Least Squares Approximation

1 Introduction

In many applications we want to find an approximation for a function, for example for differential equations.

Example problem: We want to understand how a calculator or computer can evaluate $\sin x$ for a given value x. The processor can essentially only perform addition, multiplication, division. Therefore we can try to approximate the function with a polynomial.

We use the notation $\mathscr{P}_n$ for polynomials of degree$\leq n$:

$$\mathscr{P}_n = \left\{ c_0 + c_1 x + \cdots + c_n x^n \mid c_j \in \mathbb{R} \right\}$$

It is sufficient to consider $x \in [0, \pi/2]$. Instead of approximating $\sin x$ for $x \in [0, \pi/2]$ we can consider the function $u(x) = \sin(\frac{\pi}{2} x)$ for $x \in [0,1]$.

We want to find a polynomial $w(x) \in \mathscr{P}_n$ such that the error is minimal in the least squares sense:

$$\|u - w\|_2^2 = \int_0^1 |u(x) - w(x)|^2 \, dx \text{ is minimal}$$

1.1 Abstract version of the least squares problem

Let $\mathscr{V}$ denote a **vector space** which is equipped with an **inner product** $(\cdot, \cdot)$. This defines a norm $\|v\| := (v, v)^{1/2}$.
We are given: a vector $u \in \mathscr{V}$ and an n-dimensional subspace

$$\mathscr{W} := \operatorname{span}\left\{ a^{(1)}, \ldots, a^{(n)} \right\}$$

where the vectors $a^{(1)}, \ldots, a^{(n)}$ are **assumed to be linearly independent** (otherwise some of them are redundant and can be dropped).

Least Squares Problem: Find a vector $w = c_1 a^{(1)} + \cdots + c_n a^{(n)}$ such that $\|u - w\|$ is minimal.

2 Algorithm 1: Normal Equations

We are given a point u outside of the subspace $\mathscr{W}$. We want to find the closest point w in this subspace.

Geometric intuition tells us that $w - u$ should be orthogonal on the subspace $\mathscr{W}$.

This orthogonality condition is called

Normal Equations: Find $w = c_1 a^{(1)} + \cdots + c_n a^{(n)}$ such that $\left(w - u, a^{(j)}\right) = 0$ for $j = 1, \ldots, n$.

The normal equations are n linear equations for n unknowns $c_1, \ldots, c_n$. We can write them in matrix-vector notation as

$$Mc = b$$

Here $c = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \in \mathbb{C}^n$, $M \in \mathbb{C}^{n\times n}$ is the so-called **Gram matrix** with entries $M_{jk} = \left(a^{(k)}, a^{(j)}\right)$, and the right-hand side vector $b \in \mathbb{C}^n$ has the entries $b_j = \left(u, a^{(j)}\right)$ for $j = 1, \ldots, n$.

After we have computed the entries of M and b we have to solve the $n \times n$ linear system $Mc = b$. We can use Gaussian elimination with pivoting for this. In Matlab we can use `c=M\b` .

The linear system has a unique solution if and only if the matrix M is nonsingular.

Proposition 1. *M is nonsingular.*

Proof. We have to show that $Mv = 0$ implies $v = 0$. □

We claim that the unique solution c of the normal equations is the solution of the least squares problem.

Proposition 2. *The solution vector c of the normal equations yields the unique solution of the least squares problem.*

Proof. Let $w = c_1 a^{(1)} + \cdots + c_n a^{(n)}$. We want to show that any $\tilde{w} \in \mathscr{W}$ different from w will have $\|\tilde{w} - u\| > \|w - u\|$. We can write $\tilde{w} = w + v$ with some $v \in W$, then multiplying out and using the normal equations we obtain

$$\|u - \tilde{w}\|^2 = (u - w - v, u - w - v) = (u - w, u - w) - \underbrace{(v, u - w)}_{0} - \underbrace{(u - w, v)}_{0} + (v, v)$$

$$\|u - \tilde{w}\|^2 = \|u - w\|^2 + \|v\|^2$$

which is actually Pythagoras' theorem for the triangle with vertices $u, w, \tilde{w}$. Hence $\|u - \tilde{w}\|^2 \geq \|u - w\|^2$ and we have equality only if $v = d_1 a^{(1)} + \cdots + d_n a^{(n)} = 0$. Since $a^{(1)}, \ldots, a^{(n)}$ are by assumption linearly independent we must have $d_1 = \cdots = d_n = 0$. Hence the coefficients $c_1, \ldots, c_n$of the least squares solution are unique. □

2.1 Case of $\mathscr{V} = \mathbb{C}^m$

We now consider the special case $\mathscr{V} = \mathbb{C}^m$ with the inner product

$$(u, v) = u_1 \overline{v}_1 + \cdots + u_n \overline{v}_n = [\overline{v}_1, \ldots, \overline{v}_n] \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

where we use the notation $v^H := \overline{v}^\top$. Note that in Matlab v^H is obtained by `v'` .

