 (Orthogonal/Unitary matrices)**.** $Q \in \mathcal{M}_n(\mathbb{R})$ is orthogonal if and only if $Q^T = Q^{-1}$. $Q \in \mathcal{M}_n(\mathbb{C})$ is unitary if and only if $Q^* = Q^{-1}$.

Example. The matrix representation of the rotation map $[R_\alpha]_{\mathbf{e},\mathbf{e}}$ from Example 1.6 is orthogonal.

It is easy to verify that any unitary/orthogonal matrices Q verify the following important properties:

1. The rows and columns of Q form an orthonormal basis of $\mathbb{K}^n$ with respect to the canonical inner product.
2. The (unitary) transformation $L_Q : x \in \mathbb{K}^n \to Qx \in \mathbb{K}^n$ induced by a unitary matrix Q preserves lengths as measured by the Euclidean norm $\|\cdot\|_2$, i.e.
$$\|L_Q(x)\|_2 = \|Qx\|_2 = \|x\|_2 \,, \quad \forall\, x \in \mathbb{K}^n.$$
3. The (orthogonal) transformation $L_Q : x \in \mathbb{R}^n \to Qx \in \mathbb{R}^n$ induced by an orthogonal matrix Q preserve the angle $\theta_{x,y}$ between any two vectors $x, y \in \mathbb{R}^n$, i.e.
$$\theta_{x,y} \coloneqq \arccos\left(\frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2}\right) = \arccos\left(\frac{\langle Qx, Qy \rangle}{\|Qx\|_2 \|Qy\|_2}\right) .$$
4. The product of orthogonal (resp. unitary) matrices is orthogonal (resp. unitary).

1.8 Eigenvectors and eigenvalues

Definition 1.8.1 (Eigenvectors, eigenvalues and spectrum)**.** Let $A \in \mathcal{M}_n(\mathbb{C})$. A scalar $\lambda \in \mathbb{C}$ is an eigenvalue of A if and only if there exists a vector $v \in \mathbb{C}^n \setminus \{0\}$ such that

$$Av = \lambda v\,.$$

In this case, we say that v is the eigenvector associated with λ. The spectrum of A is the set $\mathrm{Sp}(A)$ of all the eigenvalues of A.

Equivalently, the eigenvalues are the scalars λ such that $\mathrm{Ker}(A - \lambda I) \neq \{0\}$. Moreover, the eigenvectors associated with an eigenvalue λ are precisely the elements of $\mathrm{Ker}(A - \lambda I) \setminus \{0\}$. Therefore we have:

Theorem 1.8.2. *The eigenvalues of $A \in \mathcal{M}_n(\mathbb{C})$ are the roots of the characteristic polynomial*

$$\det(A - \lambda I) = 0\,.$$

Corollary 1.8.3. *A matrix $A \in \mathcal{M}_n(\mathbb{C})$ has at least one and at most n distinct eigenvalues.*

Definition 1.8.4 (Spectral radius)**.** The spectral radius of a matrix $A \in \mathcal{M}_n(\mathbb{C})$ is defined as follows:

$$\rho(A) \coloneqq \max_{\lambda \in \mathrm{Sp(A)}} |\lambda| \in \mathbb{R}^+$$

1.9 Diagonalization and spectral theorem

Some special matrices. Let $A \in \mathcal{M}_n(\mathbb{K})$,

- A is diagonal if $a_{i,j} = 0$ if $i \neq j$;
- A is upper triangular if $a_{i,j} = 0$ if $i > j$;
- A is lower triangular if $a_{i,j} = 0$ if $i < j$;

Definition 1.9.1 (Diagonalizable matrix)**.** Let $A \in \mathcal{M}_n(\mathbb{C})$. A is diagonalizable if and only if there exists an invertible matrix $S \in \mathcal{M}_n(\mathbb{C})$ and a diagonal matrix $D \in \mathcal{M}_n(\mathbb{C})$ such that $A = SDS^{-1}$.

Theorem 1.9.2 (Diagonalization)**.** *Let $A \in \mathcal{M}_n(\mathbb{C})$. A is diagonalizable if and only if it has n linearly independent eigenvectors $\{v_1, \ldots, v_n\}$. Then*

$$A = SDS^{-1}$$

where D is a diagonal matrix such that, for all $1 \leq i \leq n$, D_{ii} is the eigenvalue associated to the eigenvector v_i and the ith column of S is v_i.

Proof. It suffices to observe that $\{v_1, \ldots, v_n\}$ is a set of linearly independent eigenvectors if and only if the matrix S defined in the statement is invertible, and there exists a diagonal matrix D such that $AS = SD$. □

Remark 1.9.3. *Invertible ≠ diagonalizable. Consider for example the matrix:*

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

This is invertible and it has one eigenvalues $\lambda = 1$ which is a double root of the characteristic polynomial. However the eignevectors span a one dimensional space $\mathrm{Ker}(A - I) = \mathrm{span}\{(0, 1)\}$ so A cannot be diagonalizable. On the other hand, the null matrix is diagonal but not invertible.

Corollary 1.9.4. *If all eigenvalues of $A \in \mathcal{M}_n(\mathbb{C})$ are distinct then A is diagonalizable.*

Proof. We need to prove that given n eigenvectors $v_1, \ldots, v_n$ associated with the n distinct eigenvectors $\lambda_1, \ldots, \lambda_n$ then these are linearly independent. If this was not the case, we can suppose that $v_1, \ldots, v_m$ are linearly independent with $1 \leq m < n$ and that there exists constants $\alpha_1, \ldots, \alpha_m$ not all zero such that

$$v_{m+1} = \sum_{i=1}^{m} \alpha_i v_i \,.$$

Then

$$Av_{m+1} = A \sum_{i=1}^{m} \alpha_i v_i = \sum_{i=1}^{m} \alpha_i \lambda_i v_i$$

and

$$Av_{m+1} = \lambda_{m+1} v_{m+1} = \sum_{i=1}^{m} \alpha_i \lambda_{m+1} v_i$$

Subtracting the two equations we get

$$\sum_{i=1}^{m} \alpha_i (\lambda_{m+1} - \lambda_i) v_i = 0$$

with $\lambda_{m+1} - \lambda_i \neq 0$ for all $1 \leq i \leq m$, which is a contradiction since $\{v_1, \ldots, v_m\}$ was supposed linearly independent. □

