Machine learning and data mining 2

Matrix decompositions

Radim Bělohlávek

DEPARTMENT OF COMPUTER SCIENCE
PALACKÝ UNIVERSITY, OLOMOUC

Literature

- D. Skillicorn, Understanding Complex Datasets. Data Mining with Matrix Decompositions. Chapman & Hall/CRC, Boca Raton, FL, 2007.
- L. Eldén: Matrix Methods in Data Mining and Pattern Recognition. SIAM, 2007.
- M. Hladík: Lineární algebra (nejen) pro informatiky. MatfyzPress, Praha, 2019.
- L. Barto, J. Tůma: Lineární algebra. Skriptum MFF UK, `https://www2.karlin.mff.cuni.cz/~barto/LinAlg/skripta_la6.pdf`.

Why study matrix decompositions?

- Various benefits (see below)
- Our work in decompositions of Boolean matrices and matrices with degrees (see additional material)

Preliminaries in linear algebra

a modern source for computer science (in Czech) = textbook by Hladík

Matrices

- matrix = rectangular scheme of real numbers
- $A \in \mathbb{R}^{m \times n} \sim$ matrix A with m rows and n columns
- $A = \begin{pmatrix} 1 & 2 & 3 & 3 \\ 1 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ is a matrix in $\mathbb{R}^{3 \times 4}$
- A_{ij} ... element at row i and column j (sometimes a_{ij})
- $A_{i*} \sim$ row i of A
- $A_{*j} \sim$ column j of A
- square matrix: $m = n$

special matrices

- 0 or $0_{m \times n} \sim$ zero matrix, i.e. all entries $= 0$
- I or $I_n \sim$ identity matrix, i.e. square matrix with $I_{ii} = 1$ and $I_{ij} = 0$ for $i \neq j$
- diagonal matrix $\sim$ square matrix with $A_{ij} = 0$ for $i \neq j$
- upper triangular matrix $\sim A_{ij} = 0$ for $i > j$
- symmetric matrix $\sim$ square matrix with $A_{ij} = A_{ji}$

relations and operations

- equality: $A = B$ iff same dimension and $A_{ij} = B_{ij}$ for all i, j
- addition of $A, B \in \mathbb{R}^{m \times n}$:

$$(A + B)_{ij} = A_{ij} + Bij$$

- multiplication of $A \in \mathbb{R}^{m \times n}$ by $c \in \mathbb{R}$:

$$(c \cdot A)_{ij} = c \cdot A_{ij}$$

- multiplication of matrices $A \in \mathbb{R}^{m \times p}$ and $B \in \mathbb{R}^{p \times n}$:

$$(A \cdot B)_{ij} = \sum_{k=1}^{p} A_{ik} \cdot B_{kj}$$

note: two views of matrix multiplication (example)

properties of addition and scalar multiplication

$$
\begin{aligned}
A + B &= B + A \\
(A + B) + C &= A + (B + C) \\
A + 0 &= 0 \\
A + (-1)A &= 0, \\
1A &= A, \\
a(A + B) &= aA + AB, \\
(a + b)A &= aA + bA
\end{aligned}
$$

properties of matrix multiplication

$$\begin{aligned}
(AB)C &= A(BC), \\
A(B+C) &= AB + AC, \\
(A+B)C &= AC + BC, \\
a(AB) &= (aA)B = A(aB), \\
A0 &= 0A = 0, \\
I_m A &= AI_n = A.
\end{aligned}$$

In general, $AB = BA$ does not hold.

matrix transposition of $A \in \mathbb{R}^{n \times m}$ is $A^T \in \mathbb{R}^{m \times n}$ defined by

$$(A^T)_{ij} = A_{ji}.$$

properties

$$\begin{aligned}
(A^T)^T &= A, \\
(A + B)^T &= A^T + B^T, \\
(aA)^T &= aA^T, \\
(AB)^T &= B^T A^T.
\end{aligned}$$

real vectors

- row vector: matrix $A \in \mathbb{R}^{1 \times n}$
- column vector: matrix $A \in \mathbb{R}^{m \times 1}$
- default (mostly): vector $=$ column vector and $\mathbb{R}^n$ stands for $\mathbb{R}^{n \times 1}$

further concepts in matrices

- rank of $A \sim$ max number of linearly independent columns (equivalently, rows) $\mathrm{rank}(A)$
- each $A \in \mathbb{R}^{m \times n}$ induces a mapping

$$x \in \mathbb{R}^n \mapsto Ax \in \mathbb{R}^m$$

linearity, composition represented by matrix multiplication

- $A \in \mathbb{R}^{n \times n}$ is regular $\sim Ax = 0$ has a unique solution

  otherwise A is singular
- $A \in \mathbb{R}^{n \times n}$ is regular iff $\mathrm{rank}(A) = n$
- if A and B are regular, then AB is regular

  if A or B is singular, then AB is singular
- given $A \in \mathbb{R}^{n \times n}$, $A^{-1} \in \mathbb{R}^{n \times n}$ is called the inverse of A if

$$AA^{-1} = A^{-1}A = I_n.$$

- each regular A has an inverse, which is unique

  if A has an inverse, then it is regular

  if A is regular then A^T is regular

  for $A, B \in \mathbb{R}^{n \times n}$, if $AB = I_n$ then both A and B are regular and mutually inverse

- for regular $A, B \in \mathbb{R}^{n \times n}$:

$$\begin{aligned} (A^{-1})^{-1} &= A, \\ (A^T)^{-1} &= (A^{-1})^T, \\ (aA)^{-1} &= \frac{1}{a} A^{-1} \text{ for } a \neq 0, \\ (AB)^{-1} &= B^{-1} A^{-1} \end{aligned}$$

- if $A \in R^{n \times n}$ is regular, then $Ax = b$ has a unique solution $x = A^{-1} b$

Vector spaces

- vector space over a field T with $+$, $\cdot$, and with 0 and 1 (identities for $+$ and $\cdot$) is a structure consisting of a set V (vectors) and operations $+ : V \times V \to V$ (vector addition) and $\cdot : T \times V \to V$ (scalar multiplication) satisfying

$$\begin{aligned} \langle V, + \rangle & \quad \text{is a commutative group,} \\ a(bv) &= (ab)v, \\ 1v &= v, \\ (a+b)v &= av + bv, \\ a(u+v) &= au + av. \end{aligned}$$

examples of vector spaces

- matrices of $m \times n$: $T = \mathbb{R}$, $V = \mathbb{R}^{m \times n}$,
- functions: $T = \mathbb{R}$, $V = \{f \mid f : \mathbb{R} \to \mathbb{R}\}$

basic properties

$$\begin{aligned} 0v &= o, \\ ao &= o, \\ av = o \text{ implies } & a = 0 \text{ or } v = o, \\ (-1)v &= -v. \end{aligned}$$

- subspace $\sim$ contains o, closed under addition and scalar multiplication
- intersection of subspaces is a subspace
- concept of subspace generated by a set W: $\mathrm{span}(W)$
- W generates space V if $\mathrm{span}(W) = V$
- linear combination of vectors $v_1, \ldots, v_n$: any vector $\sum_{i=1}^{n} a_i v_i$
- $\mathrm{span}(W) =$ the set of all linear combinations of vectors in W
- linear independence of a set of vectors:

$$\sum_{i=1}^{n} a_i v_i = o \text{ implies } a_1 = \cdots = a_n = 0$$

linearly dependent = not l. independent

independence of infinite set: each finite set independent

- $v_1, \ldots, v_n$ are lin. dependent iff $\exists k$ s.t. $v_k \in \operatorname{span}(v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n)$
- base (basis) of a vector space V: subset of V that is generates V and is linearly independent
- example: the set $\{e_i \mid i = 1, \ldots, n\}$ is a base of $\mathbb{R}^n$
  (e_i is a vector of 0s with 1 at position i)
- uniqueness of coordinates w.r.t. base (basis a linearly ordered set of vectors):
  $\{v_1, \ldots, v_n\}$ basis, then for each $v \in V$ there exist unique a_i s.t. $v = \sum_{i=1}^{n} a_i v_i$
- each vector space has a base (finite or infinite)
- all bases of a finitely generated vector space have the same size