We need to have $m \geq n$, otherwise the vectors $a^{(1)}, \ldots, a^{(n)} \in \mathbb{C}^m$ are linearly dependent. We define the matrix $A \in \mathbb{C}^{m\times n}$ using the columns $a^{(1)}, \ldots, a^{(n)}$

$$A = \left[a^{(1)}, \ldots, a^{(n)}\right]$$

Then we can write $w = c_1 a^{(1)} + \cdots + c_n a^{(n)} = Ac$.

Least Squares Problem: Given $A \in \mathbb{C}^{m\times n}$ and $u \in \mathbb{C}^n$ find a vector $c \in \mathbb{C}^n$ such that

$$\|Ac - u\|_2 \text{ is minimal}$$

Note that we can write the Gram matrix M and the right-hand side vector b as

$$M = A^H A, \qquad b = A^H u$$

Hence we can write the normal equations as

$$A^H A c = A^H u.$$

2.2 Notation $A = \left[a^{(1)}, \ldots, a^{(n)}\right]$

We can use the "matrix-vector" notation also in the general case of a vector space $\mathscr{V}$, e.g. for functions in $\mathscr{V} = L^2([0,1])$. For n vectors $a^{(1)}, \ldots, a^{(n)} \in \mathscr{V}$ we denote this n-tupel by $A = \left[a^{(1)}, \ldots, a^{(n)}\right]$. We can then write a linear combination with coefficient vector $c \in \mathbb{C}^n$ with the "matrix-vector" notation

$$c_1 a^{(1)} + \cdots + c_n a^{(n)} = \left[a^{(1)}, \ldots, a^{(n)}\right] \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = Ac$$

Note that we have for vectors $v = Ac$ and $\tilde{v} = A\tilde{c}$ with $c, \tilde{c} \in \mathbb{C}^n$ that

$$(v, \tilde{v}) = (Ac, A\tilde{c}) = \tilde{c}^H M c = (Mc, \tilde{c})$$

with the Gram matrix M.

3 Algorithm 2: Orthogonalization

We can use a new basis $p^{(1)}, \ldots, p^{(n)}$ for $\mathscr{W} = \operatorname{span}\left\{a^{(1)}, \ldots, a^{(n)}\right\}$. Then we can write $w \in \mathscr{W}$ as $w = d_1 p^{(1)} + \cdots + d_n p^{(n)}$ and we obtain the normal equations

$$\begin{bmatrix} \left(p^{(1)}, p^{(1)}\right) & \cdots & \left(p^{(n)}, p^{(1)}\right) \\ \vdots & & \vdots \\ \left(p^{(1)}, p^{(n)}\right) & \cdots & \left(p^{(n)}, p^{(n)}\right) \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} = \begin{bmatrix} \left(u, p^{(1)}\right) \\ \vdots \\ \left(u, p^{(n)}\right) \end{bmatrix}$$

If the vectors $p^{(1)}, \ldots, p^{(n)}$ are orthogonal on each other, the matrix is diagonal and the linear system is very easy to solve.

3.1 Gram-Schmidt orthogonalization and decomposition $A = PS$

Assume the vectors $a^{(1)}, \ldots, a^{(n)} \in \mathscr{V}$ are linearly independent. We want to find vectors $p^{(1)}, \ldots, p^{(n)}$ which are orthogonal on each other and satisfy

$$\operatorname{span}\left\{p^{(1)}, \ldots, p^{(k)}\right\} = \operatorname{span}\left\{a^{(1)}, \ldots, a^{(k)}\right\} \qquad \text{for } k = 1, \ldots, n$$

We define

$$\begin{aligned} p^{(1)} &:= a^{(1)} \\ p^{(2)} &:= a^{(2)} - s_{12} p^{(1)} \\ p^{(3)} &:= a^{(3)} - s_{13} p^{(1)} - s_{23} p^{(2)} \\ &\vdots \end{aligned}$$

where we choose the coefficients s_{kj} such that we have $\left(p^{(j)}, p^{(k)}\right) = 0$ for $j = 1, \ldots, n$, $k = 1, \ldots, j-1$.