Theorem 1.9.5 (Schur)**.** *Let $A \in \mathcal{M}_n(\mathbb{C})$, then there exists a unitary matrix $Q \in \mathcal{M}_n(\mathbb{C})$ and an upper triangular matrix $T \in \mathcal{M}_n(\mathbb{C})$ such that $A = QTQ^{-1}$.*

Theorem 1.9.6 (Spectral theorem)**.** *Let $A \in \mathcal{M}_n(\mathbb{C})$ be hermitian. Then,*

- *all eigenvalues of A are real;*
- *A has n mutually orthogonal eigenvectors;*

- *there exists a unitary matrix $Q \in \mathcal{M}_n(\mathbb{C})$ such that $A = QDQ^*$ where D is diagonal and real, and Q is real if A is real.*

Example (**Convolution and DFT**). Given two vectors $h = (h_i)_i, x = (x_i)_i \in \mathbb{K}^n$, the cyclic (or periodic) convolution of h and x is the vector $h * x \in \mathbb{K}^n$ whose $(i+1)$th coordinate $(h * x)_{i+1}$ (with respect to the canonical basis) is

$$(h * x)_{i+1} \coloneqq \sum_{j=0}^{n-1} x_{j+1} h_{(i-j)_n+1} \tag{1.9.1}$$

where $(k)_n \coloneqq k \bmod n$ for all $k \in \mathbb{N}$. Setting $k = (i-j)_n$ in equation (1.9.1), we find $j = (i-k)_n$ and therefore $h * x = x * h$.

The map $\mathrm{conv}_h : x \to h * x$ is linear. Moreover, suppose that for all $0 \leq i \leq n-1$

$$h_{(i)_n+1} = \overline{h_{(-i)_n+1}}, \tag{1.9.2}$$

then conv_h is self-adjoint since for all $x, y \in \mathbb{K}^n$

$$\langle h * x, y \rangle = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \overline{x_{j+1} h_{(i-j)_n+1}}\, y_{i+1} = \sum_{j=0}^{n} \overline{x_{j+1}} \left(\sum_{i=0}^{n-1} y_{i+1} h_{(j-i)_n+1} \right) = \langle x, h * y \rangle .$$

The same can be seen from the matrix representation of conv_h with respect to the canonical basis $\mathbf{e}$, which is the *circulant* matrix C_h given by

$$C_h = [\mathrm{conv}_h]_{\mathbf{e},\mathbf{e}} = \left[\begin{array}{c|c|c|c} & & & \\ [C^h(e_1)]_{\mathbf{e}} & [C^h(e_2)]_{\mathbf{e}} & \cdots & [C^h(e_n)]_{\mathbf{e}} \\ & & & \end{array}\right] = \begin{bmatrix} h_1 & h_n & \cdots & h_2 \\ h_2 & h_1 & \cdots & h_3 \\ \vdots & \vdots & \ddots & \vdots \\ h_n & h_{n-1} & \cdots & h_1 \end{bmatrix}.$$

One has that

$$h * x = \mathrm{DFT}^{-1}(\sqrt{n}\, \mathrm{DFT}(h) \odot \mathrm{DFT}(x)) \tag{1.9.3}$$

where $\odot$ denotes the componentwise multiplication. In fact,

$$\begin{aligned} [\mathrm{DFT}(h * x)]_{k+1} &= \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} \sum_{l=0}^{n-1} x_{l+1} h_{(j-l)_n+1} \exp\left(-i \frac{2\pi k}{n} j\right) \\ &= \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} \sum_{l=0}^{n-1} x_{l+1} \exp\left(-i \frac{2\pi k}{n} l\right) h_{(j-l)_n+1} \exp\left(-i \frac{2\pi k}{n} (j-l)\right) \end{aligned}$$

Renaming $q = (j-l)_n$ we get

$$\begin{aligned} [\mathrm{DFT}(h * x)]_{k+1} &= \frac{1}{\sqrt{n}} \sum_{q=0}^{n-1} \sum_{l=0}^{n-1} x_{l+1} \exp\left(-i \frac{2\pi k}{n} l\right) h_{q+1} \exp\left(-i \frac{2\pi k}{n} q\right) \\ &= \sqrt{n}\, [\mathrm{DFT}(h)]_{k+1} [\mathrm{DFT}(x)]_{k+1} \end{aligned}$$

Denoting by $U \in \mathcal{M}_n(\mathbb{K})$ the unitary matrix whose columns are the vectors $u_1, \dots, u_n$ defined in (1.2.2), equation (1.9.3) can be equivalently written as a diagonalization of the matrix C_h:

$$C_h = U \mathrm{diag}(\sqrt{n}\,\hat{h}) U^* \tag{1.9.4}$$

where $\hat{h} = \mathrm{DFT}(h)$.

Note that if h verifies (1.9.2), then C_h is hermitian and one can verify that $\hat{h} \in \mathbb{R}^n$ by definition of the DFT. On the other hand, if h does not verify (1.9.2) C_h is not unitary but the decomposition (1.9.4) still holds.

2 Singular value decomposition and applications

2.1 Singular value decomposition

The singular value decomposition allows us to decompose a matrix in a sum of simpler rank-one matrix. This is very useful in different contexts, since it allows us to have a more compact and easier to store representation of data, linear operators, images, etc, but also for the solution of inverse problems. Consider the following example (from Strang's book):

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

which may represent a black image of 3×4 pixels. We can write A in the form $A = uv^T$ where u and v are the column vectors $u = [1, 1, 1]^T$ and $v = [1, 1, 1, 1]^T$. Of course, the same type of decomposition holds if A were an $m \times n$ matrix: in this case it implies that in order to store A, we only need to store $m + n$ values instead of mn values.

Before introducing the definition of singular value, we state two useful results on the eigenvalues of matrix products.

Lemma 2.1.1. *Let $A \in \mathcal{M}_{m,n}(\mathbb{K})$, then A^*A is hermitian and it has non-negative eigenvalues.*

Lemma 2.1.2. *Let $A \in \mathcal{M}_{m,n}(\mathbb{K})$ and $B \in \mathcal{M}_{n,m}(\mathbb{K})$ then AB and BA have the same non-zero eigenvalues.*

Definition 2.1.3 (Singular values)**.** The singular values of a matrix $A \in \mathcal{M}_{m,n}(\mathbb{K})$ are the square roots of the eigenvalues of A^*A.