- dimension $\dim V$ of a finitely generated vector space V = size of its base (of any of its bases)
- $\dim \mathbb{R}^n = n$

  $\dim \mathbb{R}^{m \times n} = mn$

  $\dim \{o\} = 0$

  $\dim \mathcal{F} = \infty$ (space of real functions, not finitely generated)
- $v_1, \ldots, v_n$ linearly independent then $n \leq \dim V$

  $v_1, \ldots, v_n$ generate V then $n \geq \dim V$
- each linearly independent set may be extended to a base

- if W is a subspace of V, then $\dim W \leq \dim V$

  if, moreover, $\dim W = \dim V$ then $W = V$

- let U, V be subspaces of W; the sum $U + V$ is

$$U + V = \operatorname{span}(U \cup V)$$

  $U + V$ is the least subspace containing both U and V and it holds:
  $U + V = \{u + v \mid u \in U, v \in V\}$

- examples:

  $\mathbb{R}^2 = \operatorname{span}\{\langle 0, 1\rangle\} + \operatorname{span}\{\langle 2, 0\rangle\}$

  $\mathbb{R}^2 = \operatorname{span}\{\langle 0, 1\rangle\} + \operatorname{span}\{\langle 2, 1\rangle\}$

  $\mathbb{R}^3 = \operatorname{span}\{\langle 0, 0, 1\rangle, \langle 0, 1, 0\rangle\} + \operatorname{span}\{\langle 1, 0, 0\rangle\}$

- $\dim(U) + \dim(V) = \dim(U + V) + \dim(U \cap V)$
- if $U \cap V = \{o\}$ then $U + V$ is called the direct sum of U and V
  then $\dim(U) + \dim(V) = \dim(U + V)$

Matrix spaces

matrix spaces = vector spaces where vectors are matrices

- for matrix $A \in \mathbb{R}^{m \times n}$:

  $\mathcal{C}(A) = \mathrm{span}\{A_{*1}, \ldots, A_{*n}\}$ (space generated by columns of A)

  $\mathcal{R}(A) = \mathrm{span}\{A^T_{1*}, \ldots, A^T_{m*}\}$ (space generated by rows of A)

  equivalently, $\mathcal{R}(A) = \mathcal{C}(A^T)$

  $\ker(A) = \{x \in \mathbb{R}^n \mid Ax = o\}$ (kernel of A)

- $\mathcal{C}(A)$ and $\mathcal{R}(A)$ are also called the column and row space of A

  due to the properties of span:

  $\mathcal{C}(A) = \{Ax \mid x \in \mathbb{R}^n\}$ (indeed, Ax is a linear combination of columns of A)

  $\mathcal{R}(A) = \{A^T x \mid x \in \mathbb{R}^m\}$

  $\ker(A)$ is a subspace of $\mathbb{R}^n$

matrix A as a linear mapping

- each matrix $A \in \mathbb{R}^{m \times n}$ represents a mapping of $\mathbb{R}^n$ to $\mathbb{R}^m$ defined by

$$x \mapsto Ax$$

- this mapping is conveniently identified with A
- the mapping is linear in that for each $x, y \in \mathbb{R}^n$ and $a \in \mathbb{R}$:

  $A(x + y) = Ax + Ay$

  $A(ax) = a(Ax)$

- each subspace of $\mathbb{R}^n$ is a column (row) space of some matrix:

  for each subspace $V \subseteq \mathrm{R}^n$ there is $A \in \mathrm{R}^{n \times m}$ such that $V = \mathcal{C}(A)$

  for each subspace $V \subseteq \mathrm{R}^n$ there is $A \in \mathrm{R}^{m \times n}$ such that $V = \mathcal{R}(A)$

  for each subspace $V \subseteq \mathrm{R}^n$ there is $A \in \mathrm{R}^{m \times n}$ such that $V = \ker(A)$

fundamental properties of dimension:

- $\operatorname{rank}(A) = \dim \mathcal{C}(A) = \dim \mathcal{R}(A)$
- $\operatorname{rank}(A) = \operatorname{rank}(A^T)$
- $\dim \ker(A) + \operatorname{rank}(A) = n$

further useful properties:

- for any $Q \in \mathbb{R}^{p \times m}$ and $A \in \mathbb{R}^{m \times n}$

$$\mathcal{R}(QA) \text{ is a subspace of } \mathcal{R}(A)$$

- for a regular $Q \in \mathbb{R}^{p \times m}$ and any $A \in \mathbb{R}^{m \times n}$

$$\mathcal{R}(QA) = \mathcal{R}(A)$$

Scalar product, norm, and related

standard scalar product and norm:

- scalar product of $x, y \in \mathbb{R}^n$:

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$$

- example: $\langle \langle 1, 2, 3 \rangle, \langle 1, -1, 2 \rangle \rangle = 1 \cdot 1 - 2 \cdot 1 + 3 \cdot 2 = 5.$
- norm of $x \in \mathbb{R}^n$:

$$||x|| = \sqrt{\langle x, x \rangle}$$

- example: $||\langle 3, 0, 4 \rangle|| = \sqrt{3^2 + 0^2 + 4^2} = \sqrt{25} = 5.$
- if φ is the angle between x and y, we have:

$$\langle x, y \rangle = ||x|| \cdot ||y|| \cdot \cos(\varphi)$$

in particular, x and y are orthogonal (i.e. geometrically perpendicular), iff

$$\langle x, y \rangle = 0$$

general notions in a vector space V:

- scalar product in V is a binary operation $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ satisfying

  $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ iff $x = o$
  $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
  $\langle ax, y \rangle = a\langle x, y \rangle$
  $\langle x, y \rangle = \langle y, x \rangle$

- standard scalar product is a scalar product in the above general sense
- norm in V is a mapping $|| \cdot || : V \to \mathbb{R}$ satisfying

  $||x|| \geq 0$, and $||x|| = 0$ iff $x = o$
  $||ax|| = a||x||$ for each $x \in V$ and $a \in \mathbb{R}$
  $||x + y|| \leq ||x|| + ||y||$

- on a vector space V with a scalar product, vectors $x, y \in V$ are orthogonal (perpendicular) if
$$\langle x, y \rangle = 0.$$
- Pythagoras T: if x, y are orthogonal, then
$$||x + y||^2 = ||x||^2 + ||y||^2$$
- angle between vectors $x, y \in V$ in a space with scalar product is defined as $\alpha \in [0, 2\pi]$ such that:
$$\cos(\alpha) = \frac{\langle x, y \rangle}{||x|| \cdot ||y||}$$

general notions in a vector space V:

- Theorem: For any scalar product $\langle \cdot, \cdot \rangle$ on V, the mapping

$$||x|| = \sqrt{\langle x, x \rangle}$$

  is a norm (norm induced by scalar product).

- Hence the standard norm is a norm in the general sense.