E.g., the condition $\left(p^{(2)}, p^{(1)}\right)$ gives

$$s_{12} = \frac{\left(a^{(2)}, p^{(1)}\right)}{\left(p^{(1)}, p^{(1)}\right)}$$

The conditions $\left(p^{(3)}, p^{(1)}\right) = 0$ and $\left(p^{(3)}, p^{(2)}\right) = 0$ give

$$s_{13} = \frac{\left(a^{(3)}, p^{(1)}\right)}{\left(p^{(1)}, p^{(1)}\right)}, \qquad s_{23} = \frac{\left(a^{(3)}, p^{(2)}\right)}{\left(p^{(2)}, p^{(2)}\right)}$$

Therefore we obtain

Gram-Schmidt orthogonalization algorithm:

For $j = 1, \ldots, n$:

$$p^{(j)} := a^{(j)} - \sum_{k=1}^{j-1} s_{kj} p^{(k)} \text{ where } s_{kj} := \frac{\left(a^{(j)}, p^{(k)}\right)}{\left(p^{(k)}, p^{(k)}\right)}$$

We see that $p^{(j)} \in \operatorname{span}\left\{a^{(1)}, \ldots, a^{(j)}\right\}$ for $j = 1, \ldots, n$. All the vectors $p^{(j)}$ are nonzero: E.g., if obtained $p^{(3)} = 0$ this would mean that $a^{(3)} \in \operatorname{span}\left\{p^{(1)}, p^{(2)}\right\} \subset \operatorname{span}\left\{a^{(1)}, a^{(2)}\right\}$, hence the vectors $a^{(1)}, a^{(2)}, a^{(3)}$ are linearly independent, in contradiction to our assumption that $a^{(1)}, \ldots, a^{(n)}$ are linearly independent.

Note that we have

$$\begin{aligned} a^{(1)} &:= p^{(1)} \\ a^{(2)} &:= p^{(2)} + s_{12} p^{(1)} \\ a^{(3)} &:= p^{(3)} + s_{13} p^{(1)} + s_{23} p^{(2)} \\ &\vdots \end{aligned}$$

i.e., we have the decomposition

$$\left[a^{(1)}, \ldots, a^{(n)}\right] = \left[p^{(1)}, \ldots, p^{(n)}\right] \begin{bmatrix} 1 & s_{12} & \cdots & s_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & s_{n-1,n} \\ & & & 1 \end{bmatrix}$$

$$A = PS$$

with the upper triangular matrix $S \in \mathbb{C}^{n \times n}$. This is the "QR decomposition **without** normalization".

We can write $w \in \mathscr{W}$ as

$$w = Ac = Pd$$

Because of $A = PS$ the vectors $c, d \in \mathbb{C}^n$ are related by $Sc = d$.

This gives the following **algorithm for solving the least squares problem**:

1. Given $a^{(1)}, \ldots, a^{(n)}$ use the Gram-Schmidt algorithm to find $p^{(1)}, \ldots, p^{(n)}$ and the matrix S
2. Given $u \in \mathscr{V}$ compute the coefficients $d_1, \ldots, d_n$ from Section 3:

$$\text{For } j = 1, \ldots, n: \qquad d_j := \frac{\left(u, p^{(j)}\right)}{\left(p^{(j)}, p^{(j)}\right)}$$

3. We obtain $w \in \mathscr{W}$ with $\|w - u\| = \min$

$$w = d_1 p^{(1)} + \cdots + d_n p^{(n)}$$

We obtain the coefficient vector c with $\|Ac - u\| = \min$ by solving the linear system

$$Sc = d$$

using back substitution.

3.2 Orthonormalization and decomposition $A = QR$

Sometimes it is useful to have an **orthonormal basis** (ON basis) $q^{(1)}, \ldots, q^{(n)}$ for $\operatorname{span}\left\{a^{(1)}, \ldots, a^{(n)}\right\}$. We can obtain this by normalizing the vectors $p^{(1)}, \ldots, p^{(n)}$. If we multiply row j of the matrix S by $\left\|p^{(j)}\right\|$

$$q^{(j)} = \frac{1}{\left\|p^{(j)}\right\|} p^{(j)}, \qquad [r_{jj}, \ldots, r_{jn}] := \left\|p^{(j)}\right\| \cdot [1, s_{j,j+1}, \ldots, s_{jn}]$$

we get from $A = PS$ the so-called QR decomposition

$$\left[a^{(1)},\ldots,a^{(n)}\right] = \left[q^{(1)},\ldots,q^{(n)}\right] \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{bmatrix}$$

$$A = QR$$

where the matrix R is upper triangular with positive entries on the diagonal.

3.3 Case of $\mathscr{V} = \mathbb{C}^m$

We are given n linearly independent vectors $a^{(1)},\ldots,a^{(n)} \in \mathbb{C}^m$. Therefore we must have $m \geq n$. This corresponds to the matrix $A = \left[a^{(1)},\ldots,a^{(n)}\right] \in \mathbb{C}^{m\times n}$ with linearly independent columns.

We then obtain the QR decomposition

$$\left[a^{(1)},\ldots,a^{(n)}\right] = \underbrace{\left[q^{(1)},\ldots,q^{(n)}\right]}_{\text{orthonormal}} \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{bmatrix}$$

$$\underset{m\times n}{A} = \underset{m\times n}{Q}\ \underset{n\times n}{R}$$

We can compute this in Matlab by `[Q,R]=qr(A,0)` . Note that Matlab uses a slightly different algorithm and therefore some of the columns of Q are multiplied by -1, and the corresponding r_{jj} are negative.