Note that due to lemma 2.1.2, the non-zero singular values of A are also the square roots of the eigenvalues of AA^*.

Theorem 2.1.4 (SVD)**.** *Every $A \in \mathcal{M}_{m,n}(\mathbb{K})$ can be written as*

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

where $V \in \mathcal{M}_n(\mathbb{K})$ and $U \in \mathcal{M}_m(\mathbb{K})$ are unitary and $\Sigma \in \mathcal{M}_{m,n}(\mathbb{K})$ is diagonal and of the form

$$\Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}$$

where $\Sigma_r \in \mathcal{M}_r(\mathbb{K})$ is diagonal and $\Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ are the r non-zero singular values of A. The columns of U and V are, respectively, the left and right singular vectors of A.

Remark 2.1.5. *If we denote by u_i the ith column of U (interpreted as a $m \times 1$ matrix) and by v_i^* the ith row of V (interpreted as a $1 \times n$ matrix), the SVD can be equivalently written as follows:*

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^* ,$$

i.e. as a decomposition in a sum of r rank-one matrices. Moreover, if we denote by U_r and V_r the matrices containing the first r columns of U and V we also have

$$A = U_r \Sigma_r V_r^* .$$

*This last decomposition is sometimes refferred to as **economy size SVD**: this is because in applications we often do not the remaining columns of U and V.*

Remark 2.1.6 (Orthonormal bases of the fundamental spaces)**.** *Notice that the SVD gives orthonormal bases for the kernel and range of $A = U\Sigma V^*$ as well as to their orthogonal complements. In particular, one can check that:*

- $\mathrm{rank}(A)$ *is number of nonzero singular values of A;*
- $\mathrm{Ker}(A)$ *is the span of the last $n - r$ columns of V;*
- $\mathrm{Ker}(A)^\perp = \mathrm{Im}(A^*)$ *is the span of the first r columns of V;*
- $\mathrm{Im}(A)$ *is the space of the first r columns of U.*
- $\mathrm{Im}(A)^\perp = \mathrm{Ker}(A^*)$ *is the span of the last $m - r$ columns of U;*

Remark 2.1.7 (Relation with diagonalization of hermitian matrices)**.** *The factorisation of an hermitian matrix given by the spectral theorem 1.9.6 may not coincide with any SVD decomposition of the same matrix. This is because the singular values are non-negative real scalars whereas the eigenvalues of an hermitian matrix may be negative. A positive hermitian matrix has non-negative eigenvalues, therefore (up to a rearrangement of the eigenvalues) the two factorizations are equivalent.*

Remark 2.1.8 (Geometric interpretation)**.** *According to the SVD theorem, any linear map from $\mathbb{K}^n$ to $\mathbb{K}^m$ may be written as the composition of a unitary transformation (such as rotations or reflections) in $\mathbb{K}^n$, followed by a diagonal transformation (stretching/scaling of the axes) from $\mathbb{K}^n$ to $\mathbb{K}^m$ and a final unitary transformation in $\mathbb{K}^m$.*

2.2 Inverse problems and regularization

The content of this section is based on the course notes of Gabriel Peyré [2].

An inverse problem consists in finding a high resolution signal from low resolution noisy observations. These problems are usually ill-posed, e.g., they may not have a unique solution or the map from observation to signal may not be continuos. Typically,

the observation can be represented as a vector $y \in F$, which is the result of the application of a linear map $L \in \mathcal{L}(E, F)$ onto the signal $x \in E$, plus an acquisition noise $w \in F$, i.e.,

$$y = Lx + w\,.$$

The objective is to recover x knowing y and the linear map L but not the noise w.

It is often convenient to think of E and F as infinite dimensional function spaces, but in most applications E and F are finite dimensional and this is the only setting we will consider here. Note that even if L is invertible, when $w \neq 0$ it may not be appropriate to take $\tilde{x} = L^{-1}y$ as an approximation for y. This is beacause $\tilde{x} = x + L^{-1}w$ but L^{-1} is usually ill-conditioned, meaning that $\|L^{-1}w\|_F \gg \|w\|_F$. We will discuss this phenomenon in detail in the next chapter.

2.2.1 Moore-Penrose inverse

If $w = 0$, one can try to inverse the linear relation $y = Lx$ to find x. If L is not invertible this can still be done in a least square sense. From now on, we consider for simplicity the case where $E = \mathbb{K}^n$ and $F = \mathbb{K}^m$ equipped with the canonical inner products and norms, and we identify $L \in \mathcal{M}_{m,n}(\mathbb{K})$ (in the general case, one just needs to choose two bases $\mathbf{e}$ and $\mathbf{f}$ for E and F, respectively, and replace L with its matrix representation $[L]_{\mathbf{e},\mathbf{f}}$).

First of all, since we may have $y \notin \mathrm{Im}(L)$, we take

$$\tilde{y} \coloneqq \mathrm{Proj}_{\mathrm{Im}(L)} y = \arg\min_{z \in \mathrm{Im}(L)} \|z - y\|^2$$

Suppose that $\{u_1, \dots, u_r\}$ is an orthonormal basis for $\mathrm{Im}(L)$ (note that r is the rank of L), then any z can be written as $z = \sum_{i=1}^r z_i u_i$ and

$$\Big\| \sum_{i=1}^r z_i u_i - y \Big\|^2 = \sum_{i=1}^r \|z_i u_i - y\|^2 - (k-1)\|y\|^2$$

from which one can deduce that $z_i = \langle u_i, y \rangle$, or $\tilde{y} = \sum_{i=1}^r \langle u_i, y \rangle u_i$. Denoting by $U_r \in \mathbb{M}_{m,k}$ the matrix whose ith column is u_i, we also have

$$\tilde{y} = U_r U_r^* y\,.$$

Solving $Lx = \tilde{y}$ in least square sense means solving

$$\inf_{Lx = \tilde{y}} \|x\|^2 \tag{2.2.1}$$

The solution to this problem involves the Moore-Penrose inverse of L which is defined as follows:

Definition 2.2.1 (Moore-Penrose inverse)**.** Let $L \in \mathcal{M}_{m,n}(\mathbb{K})$ and let $L = U\Sigma V^*$ an SVD of L, where $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$. The Moore-Penrose inverse of L is the matrix $L^\dagger \coloneqq V\Sigma^\dagger U^*$, where

$$\Sigma^\dagger \coloneqq \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} \in \mathcal{M}_{n,m}\,, \quad \Sigma_r^{-1} = \mathrm{diag}\left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r}\right)\,.$$

The Moore-Penrose inverse is a generalized inverse of L. If $m = n$ and L is invertible, one clearly has $L^{-1} = L^\dagger$.