- p-norm for $p = 1, 2, \ldots$:

$$||x||_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|}$$

  $p = 2$: Euclidean norm (standard)

  $p = 1$: Manhattan norm

  $p = \infty$ (using limit): maximum (Chebyshev) norm $||x||_\infty = \max_i |x_i|$

Digression: Cosine similarity and information retrieval

- problem: given a collection $\{d_i \mid i \in I\}$ of documents d_i and a document q (q may be a regular document or a user query), find document d_i most similar to q
- idea: each document d is regarded as a vector $x(d)$ in a vector space in which coordinates correspond to keywords (or all words, or cleaned words) which may appear in documents
- for keyword w, the coordinate $x(d)_w$ is defined as

$$x(d)_w = \text{ number representing significance of } w \text{ in } d$$

example: $x(d)_w$=frequency of work w in d

- more sophisticated definitions of vectors $x(d)$:

  e.g. tf-idf = term-frequency–inverse document frequency

$$\begin{aligned} \mathrm{tf} &= \frac{\text{no. occurences of } w \text{ in } d}{\text{no. words in } d} \\ \mathrm{idf} &= \log \frac{\text{no. documents in the database}}{\text{no. documents containing word } w} \\ x(d)_w &= \mathrm{tf} \cdot \mathrm{idf} \end{aligned}$$

- given a query document q, one looks for document d for which

$$\cos(\alpha) = \frac{\langle q, d \rangle}{||q|| \cdot ||d||} \text{ is maximal}$$

- note: not $||q - d||$ minimal, but rather smallest angle α between q and d
- example:

end of digression

Orthonormal base and projection

- a system $\{x_1, \ldots, x_n\} \subseteq V$ is orthogonal if $\langle x_i, x_j \rangle = 0$ for $i \neq j$

  if, moreover, $||x_i|| = 1$ for each i, the system is orthonormal
- obviously, if $\{x_1, \ldots, x_n\}$ is orthogonal, then $\{\frac{1}{||x_1||}x_1, \ldots, \frac{1}{||x_n||}x_n\}$ is orthonormal
- example in R^2:

  $\{\langle 0, 1\rangle, \langle 1, 0\rangle\}$ is orthonormal

  $\{\frac{\sqrt{2}}{2}\langle 1, 1\rangle, \frac{\sqrt{2}}{2}\langle -1, 1\rangle\}$ is orthonormal
- Theorem (Fourier coefficients). If $\{v_1, \ldots, v_n\}$ is an orthonormal base, then for each $x \in V$:

$$x = \sum_{i=1}^{n} \langle x, v_i \rangle v_i$$

- example: express using the theorem the vector $\langle 2, 3\rangle$ in the two above bases
- Theorem (Gram-Schmidt orthogonalization). Each finitely-dimensional vector space with a scalar product has an orthonormal base.

- The orthogonal complement of $M \subseteq V$ is the set

$$M^\perp = \{v \in V \mid \text{ for each } x \in M : \langle x, v \rangle = 0\},$$

  i.e. $M^\perp$ consists of all vectors perpendicular to each $x \in M$

- $M^\perp$ is a vector subspace of V (easy to check)

  $M \subseteq N$ implies $M^\perp \supseteq N^\perp$

  $M^\perp = (\mathrm{span}(M)^\perp)$

- examples:

  $\{\langle 1, -3 \rangle\}^\perp$ is the space generated by $\langle 3, 1 \rangle$, i.e. $\{\langle 1, -3 \rangle\}^\perp = \mathrm{span}\{\langle 3, 1 \rangle\}$

  $\{\langle 1, 0, 0 \rangle\}^\perp = \mathrm{span}\{\langle 0, 1, 0 \rangle, \langle 0, 0, 1 \rangle\}$

- Definition. A projection of $x \in V$ to a subspace $U \subseteq V$ is a vector $x_U \in U$ such that

$$||x - x_U|| = \min_{y \in U} ||x - y||.$$

- Theorem. Let U be a subspace of a finitely-dimensional V.
  (a) If for $x \in V$ and $y \in U$ we have $x - y \in U^{\perp}$, then $||x - y|| < ||x - z||$ for each $z \in U - \{y\}$, i.e. y is a unique projection of x to U.
  (b) If $\{v_1, \ldots, v_m\}$ is an orthonormal base of U, then

$$x_U = \sum_{i=1}^{m} \langle x, v_i \rangle v_i$$

  is the unique projection of x to U.

- Proof.

  (a) Since $y, z \in U$, we have $y - z \in U$, hence $(x - y) \perp (y - z)$.

  Hence due to Pythagoras t.,

  $$||x - z||^2 = ||x - y||^2 + ||y - z||^2 \geq ||x - y||^2,$$

  with equality iff $||y - z||^2$, i.e. iff $y = z$.

  (b) Let $\{v_1, \ldots, v_m, v_{m+1}, \ldots, v_n\}$ be an extension to an orthonormal base of V. We check that $y = \sum_{i=1}^{m} \langle x, v_i \rangle v_i$ satisfies $x - y \in U^\perp$:

  $$x - y = \sum_{i=1}^{n} \langle x, v_i \rangle v_i - \sum_{i=1}^{m} \langle x, v_i \rangle v_i = \sum_{i=m+1}^{n} \langle x, v_i \rangle v_i \in U^\perp.$$

  The proof is finished due to (a).

Examples of projection.

- Projection of $x = \langle 1, 2, 4, 5 \rangle^T$ to a subspace U generated by $x_1 = \langle 1, 0, 1, 0 \rangle^T$, $x_2 = \langle 1, 1, 1, 1 \rangle^T$, and $x_3 = \langle 1, 0, 0, 1 \rangle^T$:

  One may obtain an orthonormal basis (we omit details):

  $$v_1 = \frac{\sqrt{2}}{2}\langle 1, 0, 1, 0 \rangle^T, v_2 = \frac{\sqrt{2}}{2}\langle 0, 1, 0, 1 \rangle^T, v_3 = \frac{1}{2}\langle 1, -1, -1, 1 \rangle^T.$$

  According to the theorem,

  $$x_U = \sum_{i=1}^{3} \langle x, v_i \rangle v_i = \frac{1}{2}\langle 5, 7, 5, 7 \rangle$$

- Projection on a line (essential for QR decomposition):

  Let $o \neq a \in \mathbb{R}^n$ (line). Projection of $x \in \mathbb{R}^n$ to the space U generated by a, i.e. $U = \text{span}\{a\}$.

  Clearly,

  $$\frac{1}{||a||}a$$

  provides an orthonormal basis of U. According to the theorem,

  $$x_U = \langle x, v_1 \rangle v_1 = \langle x, \frac{1}{||a||}a \rangle \frac{1}{||a||}a = \frac{1}{||a||^2}\langle x, a \rangle a = \frac{x^T a}{a^T a}a.$$

- in particular, consider $x = \langle 2, 1 \rangle$ and $a = \langle 1, 0 \rangle$; then

  $$x_U = \frac{x^T a}{a^T a}a = \frac{\langle \langle 2, 1 \rangle, \langle 1, 0 \rangle \rangle}{\langle \langle 1, 0 \rangle, \langle 1, 0 \rangle \rangle}\langle 1, 0 \rangle = \langle 2, 0 \rangle.$$

- now, consider $x = \langle 2, 0 \rangle$ and $a = \langle 1, 1 \rangle$; then

  $$x_U = \frac{x^T a}{a^T a}a = \frac{\langle \langle 2, 0 \rangle, \langle 1, 1 \rangle \rangle}{\langle \langle 1, 1 \rangle, \langle 1, 1 \rangle \rangle}\langle 1, 1 \rangle = \langle 1, 1 \rangle.$$

Matrix decompositions

- We start with QR and SVD decompositions

  = selected fundamental decompositions from linear algebra

- B. A. Cipra,"The best of the 20th century: Editors name top 10 algorithms." SIAM News, 33:1–2, 2000:

  matrix decomposition algorithms (including QR and SVD) are among 10 best algorithms

  (others: Quicksort, fast Fourier transform, Monte Carlo, Fortran compiler, ...)