The vectors $q^{(1)},\ldots,q^{(n)}$ form an orthonormal basis for range$A$. This is a subspace of dimension n of the vector space V which has dimension $m \geq n$. We can extend the orthonormal basis $q^{(1)},\ldots,q^{(n)}$ to a basis $q^{(1)},\ldots,q^{(m)}$ of the whole space $\mathbb{C}^m$. Here the additional vectors $q^{(n+1)},\ldots,q^{(m)}$ form an orthonormal basis for the orthogonal complement of range$A$.

This yields the "full version" of the QR decomposition:

$$\left[a^{(1)},\ldots,a^{(n)}\right] = \underbrace{\left[q^{(1)},\ldots,q^{(n)},q^{(n+1)},\ldots,q^{(m)}\right]}_{\text{orthonormal}} \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$$

$$\underset{m\times n}{A} = \underset{m\times m}{\tilde{Q}}\ \underset{m\times n}{\tilde{R}}$$

We can compute this in Matlab by `[Qt,Rt]=qr(A,0)` .

Note that the additional part $\left[q^{(n+1)},\ldots,q^{(m)}\right]$ does not contribute anything to the decomposition since it is multiplied by zero.

In many applications (e.g. solving a least squares problem) the "economy size version" `[Q,R]=qr(A,0)` is sufficient. Only use the "full version" `[Qt,Rt]=qr(A,0)` if you need a basis for the orthogonal complement of range$A$.

The most accurate way to compute the QR decomposition in machine arithmetic is to transform the columns of A step by step to an upper triangular matrix $\tilde{R}$ using so-called "Householder reflections". This is the algorithm used by Matlab for the `qr` command.

4 Review of Hermitian matrices and matrix norms

4.1 Review: Hermitian matrices, positive definite matrices

We call a matrix $A \in \mathbb{C}^{n\times n}$ **Hermitian** if $A^H = A$. For a real matrix this is the same as symmetric.

Theorem 1. *Assume $A \in \mathbb{C}^{n\times n}$ is Hermitian. Then the eigenvalues $\lambda_1,\dots,\lambda_n$ are real, and there exists an orthonormal basis of eigenvectors $v^{(1)},\dots,v^{(n)}$: With $V = \left[v^{(1)},\dots,v^{(n)}\right]$ we have*

$$A = V \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} V^H, \qquad V^H V = I$$

We call a Hermitian matrix A **positive definite** if for all nonzero vectors v

$$v^H A v > 0$$

Note that a Hermitian matrix is positive definite if and only if all eigenvalues are positive.

Consider a matrix $A = \left[a^{(1)},\dots,a^{(n)}\right] \in \mathbb{C}^{m\times n}$ with linearly independent columns. Then the Gram matrix M with $M_{jk} = \left(a^{(k)}, a^{(j)}\right)$ is positive definite.

4.2 Matrix norms: 2-norm, Frobenius norm

A matrix $A \in \mathbb{C}^{m\times n}$ corresponds to a linear mapping $\mathbb{C}^n \to \mathbb{C}^m$, $c \mapsto Ac$.

We would like to have a bound such that

$$\forall v \in \mathbb{C}^n: \qquad \|Av\| \le C\,\|v\|$$

with the smallest possible constant C.

We can define a norm as the "maximal magnification factor":

$$\|A\| = \sup_{\substack{v\in\mathbb{C}^n \\ \|v\|=1}} \|Av\| = \sup_{\substack{v\in\mathbb{C}^n \\ \|v\|\le 1}} \|Av\|$$

Since the unit ball in $\mathbb{C}^n$ is compact there exists v with $\|v\| = 1$ such that $\|Av\|$ is maximal, and we can write max instead of sup.

For the vector norms $\|v\|_\infty$, $\|v\|_1$, $\|v\|_2$ we obtain in this way matrix norms $\|A\|_\infty$, $\|A\|_1$, $\|A\|_2$. One can show that

$$\|A\|_\infty = \max_{j=1,\dots,n} \sum_{k=1}^{n} \left|a_{jk}\right| \qquad \text{maximal } \textbf{row sum} \text{ of absolute values}$$

$$\|A\|_1 = \max_{k=1,\dots,n} \sum_{j=1}^{n} \left|a_{jk}\right| \qquad \text{maximal } \textbf{column sum} \text{ of absolute values}$$

$$\|A\|_2^2 = \lambda_{\max}\left(A^H A\right) \qquad \text{maximal eigenvalue of the Gram matrix } M = A^H A$$

The last line follows from

$$\|Ac\|_2^2 = \left\|c_1 a^{(1)} + \dots + c_n a^{(n)}\right\|_2^2 = \sum_{j=1}^{n}\sum_{k=1}^{n} c_j \bar{c}_k \left(a^{(j)}, a^{(k)}\right) = c^H M c \le \lambda_{\max}(M)\,\|c\|_2^2$$