Proposition 2.2.2. *The unique solution of problem* (2.2.1) *is* $\tilde{x} = L^\dagger y \in \mathrm{Ker}(L)^\perp$.

Proof. First of all, observe that for any $x \in \mathbb{K}^n$ can be decomposed as $x = z + r$ where $r \in \mathrm{Ker}(L)$ and $z \in \mathrm{Ker}(L)^\perp$, and $\|x\|^2 = \|z\|^2 + \|r\|^2$. The minimisation problem is equivalent to

$$\inf_{Lz=\tilde{y}} \|z\|^2 + \|r\|^2$$

and we must have $r = 0$. We need to find a solution to $Lz = \tilde{y}$, with $z \in \mathrm{Ker}(L)^\perp$. Using the SVD of L, we need to solve

$$U_r \Sigma_r V_r^* z = U_r U_r^* y \,.$$

Since $U_r^* U_r$ is the identity matrix, this implies that $V_r^* z = \Sigma_r^{-1} U_r^* y$. Now $z \in \mathrm{Ker}(L)^\perp$ if and only if there exists a vector $\alpha \in \mathbb{K}^r$ such that $z = \sum_{i=1}^r \alpha_i v_i = V_r \alpha$. Since $V_r^* V_r$ is the identity matrix, we have

$$\alpha = \Sigma_r^{-1} U_r^* y \implies z = V_r \Sigma_r^{-1} U_r^* y \,.$$

□

2.2.2 Ridge regression

If the noise $w \neq 0$, then it is not advisable to use the Moore-Penrose to recover the signal x. For example, using the notation of the previous section, if $w = \alpha u_r$ for some $\alpha \in \mathbb{R}$ then

$$\|L^\dagger y - L^\dagger L x\| = \|L^\dagger w\| = \frac{|\alpha|}{\sigma_r}$$

which may be very large if σ_r (the smallest non-zero singular value) is small. In this case $L^\dagger y$ may be very far from the solution we would obtain without noise, even if this latter is small (i.e., even if $|\alpha|$ is small).

Instead of solving the problem with the Moore-Penrose inverse, let us consider the following variational problem

$$\inf_x \|Lx - y\|^2 + \lambda \|x\|^2 \,, \tag{2.2.2}$$

where $\lambda > 0$ is a parameter multiplying a quadratic regularization term. We can find the solution of such problem following similar steps as those in the proof of proposition 2.2.2. Specifically, if we decompose $x = z + r$ with $z \in \mathrm{Ker}(L)^\perp$ and $r \in \mathrm{Ker}(L)$, we immediately see that we must have $r = 0$ and therefore we need to solve

$$\inf_{x \in \mathrm{Ker}(L)^\perp} \|Lx - \tilde{y}\|^2 + \|y'\|^2 + \lambda \|x\|^2$$

where we have also used the decomposition $y = \tilde{y} + y'$ where $\tilde{y} \in \mathrm{Im}(L)$ and $y' \in \mathrm{Im}(L)^\perp$. Setting $x = V_r\alpha$ and neglecting the term $\|y'\|^2$ which does not depend on x, we get

$$\inf_{\alpha \in \mathbb{K}^r} \|LV_r\alpha - U_rU_r^*y\|^2 + \lambda\|V_r\alpha\|^2\,, \tag{2.2.3}$$

Using the SVD of $L = U_r\Sigma_rV_r^*$, and the fact that $V_r^*V_r$ and $U_r^*U_r$ are the identity matrix, each coordinate α_i of the vector α needs to minimize

$$(\sigma_i\alpha_i - (U_r^*y)_i)^2 + \lambda\alpha_i^2$$

Hence, we obtain that the unique solution of problem (2.2.3) is

$$\alpha = \mathrm{diag}\left(\frac{\sigma_1}{\sigma_1^2+\lambda}, \ldots, \frac{\sigma_r}{\sigma_r^2+\lambda}\right) U_r^*y\,.$$

We have proven the following result:

Proposition 2.2.3. *Problem (2.2.2) admits a unique solution given by $\tilde{x} = L_\lambda^\dagger y$, where*

$$L_\lambda^\dagger \coloneqq V_r\mathrm{diag}\left(\frac{\sigma_1}{\sigma_1^2+\lambda}, \ldots, \frac{\sigma_r}{\sigma_r^2+\lambda}\right) U_r^*\,.$$

Remark 2.2.4. *A more general approach, which bypasses the definition of an auxiliary variational problem, consists in replacing $L^\dagger = V_r\Sigma_r^{-1}U_r^*$ with*

$$L_\lambda^\dagger \coloneqq V_r\mathrm{diag}(\mu_\lambda(\sigma_1), \ldots, \mu_\lambda(\sigma_r))U_r^*\,,$$

where $\mu_\lambda : [0,\infty) \to [0,\infty)$ is a function depending on a regularization parameter $\lambda > 0$, such that

$$\mu_\lambda(\sigma) \leq C_\lambda\,, \quad \lim_{\lambda\to 0} \mu_\lambda(\sigma) = \frac{1}{\sigma}\,,$$

where $C_\lambda > 0$ is a constant depending on λ. The quadratic regularization corresponds to the choice:

$$\mu_\lambda(\sigma) = \frac{\sigma}{\sigma^2+\lambda}$$

Remark 2.2.5. *Note that the solution $\tilde{x}$ obtained by the Moore-Penrose inverse or the quadratic regularization belongs to $\mathrm{Ker}(L)^\perp$. This means that we have no hope of reconstructing a signal that is not orthogonal to $\mathrm{Ker}(L)$. This fact can be alleviated by means of more general regularization terms which may used to include additional information on the solution such as sparsity.*