- QR and SVD

  useful both from several pure and applied linear algebra viewpoint

  including machine learning and data mining applications

- decomposition

$$A = QR$$

of an arbitrary matrix $A \in \mathbb{R}^{m \times n}$ into an orthoginal matrix $Q \in \mathbb{R}^{m \times m}$ and an upper triangular matrix $R \in \mathbb{R}^{m \times n}$ with nonnegative main diagonal

- example for $m = n = 3$:

$$\begin{pmatrix} 0 & -20 & -14 \\ 3 & 27 & -4 \\ 4 & 11 & -2 \end{pmatrix} = \begin{pmatrix} 0 & -20/25 & -15/25 \\ 15/25 & 12/25 & -16/25 \\ 20/25 & -9/25 & 12/25 \end{pmatrix} \cdot \begin{pmatrix} 5 & 25 & -4 \\ 0 & 25 & 10 \\ 0 & 0 & 10 \end{pmatrix}$$

- $QR =$ useful form providing a better view of A
- provide basic results, application to the least squares problem, overview of further applications

Orthogonal matrices

- Definition. A matrix $Q \in \mathbb{R}^{n \times n}$ is orthogonal if

$$Q^T Q = I_n.$$

- Example:

$$Q = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Then

$$Q^T Q = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- basic concept in linear algebra, many useful properties, utilized in both QR and SVD

Theorem (equivalent conditions)

The following conditions are equivalent for any $Q \in \mathbb{R}^{n \times n}$:

(a) *Q is orthogonal.*

(b) *Q is regular and $Q^{-1} = Q^T$.*

(c) *$QQ^T = I_n$.*

(d) *Q^T is orthogonal.*

(e) *An inverse Q^{-1} of Q exists and is orthogonal.*

(f) *Rows of Q form an orthonormal basis of $\mathbb{R}^n$.*

(g) *Columns of Q form an orthonormal basis of $\mathbb{R}^n$.*

Proof.

(a) $\Rightarrow$ (b): Recall first the following claim from linear algebra:

(C) If AB is regular for $A, B \in \mathbb{R}^{n\times n}$, then A and B are regular.

Now, if $Q^TQ = I_n$, then due to (C), Q is regular and thus has an inverse Q^{-1}, and

$$Q^{-1} = I_nQ^{-1} = (Q^TQ)Q^{-1} = Q^T(QQ^{-1}) = Q^TI_n = Q^T.$$

(b) $\Rightarrow$ (a) follows from the definition of inverse.

(b) $\Leftrightarrow$ (c): similar to (a) $\Leftrightarrow$ (b).

(a) $\Rightarrow$ (d): (a) implies (c), i.e. $(Q^T)^TQ^T = QQ^T = I_n$, i.e. Q^T is orthogonal.
(d) $\Rightarrow$ (a): This is just the proved claim (a) $\Rightarrow$ (d) applied to Q^T.

Proof.

(a) $\Rightarrow$ (e): (a) implies (b), hence Q^{-1} exists and $Q^{-1} = Q^T$. Thus $(Q^{-1})^T Q^{-1} = (Q^T)^T Q^T = QQ^T = I_n$ due to (c).

(e) $\Rightarrow$ (a): (e) means $(Q^{-1})^T Q^{-1} = I_n$ hence, due to $(A^{-1})^T = (A^T)^{-1}$, $(Q^T)^{-1} Q^{-1} = I_n$. Applying the operation of inverse and using $(AB)^{-1} = B^{-1}A^{-1}$, we get for the left hand side $((Q^T)^{-1}Q^{-1})^{-1} = (Q^{-1})^{-1}((Q^T)^{-1})^{-1} = QQ^T$, and for the right hand side $I_n^{-1} = I_n$, proving $QQ^T = I_n$, which is equivalent to (a) due to the proved equivalences.

(a) $\Rightarrow$ (f): $Q^T Q = I_n$ tells that the scalar product $\langle Q_{i*}, Q_{j*} \rangle = 1$ for $i = j$ and $\langle Q_{i*}, Q_{j*} \rangle = 0$ for $i \neq j$. Hence, the rows form an orthonormal system. Vectors of orthonormal systems are linearly independent, hence the n rows form a base.

(f) $\Rightarrow$ (a): $Q^T Q = I_n$ follows from orthonormality of the rows.

(a) $\Leftrightarrow$ (g): Symmetrically as for (a) $\Leftrightarrow$ (f). □

Theorem (orthogonal matrices)

Let $Q \in \mathbb{R}^{n \times n}$ be orthogonal.

(a) *If R is orthogonal then QR is orthogonal*

(b) $\langle Qx, Qy \rangle = \langle x, y \rangle$ *for each* $x, y \in \mathbb{R}^n$.

(c) $||Qx|| = ||x||$ *for each* $x \in \mathbb{R}^n$.

(d) $|Q_{ij}| \leq 1$ *for each* i, j.

Proof.

(a): $(QR)^T(QR) = R^TQ^TQR = R^TR = I_n$.
(b): $\langle Qx, Qy\rangle = (Qx)^TQy = x^TQ^TQy = x^Ty = \langle x, y\rangle$.
(c): $||Qx|| = \sqrt{\langle Qx, Qx\rangle} = \sqrt{\langle x, x\rangle} = ||x||$.
(d): Since rows of Q form an orthonormal base (see above), we have $||Q_{i*}|| = 1$, hence $1 = ||Q_{i*}||^2 = \sum_{j=1}^{n} Q_{ij}$, from which $Q_{ij} \leq 1$ follows.

□

Examples of orthogonal matrices 1: rotation

- matrix of rotation counter-clockwise by angle φ:

$$\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$

- check orhogonality and check that it is a matrix of rotation (use $\sin(\alpha+\beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$ and $\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$)
- In general, rotation in $\mathbb{R}^n$ in the plane given by axes x_i and x_j is accomplished by the Givens matrix $G_{ij}(c,s)$ for $c^2 + s^2 = 1$:

$$\begin{pmatrix} I & & & & \\ & c & & -s & \\ & & I & & \\ & s & & c & \\ & & & & I \end{pmatrix}$$

Examples of orthogonal matrices 2: Householder matrices

- $o \neq a \in \mathbb{R}^n$, Householder matrix given by $o \neq a \in \mathbb{R}^n$:

$$H(a) = I_n - \frac{2}{a^T a} a a^T$$

- based on projection, let us recall appropriate notions and derive the form of $H(a)$ (in particular, orthogonal projection and projection on the line; see slides above)

Note (without proof):

- Each orthogonal matrix is the product of at most $\binom{n}{2}$ Givens matrices.
- Each orthogonal matrix is the product of at most n Householder matrices.

Lemma

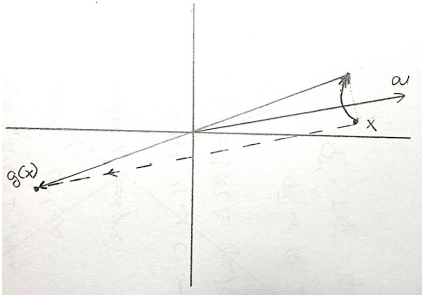
For each $o \neq a \in \mathbb{R}^n$, the Householder matrix $H(a)$ is orthogonal.

Proof.