We can also define the **Frobenius norm** (a.k.a. Hilbert-Schmidt norm)

$$\|A\|_F := \left(\left\|a^{(1)}\right\|_2^2 + \dots + \left\|a^{(n)}\right\|_2^2\right)^{1/2} = \left(M_{11} + M_{22} + \dots + M_{nn}\right)^{1/2}$$

Note that

$$\begin{aligned} \|Ac\|_2^2 &= \sum_{j=1}^{n}\sum_{k=1}^{n} c_j \bar{c}_k \left(a^{(j)}, a^{(k)}\right) \le \sum_{j=1}^{n}\sum_{k=1}^{n} \left|c_j\right| \left|c_k\right| \left\|a^{(j)}\right\|_2 \left\|a^{(k)}\right\|_2 \\ &= \left(\sum_{j=1}^{n} \left|c_j\right| \left\|a^{(j)}\right\|_2\right)^2 \le \left(|c_1|^2 + \dots + |c_n^2|\right) \underbrace{\left(\left\|a^{(1)}\right\|_2^2 + \dots + \left\|a^{(n)}\right\|_2^2\right)}_{\|A\|_F^2} \end{aligned}$$

$$\|Ac\|_2 \le \|A\|_F\,\|c\|_2$$

Hence we have $\|A\|_2 \le \|A\|_F$.

4.3 Case $A = \left[a^{(1)}, \ldots, a^{(n)}\right]$ with $a^{(j)} \in \mathscr{V}$

For a general inner product space $\mathscr{V}$ we can consider $A = \left[a^{(1)}, \ldots, a^{(n)}\right]$ with $a^{(j)} \in \mathscr{V}$. For $c \in \mathbb{C}^n$ we can consider the mapping $c \mapsto Ac = c_1 a^{(1)} + \cdots + c_n a^{(n)}$, $\mathbb{C}^n \to \mathscr{V}$ and define the norm

$$\|A\| = \sup_{\substack{c \in \mathbb{C}^n \\ \|c\|_2 = 1}} \|Ac\| = \sup_{\substack{c \in \mathbb{C}^n \\ \|c\|_2 \leq 1}} \|Ac\|$$

Since the unit ball in $\mathbb{C}^n$ is compact there exists v with $\|v\| = 1$ such that $\|Av\|$ is maximal, and we can write max instead of sup. By the same argument as above we have

$$\|A\|^2 = \lambda_{\max}(M)$$

with the Gram matrix $M \in \mathbb{C}^{n \times n}$. With the Frobenius norm

$$\|A\|_F := \left(\left\| a^{(1)} \right\|^2 + \cdots + \left\| a^{(n)} \right\|^2 \right)^{1/2}$$

we have with the same arguments as above

$$\|Ac\| \leq \|A\|_F \, \|c\|_2$$

and $\|A\| \leq \|A\|_F$.

5 The Singular Value Decomposition

5.1 Motivation

We are given vectors $a^{(1)}, \ldots, a^{(n)}$ in an inner product space $\mathscr{V}$. Assume that $a^{(1)}, \ldots, a^{(n)}$ are linearly independent, i.e. $A = \left[a^{(1)}, \ldots, a^{(n)}\right]$ has rank n.

We want to know: Are the vectors $a^{(1)}, \ldots, a^{(n)}$ "almost linearly dependent", i.e., is there a matrix $\tilde{A}$ of lower rank which is very close to A?

First we want to find the **closest rank 1 approximation**: Pick a unit vector $u^{(1)}$ and let $\mathscr{W}_1 = \operatorname{span} u^{(1)}$. Let $\tilde{a}^{(j)}$ be the point on $\mathscr{W}_1$ which is closest to $a^{(j)}$ for $j = 1, \ldots, n$, let $\tilde{A} = \left[\tilde{a}^{(1)}, \ldots, \tilde{a}^{(n)}\right]$. We can measure the total distance of the points $a^{(1)}, \ldots, a^{(n)}$ to the subspace $\mathscr{W}_1$ by

$$\left\| a^{(1)} - \tilde{a}^{(1)} \right\|^2 + \cdots + \left\| a^{(n)} - \tilde{a}^{(n)} \right\|^2 = \left\| A - \tilde{A} \right\|_F^2$$

We pick the unit vector $u^{(1)}$ such that this expression becomes minimal.

Next we want to **find the closest rank 2 approximation**: Pick an ON basis $u^{(1)}, u^{(2)}$ and let $\mathscr{W}_2 = \operatorname{span}\left\{u^{(1)}, u^{(2)}\right\}$. Let $\tilde{a}^{(j)}$ be the point on $\mathscr{W}_2$ which is closest to $a^{(j)}$ for $j = 1, \ldots, n$, let $\tilde{A} = \left[\tilde{a}^{(1)}, \ldots, \tilde{a}^{(n)}\right]$. Now we want to pick the ON vectors $u^{(1)}, u^{(2)}$ such that

$$\left\| a^{(1)} - \tilde{a}^{(1)} \right\|^2 + \cdots + \left\| a^{(n)} - \tilde{a}^{(n)} \right\|^2 = \left\| A - \tilde{A} \right\|_F^2$$

becomes minimal.