Example (Deconvolution via the DFT). Consider the linear operator $\mathrm{conv}_h \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ performing the convolution with a vector h veryfying (1.9.2), so that its DFT, denoted $\hat{h}$, is a real vector. The deconvolution of the observed signal y via the Moore-Penrose inver quadratic regularization amounts to setting:

$$x = \mathrm{DFT}^{-1}(\hat{x})\,, \quad \hat{x}_i = \begin{cases} \dfrac{\hat{y}_i}{\hat{h}_i} & \text{if } \hat{h}_i \neq 0 \\ 0 & \text{otherwise} \end{cases}\,,$$

whereas the deconvolution obtained via the quadratic regularization is given by

$$x = \mathrm{DFT}^{-1}(\hat{x}), \quad \hat{x}_i = \frac{\hat{h}_i}{\hat{h}_i^2 + \lambda}\hat{y}_i .$$

Remark 2.2.6 (**Tikhonov regularization**). *A more general form of variational regularization is the following*

$$\inf_x \|Lx - y\|^2 + \lambda \|Bx\|^2 ,$$

where $B \in \mathcal{M}_n(\mathbb{K})$. The matrix B can be chosen to enforce desired properties on the solution x. For example, if x is a time varying signal, a typical choice for B is a discrete version of the time derivative, which enforces smoothness on the solution.

2.3 Principal component analysis

Given a set of data valued in a m-dimensional vector space E, the idea of the principal component analysis it to find the lower dimensional subspaces of E such that the orthogonal projection of the data vectors onto such subspaces has the largest possible variance.

Suppose that we are given n realization of an m-dimensional random vector, which we denote $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{K}^m$. For example these could be obtained by repeating n times the same experiment. The sample mean and variance are given by

$$\mathrm{Avg}(\mathcal{X}) = \frac{1}{n}\sum_{i=1}^n x_i , \quad \mathrm{Var}(\mathcal{X}) = \frac{1}{n-1}\sum_{i=1}^n \|x_i - \mathrm{Avg}(\mathcal{X})\|_2^2 .$$

Let us now introduce the matrix $X \in \mathcal{M}_{m,n}(\mathbb{K})$ whose columns are the n observed realizations of the random vector, normalized to have zero sample mean, i.e.

$$X = \left[\begin{array}{c|c|c|c} & & & \\ x_1 - \mathrm{Avg}(\mathcal{X}) & x_2 - \mathrm{Avg}(\mathcal{X}) & \cdots & x_n - \mathrm{Avg}(\mathcal{X}) \\ & & & \end{array}\right] \tag{2.3.1}$$

The sample covariance matrix is given by

$$S = \frac{1}{n-1} X X^* \in \mathcal{M}_{m,m}(\mathbb{K})$$

and note that $\mathrm{Var}(\mathcal{X}) = \mathrm{tr}(S)$, the trace of the matrix S.

Let us now consider the projection of the data on the subspace spanned by a unit vector $p_1 \in \mathbb{K}^m$, $\tilde{\mathcal{X}} = \{\tilde{x}_1, \dots, \tilde{x}_n\}$ where $\tilde{x}_i = \langle p_1, x_i \rangle p_1$ and with $\langle \cdot, \cdot \rangle$ being the canonical inner product. The sample mean of the new set is simply $\mathrm{Avg}(\tilde{\mathcal{X}}) = \langle p_1, \mathrm{Avg}(\mathcal{X}) \rangle p_1$, whereas the sample variance is

$$\begin{aligned} \mathrm{Var}(\tilde{\mathcal{X}}) &= \frac{1}{n-1}\sum_{i=1}^n \|\tilde{x}_i - \mathrm{Avg}(\tilde{\mathcal{X}})\|_2^2 = \frac{1}{n-1}\sum_{i=1}^n |\langle p_1, x_i - \mathrm{Avg}(\mathcal{X}) \rangle|^2 \\ &= \frac{1}{n-1}\|X^* p_1\|_2^2 = \langle S p_1, p_1 \rangle . \end{aligned}$$

The first principal component is the direction p_1 such that the variance of the projected data is maximal. This is therefore the solution to the maximization problem

$$\sup \left\{ \frac{1}{n-1} \|X^* p_1\|_2^2 \, : \, \|p_1\|_2 = 1 \right\} . \tag{2.3.2}$$

Given an SVD decomposition of $X = U\Sigma V^*$, we can express any vector p_1 in terms of the orthonormal basis $\{u_1, \dots, u_m\}$ given by the columns of U, i.e. $p_1 = \sum_{j=1}^m p_{1,j} u_j$ and we have that

$$\|p_1\|_2^2 = \sum_{j=1}^n p_{1,j}^2 = 1$$

Threfore

$$\|X^* p_1\|_2^2 = \left\| \sum_{i=1}^r v_i \sigma_i \langle u_i, p_1 \rangle \right\|_2^2 = \sum_{i=1}^r \sigma_i^2 p_{1,i}^2 \leq \sigma_1^2 \, .$$

If $\sigma_1 > \sigma_2$, the inequality is an equality if and only if $p_1 = u_1$, which is therefore the unique first principal component. On the other hand, if the first $k > 1$ singular values coincide than any unit vector spanned by the first k right singular vectors $\{u_1, \dots, u_k\}$ solves problem (2.3.2).

The following principal components can be defined in an iterative fashion. The kth principal component is the direction p_k, orthogonal to first $k-1$ principal components $p_1, \dots p_{k-1}$, such that the data projected on the space spanned by p_k has maximal variance. One can verify that the space spanned by the first k principal components is the k-dimensional subspace of $\mathbb{K}^m$ such that the projection of the data on such space has maximal variance. As a matter of fact, the variance of $\tilde{\mathcal{X}}$, the projection of $\mathcal{X}$ onto the space spanned by the orthonormal set $\{p_1, \dots, p_k\}$ is

$$\mathrm{Var}(\tilde{\mathcal{X}}) = \frac{1}{n-1} \sum_{i=1}^k \|X^* p_i\|_2^2 \, .$$

3 Matrix norms, low rank approximations and condition number

3.1 Matrix norms

The space of matrices $\mathcal{M}_{m,n}(\mathbb{K})$ is isomorphic to $\mathbb{K}^{mn}$, since a matrix in $\mathcal{M}_{m,n}(\mathbb{K})$ can be simply regarded as an ordered collection of mn values in $\mathbb{K}$ and such an identification is linear. Therefore, we could define norms on $\mathcal{M}_{m,n}(\mathbb{K})$ simply using the norms we know on $\mathbb{K}^{mn}$. However, this approach gives us norms that in general are not well-suited for computations, since they do not behave well under the operations of matrix-vector and matrix-matrix multiplication. The two main properties that are specific for norms on $\mathcal{M}_{m,n}(\mathbb{K})$ are given in the following definitions.