By checking $H(a)^T H = I_n$. □

Obtaining Householder matrix:

- $H(a)$ is the matrix of the following transformation $g(x)$:



1. rotate x around a (to obtain $r(x)$)
2. then transpose $r(x)$ around the origin $\langle 0, 0 \rangle$

- denote x' the projection of x on the line given by a;

  then, clearly, $r(x) = x + 2(x' - x)$
- now recall: projection x' of $x \in \mathbb{R}^n$ on the line given by $a \in \mathbb{R}^n$ (i.e. on $U = \text{span}\{a\}$) is given by (see above)

$$x' = \frac{x^T a}{a^T a} a, \text{ which is rewritten as } x' = \frac{a^T x}{a^T a} a = \frac{a(a^T x)}{a^T a} = \frac{(aa^T)x}{a^T a} = \frac{aa^T}{a^T a} x$$

- hence, $r(x) = x + 2(\frac{aa^T}{a^T a}x - x) = 2\frac{aa^T}{a^T a}x - x = (2\frac{aa^T}{a^T a} - I_n)x$
- the result $g(x)$, clearly, equals $-r(x)$, i.e.

$$g(x) = (I_n - 2\frac{aa^T}{a^T a})x,$$

  hence $g(x)$ is obtains by transformation by the Householder matrix $H(a)$.

goal: for any $x \neq y \in \mathbb{R}^n$ with $||x|| = ||y||$ find a linear transform mapping x to y ($||\cdot||$ stands for Euclidean norm)

crucial property: Householder matrices may be used

Theorem

For each $x \neq y \in \mathbb{R}^n$ with $||x|| = ||y||$ and $x - y$ one has

$$[H(x-y)]x = y.$$

Proof.

$$
\begin{aligned}
[H(x-y)]x &= \left(I_n - \frac{2}{(x-y)^T(x-y)}(x-y)(x-y)^T\right)x \\
&= x - \frac{2(x-y)^T x}{(x-y)^T(x-y)}(x-y) = x - \frac{2\|x\|^2 - 2y^T x}{(x-y)^T(x-y)}(x-y) \\
&= x - \frac{\|x\|^2 + \|y\|^2 - 2y^T x}{\|x-y\|^2}(x-y) = x - \frac{\|x-y\|^2}{\|x-y\|^2}(x-y) \\
&= x - (x-y) = y.
\end{aligned}
$$

□

Corollary (Householder projection of x onto e_1)

For each $x \in \mathbb{R}^n$ *let*

$$H = \begin{cases} H(x - \|x\|e_1) & \text{if } x \neq \|x\|e_1 \\ I_n & \text{if } x = \|x\|e_1. \end{cases}$$

Then $Hx = \|x\|e_1$.

Example. Let $x = \langle 2, 2, 1\rangle$. Then $\|x\| = 3$, $x \neq \|x\|e_1 = 3\langle 1, 0, 0\rangle$, hence $x - \|x\|e_1 = \langle 2, 2, 1\rangle - \langle 3, 0, 0\rangle = \langle -1, 2, 1\rangle$

$$
\begin{aligned}
H &= H(x - 3e_1) = I_n - \frac{2}{(x - 3e_1)^T (x - 3e_1)} (x - 3e_1)(x - 3e_1)^T \\
&= I_3 - \frac{2}{\langle -1, 2, 1\rangle^T \langle -1, 2, 1\rangle} \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} -1 & 2 & 1 \end{pmatrix} \\
&= I_3 - \frac{2}{6} \begin{pmatrix} 1 & -2 & -1 \\ -2 & 4 & 2 \\ -1 & 2 & 1 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 2 & 2 & 1 \\ 2 & -1 & -2 \\ 1 & -2 & 2 \end{pmatrix}
\end{aligned}
$$

Two examples of easy solutions to substantial problems in linear algebra by Householder transformation:

- Alternative to elimination (in solution of a system of linear equations).

  Let $A \in \mathbb{R}^{[m \times n]}$ (and e.g. consider $Ax = b$).

  Consider Househ. m. H which transforms the first column of A to e_1 (Corollary). Then HA contains in its first column 0s everywhere except $(HA)_{11}$ (H "nullifies" the first column except for its first element). Recursive aplication of this step enables to transform A to a REF (row echelon form) Details easy; omitted.

- Let $u \in \mathbb{R}^n$, $\|u\| = 1$. Construct an orthonormal base containing u (i.e. complete u to an orthonormal base).

  Reformulation: Construct an orthonormal matrix Q with its first column equal to u.

  Solution: If $u = e_1$, put $Q = I_n$. If not, put $Q = H(e_1 - u)$. Then the first column of Q is indeed u, as due to the Theorem,

$$Q_{*1} = Qe_1 = H(e_1 - u)e_1 = u.$$

QR decomposition

Now:

Theorem (existence of QR decomposition)

For each $A \in \mathbb{R}^{m \times n}$ there exists an orthogonal $Q \in \mathbb{R}^{m \times m}$ and an upper triangular $R \in \mathbb{R}^{m \times n}$ with non-negative elements on the main diagonal such that

$$A = QR.$$

(recall block multiplication of matrices)

Proof.

By induction on n (number of columns).
$n = 1$: Then $A \in \mathbb{R}^{m \times 1}$, i.e. A is a vector $a \in \mathbb{R}^m$. For the matrix $H \in \mathbb{R}^{m \times m}$ obtained as in Corollary we have $HA = Ha = \|a\|e_1$. Since H is orthogonal, H^T is its inverse, whence

$$A = I_m A = (H^T H)a = H^T(Ha) = H^T\|a\|e_1,$$

which shows that $Q = H^T$ and $R = \|a\|e_1$ represent the required decomposition.

Induction step from $n-1$ to n:
Apply Corollary to the first column of $A \in \mathbb{R}^{m \times n}$ to obtain $H \in \mathbb{R}^{m \times m}$. Then $HA_{*1} = \|A_{*1}\|e_1$. Hence HA has the following form:

$$HA = \begin{pmatrix} \|A_{*1}\| & b^T \\ o & B \end{pmatrix}$$

where $B \in \mathbb{R}^{(m-1)\times(n-1)}$, $b^T \in \mathbb{R}^{1\times(n-1)}$.

□

cntd.

Due to the assumption for $n-1$ there exist an orthogonal $Q' \in \mathbb{R}^{(m-1)\times(m-1)}$ and upper triangular $R' \in \mathbb{R}^{(m-1)\times(n-1)}$ with a non-negative main diagonal such that $B = Q'R'$.

From the above equality for HA and since $(Q')^T B = R'$, we get

$$\begin{pmatrix} 1 & o^T \\ o & Q'^T \end{pmatrix} HA = \begin{pmatrix} 1 & o^T \\ o & Q'^T \end{pmatrix} \begin{pmatrix} \|A_{*1}\| & b^T \\ o & B \end{pmatrix} = \begin{pmatrix} \|A_{*1}\| & b^T \\ o & R' \end{pmatrix}$$

Putting

$$Q = H^T \begin{pmatrix} 1 & o^T \\ o & Q' \end{pmatrix} \text{ and } R = \begin{pmatrix} \|A_{*1}\|e_1 & b^T \\ o & R' \end{pmatrix}$$

we easily check that Q is orthogonal and R upper triangular with a non-negative main diagonal. The last but one displayed equality claims $Q^T A = R$, i.e. $A = QR$. □

Example of computing QR decomposition (recursive descent as in the proof).
Let

$$A = \begin{pmatrix} 0 & -20 & -14 \\ 3 & 27 & -4 \\ 4 & 11 & -2 \end{pmatrix}$$

$n > 1$ hence by the induction step we construct H by application of Corollary to A_{*1}.
We have $A_{*1} - \|A_{*1}\|e_1 = \langle -5, 3, 4 \rangle^T$. Thus for $x = \langle -5, 3, 4 \rangle^T$ and $H = H(x)$,

$$H = I_3 - \frac{2}{x^T x} x x^T = I_3 - \frac{2}{50} \begin{pmatrix} 25 & -15 & -20 \\ -15 & 9 & 12 \\ -20 & 12 & 16 \end{pmatrix} = \frac{1}{25} \begin{pmatrix} 0 & 15 & 20 \\ 15 & 16 & -12 \\ 20 & -12 & 9 \end{pmatrix}$$

Compute HA and obtain from HA the $(m-1) \times (n-1) = 2 \times 2$ matrix B.
Obtain decomposition $B = Q'R'$ by the same procedure.
Obtain Q and R from H, Q' and R' by the formulas in the induction step of the proof.