In this way we will obtain an ON basis $u^{(1)}, \ldots, u^{(n)}$ for $\operatorname{span}\left\{a^{(1)}, \ldots, a^{(n)}\right\}$. For each $k = 1, 2, \ldots, n-1$ we also obtain the best rank k approximation $A_{(k)}$ which minimizes $\left\| A - A_{(k)} \right\|_F$.

Alternatively we can measure the error of a rank k matrix $\tilde{A}$ by the norm $\left\| A - \tilde{A} \right\|$ instead of the Frobenius norm $\left\| A - \tilde{A} \right\|_F$. It turns out that minimizing $\left\| A - \tilde{A} \right\|$ over all rank k matrices $\tilde{A}$ leads to the same matrix $A_{(k)}$ we obtained before for the Frobenius norm.

5.2 Construction of the Singular Value Decomposition

We consider an inner product space $\mathscr{V}$ and vectors $a^{(1)},\dots,a^{(n)} \in \mathscr{V}$. Let $A = [a^{(1)},\dots,a^{(n)}]$. We first assume that these vectors are linearly independent. The Gram matrix $M \in \mathbb{C}^{n\times n}$ has the entries $M_{jk} = \left(a^{(k)}, a^{(j)}\right)$. This matrix is Hermitian, i.e., $M^H = M$. Since the vectors are linearly independent it is positive definite.

Therefore the matrix M has real positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$, and an orthonormal basis of eigenvectors $v^{(1)},\dots,v^{(n)}$. Hence the matrix $V = \left[v^{(1)},\dots,v^{(n)}\right] \in \mathbb{C}^{n\times n}$ is unitary, i.e., $V^H V = I$.

Note that we have for vectors $u = c_1 a^{(1)} + \cdots + c_n a^{(n)} = Ac$ and $\tilde{u} = \tilde{c}_1 a^{(1)} + \cdots + \tilde{c}_n a^{(n)} = A\tilde{c}$ the inner product

$$(u, \tilde{u}) = \tilde{c}^H M c$$

We now define the vectors $\tilde{u}^{(j)} := Av^{(j)} \in \mathscr{V}$ for $j = 1,\dots,n$. Then we have the inner products

$$\left(\tilde{u}^{(j)}, \tilde{u}^{(k)}\right) = v^{(k)H} M v^{(j)} = v^{(k)} \lambda_j v^{(j)} = \begin{cases} 0 & \text{if } j \neq k \\ \lambda_j & \text{if } j = k \end{cases},$$

i.e., these vectors are orthogonal. We can normalize them to length 1 by defining

$$\sigma_j := \lambda_j^{1/2}, \qquad u^{(j)} := \sigma_j^{-1} A v^{(j)}$$

We can write the last equation as

$$\boxed{Av^{(j)} = \sigma_j u^{(j)}, \qquad j = 1,\dots,n} \tag{1}$$

The numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$ are called **singular values**. The vectors $v^{(1)},\dots,v^{(n)}$ are called **right singular vectors**, the vectors $u^{(1)},\dots,u^{(n)}$ are called **left singular vectors**.

Theorem 2. *Let $\mathscr{V}$ be an inner product space. Let $A = \left[a^{(1)},\dots,a^{(n)}\right]$ with $a^{(j)} \in \mathscr{V}$ linearly independent. Then we have the singular value decomposition*

$$\left[a^{(1)},\dots,a^{(n)}\right] = \left[u^{(1)},\dots,u^{(n)}\right] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} V^H$$

where $u^{(1)},\dots,u^{(n)} \in \mathscr{V}$ are orthonormal, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$, $V \in \mathbb{C}^{n\times n}$ unitary.

In other words, we have an orthonormal basis $v^{(1)},\dots,v^{(n)}$ of $\mathbb{C}^n$ and orthonormal vectors $u^{(1)},\dots,u^{(n)} \in \mathscr{V}$ such that

$$\boxed{Av^{(j)} = \sigma_j u^{(j)} \qquad j = 1,\dots,n}$$

Let $v \in \mathbb{C}^n$. We can then write v as a linear combination of $v^{(1)},\dots,v^{(n)}$:

$$v = c_1 v^{(1)} + \cdots + c_n v^{(n)} = Vc \qquad \text{with } c_j = \left(v, v^{(j)}\right), \qquad c = V^H v$$

Hence

$$Av = \left[u^{(1)},\dots,u^{(n)}\right] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \begin{bmatrix} v^{(1)H} \\ \vdots \\ v^{(n)H} \end{bmatrix} v$$

and

$$\boxed{A = \left[u^{(1)},\dots,u^{(n)}\right] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \begin{bmatrix} v^{(1)H} \\ \vdots \\ v^{(n)H} \end{bmatrix} = \sigma_1 u^{(1)} v^{(1)H} + \cdots + \sigma_n u^{(n)} v^{(n)H}}$$

This means that we express the columns $[a^{(1)},\dots,a^{(n)}]$ as a linear combination of the vectors $u^{(1)},\dots,u^{(n)}$.