Definition 3.1.1 (Consistent/compatible norm)**.** Let $\|\cdot\|$ be a norm on $\mathbb{K}^n$, $n \geq 1$. A norm $\|\cdot\|$ on $\mathcal{M}_{m,n}(\mathbb{K})$ is consistent (or compatible) with respect to $\|\cdot\|$ if

$$\|Au\| \leq \|A\|\|u\|\,, \quad \forall u \in \mathbb{K}^n\,.$$

Definition 3.1.2 (Sub-multiplicative/matrix norm)**.** A norm $\|\cdot\|$ on $\mathcal{M}_{m,n}(\mathbb{K})$ is sub-multiplicative if for all $m, n, p \geq 1$,

$$\|AB\| \leq \|A\|\|B\|\,, \quad \forall\, A \in \mathcal{M}_{m,p}(\mathbb{K})\,, B \in \mathcal{M}_{p,n}(\mathbb{K})\,. \tag{3.1.1}$$

In this case, we say that $\|\cdot\|$ is a *matrix norm*.

Note that equation (3.1.1) in the definition of a sub-multiplicative norm is to be regarded as a relation between three different norms, one defined on $\mathcal{M}_{m,p}(\mathbb{K})$, one on $\mathcal{M}_{p,n}(\mathbb{K})$ and another one on $\mathcal{M}_{m,n}(\mathbb{K})$.

Remark 3.1.3. *Two important points:*

1. *Not all norms are sub-multiplicative: take for example,*

$$\|A\|_\Delta := \max_{i,j} |a_{i,j}|,$$

and consider the case

$$A = B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

2. *All sub-multiplicative norms are compatible with respect to some vector norm. To see this, it suffices taking $n = 1$ in equation (3.1.1) and regarding B and AB as vectors in $\mathbb{K}^p$ and $\mathbb{K}^m$, respectively.*

One of the most used norms in applications is the Frobenius norm, which can be obtained interpreting a matrix $\mathcal{M}_{m,n}(\mathbb{K})$ as a vector in $\mathbb{K}^{mn}$ and applying the vector 2-norm.

Definition 3.1.4 (Frobenius norm)**.** The Frobenius norm of $A \in \mathcal{M}_{m,n}(\mathbb{K})$ is defined as follows:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{i,j}|^2}$$

As in the vector case, the Frobenius norm is the norm associated with an inner product (which coincides with the canonical inner product on $\mathbb{K}^{mn}$) and that can be written as follows:

$$\langle \cdot, \cdot \rangle : (A, B) \in \mathcal{M}_{m,n}(\mathbb{K}) \times \mathcal{M}_{m,n}(\mathbb{K}) \to \mathrm{tr}(A^* B) \in \mathbb{K}\,, \tag{3.1.2}$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator.

Remark 3.1.5. *As a consequence of Cauchy-Schwarz inequality, the Frobenius norm is sub-multiplicative, and moreover it is compatible with respect to the 2-norm $\|\cdot\|_2$ on $\mathbb{K}^n$.*

We can define matrix norms in a different way starting from a norm on $\mathbb{K}^n$, as shown in the following definition.

Definition 3.1.6 (Induced norm)**.** Let $\|\cdot\|$ be a norm on $\mathbb{K}^n$. The norm induced norm on $\mathcal{M}_{m,n}(\mathbb{K})$ associated with it is given by:

$$\|A\| \coloneqq \sup\{\|Ax\| \ , \ x \in \mathbb{K}^n\,, \|x\| = 1\}$$

It is easy to check that this definition can be expressed in other equivalent ways. In particular, one can prove that the sup is attained and moreover

$$\|A\| = \max\{\|Ax\| \ , \ x \in \mathbb{K}^n\,, \|x\| = 1\} = \max\{\|Ax\| \ , \ x \in \mathbb{K}^n\,, \|x\| \leq 1\}\,,$$

or also

$$\|A\| = \max\left\{ \frac{\|Ax\|}{\|x\|} \ , \ x \neq 0 \right\}.$$

Proposition 3.1.7 (Properties of induced norms)**.** *Let $\|\cdot\|$ be a norm on $\mathcal{M}_{m,n}(\mathbb{K})$ induced by a norm $\|\cdot\|$ on $\mathbb{K}^n$:*

- *$\|\cdot\|$ is consistent with respect to the associated vector norm;*
- *$\|\cdot\|$ is sub-multiplicative;*
- *$\|I\| = 1$.*

Note that the Frobenius norm cannot be induced by any norm, since $\|I\|_F = \sqrt{n}$ if n is the identity matrix on $\mathcal{M}_n(\mathbb{K})$.

Definition 3.1.8 (Norms induced by the p-norms)**.** The norms on $\mathcal{M}_{m,n}(\mathbb{K})$ induced by the p-norms are defined as follows

$$\|A\|_p \coloneqq \max\left\{\frac{\|Ax\|_p}{\|x\|_p}\ ,\ x \neq 0\right\}.$$

Remark 3.1.9. *Using the SVD of A (see Section 3.2.10) we find*

$$\|A\|_2 = \sigma_1(A)$$

where $\sigma_1(A)$ is the largest singular value of A. The norm $\|\cdot\|_2$ is also called the spectral norm.