Result:

$$Q = \frac{1}{25}\begin{pmatrix} 0 & -20 & -15 \\ 15 & 12 & -16 \\ 20 & -9 & 12 \end{pmatrix} \text{ and } R = \begin{pmatrix} 5 & 25 & -4 \\ 0 & 25 & 10 \\ 0 & 0 & 10 \end{pmatrix}$$

The same is obtained by the next algorithm, which results from the proof. ($R(j:m,j)$ denotes the vector $\langle r_{jj}, \ldots, r_{mj} \rangle^T$):

input: $A \in \mathbb{R}^{m \times n}$
output: $Q \in \mathbb{R}^{m \times m}$ and $R \in \mathbb{R}^{m \times n}$ providing a QR decomposition of A

1. $Q \leftarrow I_m$, $R \leftarrow A$
2. for $j \leftarrow 1$ to $\min(m, n)$ do
3. $\quad x \leftarrow R(j : m, j)$
4. $\quad$ if $x \neq \|x\| e_1$ then
5. $\quad\quad x \leftarrow x - \|x\| e_1$
6. $\quad\quad H(x) \leftarrow I_{m-j+1} - \frac{2}{x^T x} x x^T$
7. $\quad\quad H \leftarrow \begin{pmatrix} I_{j-1} & o \\ o & H(x) \end{pmatrix}$
8. $\quad\quad R \leftarrow HR$, $Q \leftarrow QH$

Theorem

Each regular matrix A has a unique QR decomposition $A = QR$ with the main diagonal of R consisting of positive reals.

- In general, QR decomposition is not unique (E.g. for $A = o$ and $R = o$ we have $A = QR$ for any orthogonal matrix Q).

Proof.

Regularity of $A \in \mathbb{R}^{n \times n}$ and $A = QR$ implies regularity of R, hence $R_{ii} \neq 0$, i.e. the diagonal of R is positive.
Let $A = Q_1 R_2$ and $A = Q_2 R_2$ be QR decompositions. We prove $Q_1 = Q_2$ and $R_1 = R_2$.

$Q_1 R_1 = Q_2 R_2$ yields $Q_2^T Q_1 = R_2 R_1^{-1}$. This matrix, denote it U, is orthogonal as it is the product $Q_2^T Q_1$ of orthogonal matrices, and is also upper triangular as it is the product $R_2 R_1^{-1}$ of upper triangular matrices (note: R_1^{-1} is upper triangular).

As $U = R_2 R_1^{-1}$, the first column of U satisfies $U_{*1} = \langle U_{11}, 0, \ldots, 0 \rangle^T$, hence $\|U_{*1}\| = 1$ yields $U_{11} = 0$, i.e. $U_{*1} = e_1$.

Since the second column U_{*2} is orthogonal to U_{*1}, one has $U_{12=0}$, and since $\|U_{*2}\| = 1$, we get $U_{22} = 1$, thus $U_{*2} = e_2$.

Repeated application of the above reasoning yields $U_{*i} = e_i$, hence $U = I_n$.

Since $Q_2^T Q_1$, Q_2^T is inverse to Q_1, hence orthogonality of Q_1 implies $Q_2^T = Q_1^T$, whence $Q_2 = Q_1$.

Since $R_2 R_1^{-1} = I_n$, hence $R_1 = R_2$.

□

Reduced QR decomposition.

- Theorem may be generalized to $A \in \mathbb{R}^{m \times n}$ having linearly independent columns (in which case $m \geq n$).
  Then the first n columns of Q are unique and the diagonal of R is positive.
- Let Q have linearly independent columns. Then the QR decomposition $A = QR$ may be written in a block form as:

$$A = QR = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_1 \\ o \end{pmatrix} = Q_1 R_1.$$

  I.e.: $Q_1 \in \mathbb{R}^{m \times n}$, $Q_2 \in \mathbb{R}^{m \times m-n}$, $R_1 \in \mathbb{R}^{n \times n}$.
  This is called a reduced QR decomposition (the full QR decomposition may clearly be reconstructed from the reduced one).

Further ways to compute QR decomposition:

- Using Gram-Schmidt orthogonalization (numeric precision viewpoint not as good as using Householder matrices).
- Using Cholesky decomposition.

Applications of QR decomposition

Least squares as selected application in detail

1. Least squares method

- Consider a system $Ax = b$, $A \in \mathbb{R}^{m \times n}$, of linear equations with no solution.

  Typically: $m > n$ or $m >> n$ (overdetermination)

- Desired is a good approximate solution, i.e. $x \in \mathbb{R}^n$ such that

$$\|Ax - b\| \text{ is minimal.}$$

  With the Euclidean norm $\|\cdot\|$, this requires

$$\|Ax - b\|^2 = \sum_{j=1}^{m} (A_{j*}x - b_j)^2 \text{ be minimal.}$$

  Hence the term "least squares."

Theorem (solutions by least squares)

Let $A \in \mathbb{R}^{m \times n}$. Then the set of approximate solutions of $Ax = b$ in the sense of least squares (i.e. minimizing $\|Ax - b\|$) is nonempty and equals the set of (exact) solutions of

$$A^T Ax = A^T b \text{ (so called normal equations).}$$

- Note: If $Ax = b$ has an exact solution then: x is an exact solution of $Ax = b$ iff x is an approximate solution of $Ax = b$ in the sense of least squares.

Proof.

Finding $x \in \mathbb{R}^n$ for which $\|Ax - b\|$ is minimal means, by definition of projection, finding a projection of b onto the column space $\mathcal{C}(A)$ of A (justify).

According to the above results, Ax is such a projection iff $Ax - b \in \mathcal{C}(A)^\perp = \ker(A^T)$, i.e. iff $A^T(Ax - b) = 0$, which means $A^T Ax = A^T b$.

The last system of equations (normal equations) has a solution because a projection of b always exists.

□

Uniqueness of the solution in the sense of least squares?
When the columns of A are linearly independent. Namely, then $A^T A$ is regular (theorem in linear algebra), hence $A^T Ax = A^T b$ has a unique solution. Therefore:

Corollary

Let the rank of $A \in \mathbb{R}^{m \times n}$ be n. Then

$$x = (A^T A)^{-1} A^T b$$

is a unique approximate solution x of $Ax = b$ in the sense of least squares.

Example, least squares (Hladík).

The evolution of world population is described by

year (α)	1950	1960	1970	1980	1990	2000
population in 10^9 (β)	2.519	2.982	3.692	4.435	5.263	6.070

How to fit a line approximately describing this dependence? An exactly firring line $\beta = \alpha x_1 + x_2$ (i.e. x_1 slope, x_2 offset) would satisfy:

$$\begin{aligned} 1950x_1 + x_2 &= 2.519 \\ &\vdots \\ 2000x_1 + x_2 &= 6.070 \end{aligned}$$

This set of equations may be rewritten as a system of linear equations:

$$Ax = b, \qquad \text{in particular} \qquad \begin{pmatrix} 1950 & 1 \\ 1960 & 1 \\ 1970 & 1 \\ 1980 & 1 \\ 1990 & 1 \\ 2000 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2.519 \\ 2.982 \\ 3.692 \\ 4.435 \\ 5.263 \\ 6.070 \end{pmatrix}$$

No exact solution exists.

Approximate solution in the sense of least squares: $x = (A^T A)^{-1} A^T b$:

$$x_1 = 0.0724, \quad x_2 = -138.84.$$

Enables interpolation, e.g. population in 1995 is cca

$$1995 \cdot 0.0724 - 138.84 = 5.598 \text{ (actual: 5.707)},$$

as well as extrapolation (one needs to be careful) , e.g.

$$2005 \cdot 0.0724 - 138.84 = 6.322 \text{ (actual: 6.512)}.$$

2. Applying QR decomposition to least squares problem

Suppose $A \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(A) = n$ (often the case).

Let $Q = Q_1 R_1$ be the reduced decomposition.

Then using the above corollary:

$$\begin{aligned} x &= (A^T A)^{-1} A^T b = ((Q_1 R_1)^T Q_1 R_1)^{-1} (Q_1 R_1)^T b \\ &= (R_1^T Q_1^T Q_1 R_1)^{-1} R_1^T Q_1^T b \\ &= (R_1^T R_1)^{-1} R_1^T Q_1^T b \\ &= R_1^{-1} (R_1^T)^{-1} R_1^T Q_1^T b \\ &= R_1^{-1} Q_1^T b \end{aligned}$$

Hence

$$R_1 x = Q_1^T b,$$

from which we obtain x simply by substitution as R_1 is an upper triangular matrix.