If we use only the first k terms we obtain the **rank k approximation** $A_{(k)}$

$$\boxed{A_{(k)} := \sigma_1 u^{(1)} v^{(1)H} + \cdots + \sigma_k u^{(k)} v^{(k)H}}$$

and

$$A - A_{(k)} = \sigma_{k+1} u^{(k+1)} v^{(k+1)H} + \cdots + \sigma_n u^{(n)} v^{(n)H}$$

Let $\tilde{c} \in \mathbb{C}^n$ and $c := V^H \tilde{c}$. Then $\tilde{c} = Vc = c_1 v^{(1)} + \cdots + c_n v^{(n)}$ and have

$$A\tilde{c} = \sigma_1 c_1 u^{(1)} + \cdots + \sigma_n c_n u^{(n)}$$
$$A_{(k)}\tilde{c} = \sigma_1 c_1 u^{(1)} + \cdots + \sigma_k c_k u^{(k)}$$

hence

$$\left(A - A_{(k)}\right)\tilde{c} = \sigma_{k+1} c_{k+1} u^{(k+1)} + \cdots + \sigma_n c_n u^{(k)}$$
$$\left\|\left(A - A_{(k)}\right)\tilde{c}\right\| = \left\| \begin{bmatrix} \sigma_{k+1} c_{k+1} \\ \vdots \\ \sigma_n c_n \end{bmatrix} \right\|_2 \leq \sigma_{k+1} \underbrace{\|c\|_2}_{\|\tilde{c}\|_2}$$

and we have equality for $c_{k+1} = 1$, all other $c_j = 0$. Therefore $\left\|A - A_{(k)}\right\| = \sigma_{k+1}$.

5.3 Singular Value Decomposition for $\mathscr{V} = \mathbb{C}^m$

Assume $m \geq n$. We have the "thin SVD" (a.k.a. economy-size SVD)

$$A = U \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} V^H$$

We can extend $U = \left[u^{(1)}, \ldots, u^{(n)}\right]$ to an orthonormal basis $\tilde{U} = \left[u^{(1)}, \ldots, u^{(n)}, u^{(n+1)}, \ldots, u^{(m)}\right]$ of $\mathbb{C}^m$ and have

$$A = \tilde{U} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} V^H$$

This is the "full size" SVD.

5.4 SVD gives rank k approximation which minimizes $\left\|A - \tilde{A}\right\|_2$ and $\left\|A - \tilde{A}\right\|_F$

Let $B \in \mathbb{C}^{m \times n}$. Assume that $m \geq n$ (the proof for $m \leq n$ works in the same way). We choose an ON basis $v^{(1)}, \ldots, v^{(n)}$ for $\mathbb{C}^n$, and an ON basis $u^{(1)}, \ldots, u^{(m)}$ for $\mathbb{C}^m$. Then the mapping with respect to the new bases is described by the matrix $\hat{B} = U^H B V$. Note that $\|B\|_2 = \left\|\hat{B}\right\|_2$ and $\|B\|_F = \left\|\hat{B}\right\|_F$.

Therefore we only have to find the best rank k approximations for a matrix of the form

$$A = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$$

Let B be a matrix of rank $k < n$. We need to show that $\|A-B\|_2 \geq \sigma_{k+1}$ and $\|A-B\|_F \geq \left(\sigma_{k+1}^2 + \cdots + \sigma_n^2\right)^{1/2}$.

The first $k+1$ columns of B must be linearly dependent, hence there exists a nonzero vector $c = \begin{bmatrix} c_1 \\ \vdots \\ c_{k+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ such that $Bc = 0$.

Let $B = A + E$, then $(A+E)c = 0$, hence

$$\sigma_{k+1}\|c\| \leq \left\| \begin{bmatrix} \sigma_1 c_1 \\ \vdots \\ \sigma_{k+1} c_{k+1} \end{bmatrix} \right\|_2 = \|Ac\|_2 = \|Ec\|_2 \leq \|E\|_2 \|c\|_2$$

Since $\|c\| \neq 0$ we obtain $\|B-A\| = \|E\| \geq \sigma_{k+1}$ for any matrix $B \in \mathbb{C}^{m\times n}$ of rank k.

Note that this proof also shows that for B of rank $k = n-1$ we have $\|A-B\|_F = \|E\|_F \geq \sigma_n$.