Remark 3.1.10 (Spectral radius)**.** *We observe that the map $A \mapsto \rho(A)$ defined on square matrices is 1-homogeneous:*

$$\rho(\alpha A) = |\alpha|\rho(A)\,, \quad \forall \alpha \in \mathbb{K}\,.$$

However it is not a norm. In fact:

- *if T is any triangular matrix with zero diagonal $\rho(T) = 0$, and*
- *the triangular inequality does not hold, e.g.:*

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad 0 = \rho(A) + \rho(A^T) < \rho(A + A^T) = 1\,.$$

3.2 Unitarily invariant norms and best low rank approximations

Definition 3.2.1 (Unitarily invariant norms)**.** A norm $\|\cdot\|$ on $\mathcal{M}_{m,n}(\mathbb{K})$ is unitarily invariant if and only if for any unitary matrix $Q \in \mathcal{M}_m(\mathbb{K})$ and $Z \in \mathcal{M}_n(\mathbb{Z})$ we have

$$\|QAZ\| = \|A\|\,, \quad \forall\, A \in \mathcal{M}_{m,n}(\mathbb{Z}).$$

The Frobenius norm $\|\cdot\|_F$ and the induced norm $\|\cdot\|_2$ are both unitarily invariant. This follows from the inner product definition (3.1.2) for the Frobenius norm and from remark (3.1.9) for the $\|\cdot\|_2$ norm. Other classical examples of unitarily invariant norm are the p-Schatten norms

Definition 3.2.2 (Schatten p-norms)**.** Let $\sigma(A) = (\sigma_1, \ldots, \sigma_n)$, the vector of singualr values of $A \in \mathcal{M}_{m,n}(\mathbb{K})$. Then, the Schatten p-norm of A is defined by

$$\|A\|_{(p)} \coloneqq \|\sigma(A)\|_p\,.$$

The Frobenius and spectral norm are particular cases of Schatten norms. We have

$$\|A\|_F = \|A\|_{(2)} = \sqrt{\sigma_1(A)^2 + \ldots + \sigma_n(A)^2}\,, \quad \|A\|_2 = \|A\|_{(\infty)} = \sigma_1(A)$$

The first equality for the Frobenius norm can be seen as a particular (although stronger) case of the following result.

Theorem 3.2.3 (Von Neumann's trace inequality)**.** *Let $A, B \in \mathcal{M}_{m,n}(\mathbb{K})$, and denote by $\sigma(A) = (\sigma_i(A))_i$ and $\sigma(B) = (\sigma_i(A))_i$ the ordered vectors of singular values of A and B respectively, then*

$$|\langle A, B\rangle| = |\mathrm{tr}(A^*B)| \leq \sigma_1(A)\sigma_1(B) + \ldots + \sigma_n(A)\sigma_n(B)\,.$$

Corollary 3.2.4. *For any matrix $A, B \in \mathcal{M}_{m,n}(\mathbb{K})$*

$$\|A - B\|_F \geq \|\sigma(A) - \sigma(B)\|_2\,.$$

Given a matrix $A \in \mathcal{M}_{m,n}(\mathbb{K})$, suppose that (one of) its SVD is given by $A = U\Sigma V^*$. For any $0 \leq k < n$, define the truncated version of such SVD by

$$\mathcal{T}_k(A) \coloneqq U_k\Sigma_k V_k^*$$

where $\Sigma_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ its the diagonal matrix with elements given by the first k largest singular values of A, and U_k and V_k are the matrices formed by the first k columns of U and V respectively.

Remark 3.2.5. *The truncated SVD $\mathcal{T}_k(A)$ is uniquely defined if and only if $\sigma_{k+1}(A) < \sigma_k(A)$. In fact, if $\sigma_{k+1}(A) = \sigma_k(A)$, one may combine the corresponding singular vectors to form different SVDs of A which give rise to different matrices $\mathcal{T}_k(A)$. For example, the identity matrix $I \in \mathcal{M}_n(\mathbb{K})$ can be written in the form $I = QQ^*$ for any unitary matrix Q, which constitute different SVDs of I. For each of these, we can set $\mathcal{T}_k(I) = Q_kQ_k^*$ where Q_k is the matrix formed by the first k columns of Q.*

Remark 3.2.6. *For any unitarily invariant norm $\|\cdot\|$,*

$$\|\mathcal{T}_k(A) - A\| = \|\mathrm{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_n)\|.$$

In particular,

$$\|\mathcal{T}_k(A) - A\|_F = \sqrt{\sigma_{k+1}^2(A) + \ldots + \sigma_n^2(A)}\,, \quad \|\mathcal{T}_k(A) - A\|_2 = \sigma_{k+1}(A)\,. \tag{3.2.1}$$

The matrix $\mathcal{T}_k(A)$ can be understood can be understood as the best approximation of A of rank less or equal to k, in the following sense.

Theorem 3.2.7. *For any $A \in \mathcal{M}_{m,n}(\mathbb{K})$, and any unitarily invariant norm $\|\cdot\|$,*

$$\|A - \mathcal{T}_k(A)\| = \min\{\|A - B\|\ ,\ \mathrm{rank}(B) \leq k\}\,. \tag{3.2.2}$$

Remark 3.2.8. *Using either the spectral or the Frobenius norm in theorem 3.2.7, by equation (3.2.1) we find that the distance of A from the set of singular matrices is precisely $\sigma_n(A)$.*

Remark 3.2.9 (Uniqueness)**.** *Note that the set of matrices A with $\mathrm{rank}(A) \leq k$ is not convex, and consequently $\mathcal{T}_k(A)$ may not be the unique solution of (3.2.2). In fact:*

- *When using the Frobenius norm, one can deduce from Corollary 3.2.4 that the truncated SVDs of A, $\mathcal{T}_k(A)$, constitute all solutions of the minimization problem (3.2.2). In particular, in this case, such problem admits a unique solution if and only if $\sigma_{k+1}(A) < \sigma_k(A)$.*
- *When using the spectral norm, problem (3.2.2) has a unique solution if and only if $\sigma_{k+1}(A) = 0$ and this is given by $\mathcal{T}_k(A) = A$. If this is not the case, one has an infinite number of solutions. For example, for any $0 \leq \delta \leq \sigma_{k+1}(A)$,*

$$U_k(\Sigma_k - \delta I_k)V_k^*$$

is also a minimizer of (3.2.2).

Remark 3.2.10 (Relation with the PCA)**.** *In the setting of Section , given a data set of n m-dimensional observations, one can construct a matrix $X \in \mathcal{M}_{m,n}(\mathbb{K})$ as in equation (2.3.1) whose column are the observed vectors normalised so that the sample mean is zero. Then, $\mathcal{T}_k(X)$ is the orthogonal projection of the normalized data-set onto the space generated by the first k principal components. In fact, assuming that $X = U\Sigma V^*$, such projection is given by*

$$\tilde{X} := U_k U_k^* X = U_k [I_k|0]\, \Sigma V^* = \mathcal{T}_k(X).$$

where $[I_k|0] \in \mathcal{M}_{m,n}(\mathbb{K})$ and $I_k \in \mathcal{M}_k(\mathbb{K})$ is the identity matrix.