Applications of QR decomposition

Further applications

- Solving systems of linear equations

  Suppose (for simplicity) a system $Ax = b$ with a regular $A \in \mathbb{R}^{n \times n}$.

  Obtaining x:

  1. Compute QR decomposition $A = QR$.
  2. Then $(QR)x = b$, i.e. $Rx = Q^T b$. Since R is upper triangular with a positive diagonal, x is obtained by consecutive substitution ($x_n = \frac{(Q^T b)_n}{R_{nn}}$, then obtain x_{n-1}, etc.).

  This is two-times slower than Gauss elimination but numerically more stable and of better precision.

- Computing an orthonormal base (alternative to Gram-Schmidt)

  Let $A \in \mathbb{R}^{m \times n}$ have linearly independent columns (these are the n given verctors subject to orthonormalization).

  We look for an orthonormal base of the column space $\mathcal{C}(A)$.

  Let $A = Q_1 R_1$ be the reduced QR decomposition.

  Regularity of R_1 implies $\mathcal{C}(A) = \mathcal{C}(Q_1)$, i.e. the columns of Q_1 form an orthonormal base of $\mathcal{C}(A)$.

  Furthermore: Q_2 forms an orthonormal base of $\ker(A^T)$, since $\ker(A^T)$ is the orthogonal complement of $\mathcal{C}(A)$.

  That is, from a QR decomposition of A (and A^T) we obtain orthonormal bases of the three spaces associated with A, i.e. $\mathcal{C}(A)$, $\mathcal{R}(A)$, and $\ker(A^T)$.

- QR algorithm

  Algorithm for computing eigenvalues of $A \in \mathbb{R}^{n \times n}$.

  Listed among 10 most important algorithms of the 20th century (see above).

  Goes back to 1960s (independently J. G. F. Francis, V. R. Kublanovskaya). Core of modern algorithms.

  Very simple using QR decomposition. Details see e.g. Hladík's book.

SVD (singular value decomposition)

presentation according to:
Hladík: Lineární algebra nejen pro informatiky (2019)
Eldén: Matrix Methods in Data Mining and Pattern Recognition (2007)

- perhaps the most useful matrix decomposition with applications in a variety of problems
- discovered independently by Eugenio Beltrami (1873), Camille Jordan (1874), James Joseph Sylvester (1889), Léon César Autonne (1915)

a decomposition

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

of $A \in \mathbb{R}^{m\times n}$ into $U \in \mathbb{R}^{m\times m}$, diagonal $\Sigma \in \mathbb{R}^{m\times n}$, and $V \in \mathbb{R}^{n\times n}$.

example (Eldén):

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} -0.2195 & 0.8073 & 0.0236 & 0.5472 \\ -0.3833 & 0.3912 & -0.4393 & -0.7120 \\ -0.5472 & -0.0249 & 0.8079 & -0.2176 \\ -0.7110 & -0.4410 & -0.3921 & 0.3824 \end{pmatrix} \begin{pmatrix} 5.7794 & 0 \\ 0 & 0.7738 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} -0.3220 & -0.9467 \\ 0.9467 & -0.3220 \end{pmatrix}$$

in MATLAB/GNU Octave:

```
octave:3> A=[1 1; 1 2; 1 3; 1 4]

A =
1 1
1 2
1 3
1 4

octave:4> [U,S,V]=svd(A)

U =
-0.219529 0.807346 0.023607 0.547214
-0.383342 0.391214 -0.439345 -0.712023
-0.547155 -0.024917 0.807869 -0.217595
-0.710969 -0.441048 -0.392131 0.382405

S =
Diagonal Matrix
5.7794 0
0 0.7738
0 0
0 0

V =
-0.3220 0.9467
-0.9467 -0.3220
```

Theorem (SVD)

Let $A \in \mathbb{R}^{m \times n}$ and $q = \min(m, n)$. There exists a diagonal matrix $\Sigma \in \mathbb{R}^{m \times n}$ with $\Sigma_{11} \geq \cdots \geq \Sigma_{qq} \geq 0$ and orthogonal matrices matrices $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ such that

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

- The positive $\Sigma_{11}, \ldots, \Sigma_{rr}$ are also denoted $\sigma_1 \ldots, \sigma_r$ (i.e. $\Sigma_{r+1,r+1}, \ldots, \Sigma_{qq} = 0$).
- One then has $r = \mathrm{rank}(A)$ (why?, see later for one of several possible proofs).
- The values $\sigma_1 \ldots, \sigma_r$ are uniquely determined and are called the singular values of A. U and V are not uniquely determined.
- Columns of U and and V are called the (left and right) singular vectors of A.
- Why not present the decomposition in the form $A = U \Sigma V$, i.e. why V^T and not V (which clearly would be possible)? V^T is a tradition and has some technical advantages.
- Proof later.

- diagonal $n \times n$ matrix with $a_1, \ldots, a_n$ on the main diagonal is denoted $\mathrm{diag}(a_1, \ldots, a_n)$.
- Different names from "SVD" in other fields:

  strongly related to PCA (statistical data analysis),

  Karhunen-Loewe expansion (image analysis)

Basic comparison to QR decomposition:

- QR decomposition $A = QR$ treats rows and columns of A differently (nonsymmetrically)
- recall:

  If $A = Q_1 R_1$ is the reduced QR decomposition, then columns of Q_1 form an orthonormal base of $\mathcal{C}(A)$ (the orthonormal base of $\mathcal{R}(A)$ needs to be obtained from a QR decomposition of A^T)
- SVD $A = U\Sigma V$ treats columns and rows symmetrically; provides more information about A
- SVD orders information contained in A so that "dominating part" becomes visible

  tis property makes SVD appealing in data mining

Theorem (SVD and rank)

Let $A \in \mathbb{R}^{m \times n}$ and let $A = U\Sigma V^T$ an SVD with singular values $\sigma_1, \ldots, \sigma_r$ (i.e. $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_q = 0$ for $q = \min(m, n)$). Then

$$r = \operatorname{rank}(A).$$

Proof.

Obviously, $\mathrm{rank}(\Sigma) = r$ (why?).

Theorem from linear algebra:

If $Q \in \mathbb{R}^{m \times m}$ is regular, then for any $S \in \mathbb{R}^{m \times n}$ we have $\mathcal{R}(S) = \mathcal{R}(QS)$.

Hence if Q is regular and $r = \mathrm{rank}(S)$, then $r = \mathrm{rank}(QS)$.

Analogously, if $Q \in \mathbb{R}^{n \times n}$ is regular, then for any $S \in \mathbb{R}^{m \times n}$ we have $\mathcal{C}(S) = \mathcal{C}(SQ)$.

Hence if Q is regular and $r = \mathrm{rank}(S)$, then $r = \mathrm{rank}(QS)$.

Now U and V are orthogonal, and hence regular. Using tha above rules,

$$r = \mathrm{rank}(\Sigma) = \mathrm{rank}(U\Sigma) = \mathrm{rank}((U\Sigma)V) = \mathrm{rank}(A).$$

□

Reduced SVD

- Consider SVD $A = U\Sigma V^T$ and let $r = \mathrm{rank}(A)$. Let $U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}$ and $V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}$ where U_1 and V_1 have r columns (and U_2 and V_2 the remaining $m - r$ and $n - r$ columns, respectively).

  Let S be the diagonal matrix with the diagonal containing, in this succession, $\sigma_1, \ldots, \sigma_r$.

- Then

$$A = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} S & o \\ o & o \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} = U_1 S V_1^T.$$

  Note the dimensions: $U_1 \in \mathbb{R}^{m \times r}$, $S \in \mathbb{R}^{r \times r}$, $V_1^T \in \mathbb{R}^{r \times n}$

- $A = U_1 S V_1^T$ is called the reduced SVD. A full SVD may be reconstructed from the reduced form.