Consider now the case $k = n-2$. For two ON vectors c, c' we have

$$\|Ec\|_2^2 + \left\|Ec'\right\|_2^2 \leq \|E\|_F^2$$

Since the null space of B has dimension 2 we can pick an ON basis c, c' of the null space. Then

$$\sigma_n^2 + \sigma_{n-1}^2 \leq \left\| \begin{bmatrix} \sigma_1 c_1 \\ \vdots \\ \sigma_n c_n \end{bmatrix} \right\|_2^2 + \left\| \begin{bmatrix} \sigma_1 c_1' \\ \vdots \\ \sigma_n c_n' \end{bmatrix} \right\|_2^2 = \|Ac\|_2^2 + \left\|Ac'\right\|_2^2 = \|Ec\|_2^2 + \left\|Ec'\right\|_2^2 \leq \|E\|_F^2$$

5.5 Application: Principal Component Analysis (PCA)

We measure m different quantities of n items (e.g., height, length, weight, age of 10 animals) and obtain n data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^m$. Typically m is large (say, $m \geq 5$), so we cannot make a plot of the points in $\mathbb{R}^m$ to understand the data.

The best approximation by a single point is the mean $\mu := \frac{1}{N}\sum_{j=1}^n x^{(j)}$. The **variation** of the data points from each other is characterized by

$$\rho := \sum_{j=1}^n \left\| x^{(j)} - \mu \right\|^2$$

Question: Are the data points approximately located on an affine space of lower dimension? E.g., I can consider a plane P through the point μ: $P = \left\{\mu + c_1 u^{(1)} + c_2 u^{(2)} \mid c_1, c_2 \in \mathbb{R}\right\}$ where $u^{(1)}, u^{(2)}$ are orthonomal vectors.

Let $\tilde{x}^{(j)}$ be the projection of $x^{(j)}$ onto the plane P. Then by Pythagoras

$$\rho := \sum_{j=1}^n \left\| x^{(j)} - \mu \right\|^2 = \underbrace{\sum_{j=1}^n \left\| \tilde{x}^{(j)} - \mu \right\|^2}_{\tilde{\rho}} + \underbrace{\sum_{j=1}^n \left\| \tilde{x}^{(j)} - x^{(j)} \right\|^2}_{\delta}$$

We now pick a plane P which makes δ minimal. In many cases we have that $\tilde{\rho}$ is now very close to ρ, e.g., $\tilde{\rho} \approx 0.95\rho$. This means that 95% of the variation in the data can be explained by the 2-dimensional plane P, i.e., only two directions $u^{(1)}, u^{(2)} \in \mathbb{C}^m$. Each of the projected points has the form

$$\tilde{x}^{(j)} = \mu + c_1^{(j)} u^{(1)} + c_2^{(j)} u^{(2)} = \mu + \left[u^{(1)}, u^{(2)}\right] c^{(j)}$$

and the 2-dimensional plot of the points $c^{(1)}, \ldots, c^{(n)} \in \mathbb{R}^2$ still contains 95% of the variation of the original data.

Let $a^{(j)} := x^{(j)} - \mu$ and $A = \left[a^{(1)}, \ldots, a^{(n)}\right]$. Then $\rho = \|A\|_F^2$.

We compute the singular value decomposition $A = U\Sigma V$ (the thin version). Then $\rho = \sum \sigma_j^2$. If $\tilde{\rho} = \sigma_1^2 + \sigma_2^2$ is large (e.g. $\approx 0.95\rho$) then the rank 2 approximation

$$A_{(2)} = \left[u^{(1)}, u^{(2)}\right] \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} \begin{bmatrix} v^{(1)H} \\ v^{(2)H} \end{bmatrix} = \left[\tilde{x}^{(1)} - \mu, \ldots, \tilde{x}^{(n)} - \mu\right] = \left[u^{(1)}, u^{(2)}\right] \left[c^{(1)}, \ldots, c^{(n)}\right]$$

captures most of the variation in the data points, and we can look at the coefficients $c^{(1)}, \ldots, c^{(n)} \in \mathbb{R}^2$

$$\left[c^{(1)}, \ldots, c^{(n)}\right] = \begin{bmatrix} u^{(1)H} \\ u^{(2)H} \end{bmatrix} A = \begin{bmatrix} \sigma_1 v^{(1)H} \\ \sigma_2 v^{(2)H} \end{bmatrix}$$

to understand the original data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^m$.

Note: In many applications the m components of each data point $x^{(j)} \in \mathbb{R}^m$ correspond to different physical quantities like length, weight, age. In this case changing the units (e.g. centimeters instead of meters) changes the result of the SVD, since we use the errors $\left\|x^{(j)} - \tilde{x}^{(j)}\right\|^2 = \sum_{k=1}^m \left|x_k^{(j)} - \tilde{x}_k^{(j)}\right|^2$. Recommendation: **First** rescale each component so that "relevant changes" have the same size for each component. **Then** apply the PCA to the rescaled data vectors.