Remark 3.2.11 (Stability of low rank approximations)**.** *Note that if $\|\cdot\|$ is unitarily invariant and A has rank $r \leq k$, then applying the triangular inequality and then using Theorem 3.2.7, we have*

$$\|\mathcal{T}_k(A+E) - A\| \leq \|\mathcal{T}_k(A+E) - (A+E)\| + \|E\| \leq 2\|E\|\,.$$

Then, if A is arbitrary

$$\begin{aligned}\|\mathcal{T}_k(A+E) - \mathcal{T}_k(A)\| &= \|\mathcal{T}_k(\mathcal{T}_k(A) + (A + E - \mathcal{T}_k(A))) - \mathcal{T}_k(A)\| \\ &\leq 2\|(A + E - \mathcal{T}_k(A))\| \leq 2\|E\| + 2\|A - \mathcal{T}_k(A)\|\,.\end{aligned}$$

This means that if $\|A - \mathcal{T}_k(A)\|$ is small and E represents small noise, then the perturbed truncated SVD $\mathcal{T}_k(A+E)$ is close to $\mathcal{T}_k(A)$.

3.3 Condition number of a matrix

Consider the linear system:

$$Ax = b \tag{3.3.1}$$

In practice, its solution is never exact because of:

- errors in the data: evaluation of the coefficients of A or b;
- rounding errors: floating point represenation of numbers on machine.

In practice we solve:

$$(A + \delta A)y = b + \delta b$$

In the following, our aim is to to quantify how far y is from x.

Remark 3.3.1 (Stability analysis via the SVD)**.** *Suppose that $A \in \mathcal{M}_n(\mathbb{Z})$ is invertible and its SVD is given by $A = U\mathrm{diag}(\sigma_1, \dots, \sigma_n)V^*$. Then the solution to system is given by*

$$y = \sum_{i=1}^{n} \frac{1}{\sigma_i} \langle u_i, b \rangle v_i \,,$$

where u_i and v_i are the ith columns of U and V, respectively. We deduce that if we let $\sigma_n \to 0$, $\|y\| \to +\infty$, and therefore the system becomes less and less stable by reducing σ_n. This is expected given the meaning of σ_n (see remark 3.2.8).

Definition 3.3.2 (Condition number)**.** Let $\|\cdot\|$ be a matrix norm on $\mathcal{M}_n(\mathbb{K})$ and A be an invertible matrix. The condition number of A is the quantity:

$$\mathrm{cond}(A) \coloneqq \|A\|\|A^{-1}\| \,.$$

The condition number associate with the induced p-norms $\|\cdot\|_p$ is denoted $\mathrm{cond}_p(A)$.

Example (Intersection of lines)*.* Consider the problem of finding the intersection of two almost parallel lines:

$$\begin{cases} x + (1+\varepsilon)y = 1 \,, \\ (1+\varepsilon)x + y = 1 \,. \end{cases}$$

where ε is a small constant. Note that a small change in ε determines a dramatic change in the intersection point. In this case, we have

$$A = \begin{bmatrix} 1 & 1+\varepsilon \\ 1+\varepsilon & 1 \end{bmatrix}, \quad \mathrm{cond}_2(A) = \rho(A)\rho(A^{-1})$$

(note that A^{-1} is symmetric since $(A^{-1})^T = (A^T)^{-1}$). Then $\mathrm{cond}(A) = (2+\varepsilon)/\varepsilon \to \infty$ as $\varepsilon \to 0$.

Theorem 3.3.3. *Let $A \in \mathcal{M}_n(\mathbb{K})$ be an invertible matrix, $b \in \mathbb{K}^n \setminus \{0\}$. Let $\|\cdot\|$ be a norm on $\mathbb{K}^n$, we denote by $\|\cdot\|$ also the induced norm on $\mathcal{M}_n(\mathbb{K})$. If $x \in \mathbb{K}^n$ solves $Ax = b$ and $\delta x \in \mathbb{K}^n$ is such that*

$$A(x + \delta x) = b + \delta b$$

for a given $\delta b \in \mathbb{K}^n$, then

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A) \frac{\|\delta b\|}{\|b\|}$$

Theorem 3.3.4. *Let $A \in \mathcal{M}_n(\mathbb{K})$ be an invertible matrix, $b \in \mathbb{K}^n$. Let $\|\cdot\|$ be a norm on $\mathbb{K}^n$, we denote by $\|\cdot\|$ also the induced norm on $\mathcal{M}_n(\mathbb{K})$. If $x \in \mathbb{K}^n$ solves $Ax = b$ and $\delta x \in \mathbb{K}^n$ is such that*

$$(A + \delta A)(x + \delta x) = b$$

then

$$\frac{\|\delta x\|}{\|x+\delta x\|} \leq \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}\,.$$

Remark 3.3.5 (Properties of the condition number)**.** *For any invertible matrix A,*

- $\mathrm{cond}(A) \geq 1$,
- $\forall \alpha \in \mathbb{K}^*, \quad \mathrm{cond}(\alpha A) = \mathrm{cond}(A)$,
- $\mathrm{cond}_2(A) = \sigma_n(A)/\sigma_1(A)$*, where* $\sigma_1(A)$ *and* $\sigma_n(A)$ *are respectively the largest and the smallest singular values of* A*,*
- *For* A *hermitian* $\mathrm{cond}_2(A) = \max_i |\lambda_i| / \min_i |\lambda_i|$*, where* $\{\lambda_i\}_i$ *are the eignevalues of* A*,*
- *For any unitary transformation* U*,* $\mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(U)$*.*

Remark 3.3.6. *The determinant is not a good indicator of the condition number of a matrix. Take, for example,*

$$A' = \frac{1}{\varepsilon}A = \frac{1}{\varepsilon}\begin{bmatrix} 1 & 1+\varepsilon \\ 1+\varepsilon & 1 \end{bmatrix}\,.$$

Bibliography

[1] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

[2] Gabriel Peyré. *Mathematical Foundations of Data Sciences*. Available at https://mathematical-tours.github.io/book-sources/FundationsDataScience.pdf. 2021.

[3] Gilbert Strang. *Linear algebra and its applications*. Belmont, CA: Thomson, Brooks/Cole, 2006.