SVD (same as above) :

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} -0.2195 & 0.8073 & 0.0236 & 0.5472 \\ -0.3833 & 0.3912 & -0.4393 & -0.7120 \\ -0.5472 & -0.0249 & 0.8079 & -0.2176 \\ -0.7110 & -0.4410 & -0.3921 & 0.3824 \end{pmatrix} \begin{pmatrix} 5.7794 & 0 \\ 0 & 0.7738 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} -0.3220 & -0.9467 \\ 0.9467 & -0.3220 \end{pmatrix}$$

corresponding reduced SVD:

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} -0.2195 & 0.8073 \\ -0.3833 & 0.3912 \\ -0.5472 & -0.0249 \\ -0.7110 & -0.4410 \end{pmatrix} \begin{pmatrix} 5.7794 & 0 \\ 0 & 0.7738 \end{pmatrix} \begin{pmatrix} -0.3220 & -0.9467 \\ 0.9467 & -0.3220 \end{pmatrix}$$

in MATLAB/GNU Octave:

```
octave:3> A=[1 1; 1 2; 1 3; 1 4]

A =
1 1
1 2
1 3
1 4

octave:4> [U,S,V]=svd(A,0)

U =
-0.219529 0.807346
-0.383342 0.391214
-0.547155 -0.024917
-0.710969 -0.441048

S =
5.7794 0
0 0.7738

V =
-0.3220 0.9467
-0.9467 -0.3220
```

Singular values vs eigenvalues

(Advanced)
Definition: Let $A \in \mathbb{R}^{n \times n}$. Then $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}^n$ are called an eigenvalue and the corresponding eigenvector of A if $Ax = \lambda x$.

- Eigenvalues (and eigenvectors) are of the most important notions in linear algebra, for instance for analyzing linear systems. For instance, $\det(A) = \lambda \cdot \dots \cdot \lambda_n$ where λ_i are all the eigenvalues of A.
- Recall spectral decomposition:

  If $A = S\Lambda S^{-1}$ with a diagonal matrix Λ and regular S, we call this decomposition a spectral decomposition.

  In this case, the diagonal of Λ contains the eigenvalues of A.

Theorem (singular values vs eigenvalues)

Let $A \in \mathbb{R}^{n\times m}$ and $r = \operatorname{rank}(A)$. Let the eigenvalues of A^TA be $\lambda_1 \geq \cdots \geq \lambda_n$. Then

$$\sigma_1 = \sqrt{\lambda_1}, \ldots, \sigma_r = \sqrt{\lambda_r}.$$

Proof.

For an SVD $A = U\Sigma V^T$, we have

$$A^TA = V\Sigma^TU^TU\Sigma V^T = V\Sigma^T\Sigma V^T = V\operatorname{diag}(\sigma_1^2, \ldots, \sigma_r^2, 0 \ldots, 0)V^T.$$

Since $A^TA = V\operatorname{diag}(\sigma_1^2, \ldots, \sigma_r^2, 0 \ldots, 0)V^T$ is a spectral decomposition of A^TA, the numbers on the diagonal in the middle matrix are the eigenvalues of A^TA. □

Relationship to matrix 2-norm

(Advanced) Recall:

- Given a vector norm $\|\cdot\|$, the corresponding matrix norm (operator norm) $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ is defined by

$$\|A\| = \sup_{x\neq o} \frac{\|Ax\|}{\|x\|}.$$

- In particular for the 2-norm (Euclidean norm, i.e. $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$), one has the matrix 2-norm. It may be proved that

$$\|A\|_2 = \sqrt{\max_{i=1}^n \lambda_i(A^T A)},$$

  where $\lambda_i(A^T A)$ are the eigenvalues of $A^T A$.

- Singular values offer a simple formula for $\|A\|_2$:

Theorem (2-norm of matrix as largest singular value)

For any matrix $A\mathbb{R}^{m\times n}$ we have $\|A\|_2 = \sigma_1$, i.e. the 2-norm of A equals the largest signular value of A.

Proof.

Assume $m \geq n$ and consider an SVD $A = U\Sigma V^T$.

Since orthogonal transformations preserve the vector norm, one easily checks that $\|A\|_2 = \|\Sigma\|$
Now, since Σ is diagonal,

$$\|\Sigma\|_2^2 = \sup_{\|y\|_2=1} \|Ay\|_2^2 = \sup_{\|y\|_2=1} \sum_{i=1}^{n} \sigma_i^2 y_i^2 \leq \sup_{\|y\|_2=1} \sigma_1^2 \sum_{i=1}^{n} y_i^2 = \sigma_1^2$$

with equality if and only if $y = e_1$. □

SVD and fundamental subspaces of A

(Compare with results for QR; orthonormal base of $\mathcal{C}(A)$.)

Theorem

Let $m \geq n$, $A \in \mathbb{R}^{m \times n}$, and $A = U\Sigma V^T$ be an SVD with $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$.

(a) *The first r columns of U form an orthonormal base of $\mathcal{R}(A)$ and*

$$\operatorname{rank}(A) = \dim(\mathcal{R}(A)) = r.$$

(b) *The first r columns of V form an orthonormal base of $\mathcal{C}(A)$. and*

(c) *The columns $U_{*r+1}, \ldots, U_{*m}$ form an orthonomal base of $\ker(A^T)$.*

(d) *The columns $V_{*r+1}, \ldots, V_{*n}$ form an orthonomal base of $\ker(A)$.*

Proof.

Later. □

SVD and matrix approximation

- basic question: Let $A \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(A) = r$.
  What is the best approximation of A among all matrices of a rank $k < r$?
- Informally: What is the best simpler matrix that approximates A?
- Formally: Pick a matrix norm $\| \cdot \|$.
  Let $k < r$. Find matrix $Z \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(Z) = k$ with the smallest value (distance from A)
  $$\|A - Z\|.$$
- Crucial for applications in data analysis.
- Consider two cases: 2-norm and Frobenius norm.

2-norm is a matrix norm $\|\cdot\|_2$ induced by the vector 2-norm (Euclidean norm), i.e. by

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \qquad \text{via} \qquad \|A\| = \sup_{x \neq o} \frac{\|Ax\|}{\|x\|}.$$

Theorem

Let $A \in \mathbb{R}^{m \times n}$ with $\operatorname{rank}(A) > k$. The approximation problem

$$\min\{\|A - Z\|_2 \,;\, A \in \mathbb{R}^{m \times n},\ \operatorname{rank}(Z) = k\}$$

has the solution $Z = A_k$ where

$$A_k = U_k \Sigma_k V_k^T,$$

where $U_k = (U_{*1}, \ldots, U_{*k})$ (first k columns of U), $\Sigma_k = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$, and $V_k = (V_{*1}, \ldots, V_{*k})$ (first k columns of V). Moreover,

$$\|A - A_k\|_2 = \sigma_{k+1}.$$

Proof.

Later. □

Frobenius is a matrix norm $\|\cdot\|_F$ defined by

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} A_{ij}^2}.$$

Theorem

Let $A \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(A) > k$. The approximation problem

$$\min\{\|A - Z\|_F\,;\, A \in \mathbb{R}^{m\times n},\ \operatorname{rank}(Z) = k\}$$

has the solution $Z = A_k$ where

$$A_k = U_k \Sigma_k V_k^T,$$

where $U_k = (U_{*1}, \ldots, U_{*k})$ (first k columns of U), $\Sigma_k = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$, and $V_k = (V_{*1}, \ldots, V_{*k})$ (first k columns of V). Moreover, for $q = \min(m, n)$,

$$\|A - A_k\|_2 = \sqrt{\sum_{i=k+1}^{q} \sigma_i^2}.$$

Proof.

Later. □