Linear Algebra Methods for Data Mining

Saara Hyvönen, Saara.Hyvonen@cs.helsinki.fi

Spring 2007

1. Basic Linear Algebra

Example 1: Term-Document matrices

	Doc1	Doc2	Doc3	Doc4	Query
Term1	1	0	1	0	1
Term2	0	0	1	1	1
Term3	0	1	1	0	0

- The documents and the query are represented by a vector in $\mathbb{R}^n$ (here $n = 3$).
- In applications matrices may be large!
  Number of terms: 10^4, number of documents: 10^6.

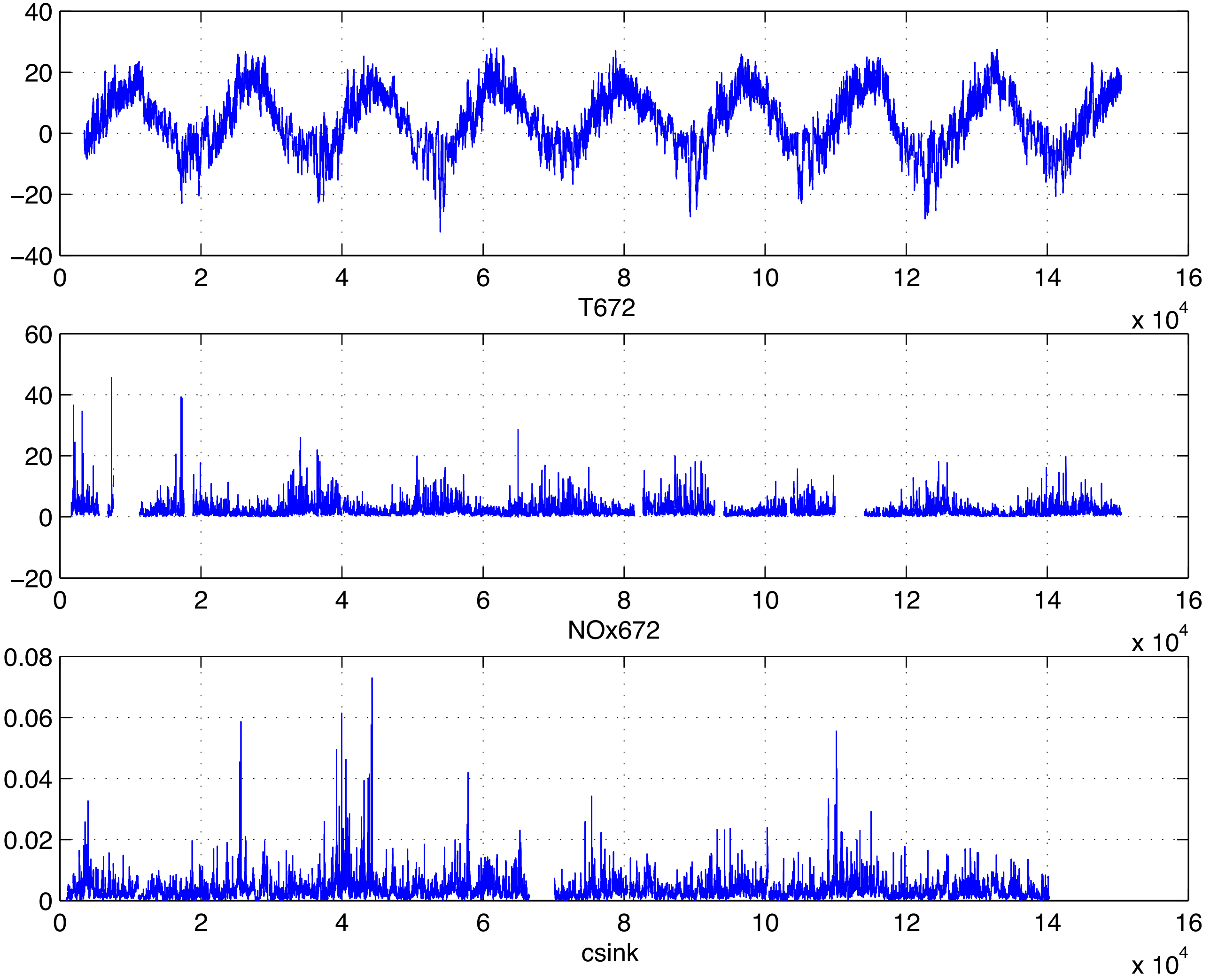
Example 1 continued: Tasks

- Find document vectors close to query.
  Use some distance measure in $\mathbb{R}^n$.

- Use linear algebra methods for
  - data compression
  - retrieval enhancement.

- Find "topics" or "concepts" from term-document matrix.

Example 2: measurement data



Example 2 continued

- In Hyytiälä Forest Field Station the 30 minute averages of some 100+ variables have measured for 10+ years...

- some 175 000 time points, 100 variables: alot of data!

- Possible question:
  - how do days vary? how do measured variables depend on each other?
  - what separates days when phenomenon X occurs from those when it doesn't?
  - are there (independent) (pollution) sources present?

Matrices

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Rectangular array of data: elements are real numbers.

Basic concepts

- vectors
- norms and distances
- eigenvalues, eigenvectors
- linearly independent vectors, basis
- orthogonal bases
- matrices, orthogonal matrices
- orthogonal matrix decompositions: SVD

Next: quick review of the following concepts:

- matrix-vector multiplication, matrix-matrix multiplication
- vector norms, matrix norms
- distances between vectors
- eigenvalues, eigenvectors
- linear independence
- basis
- orthogonality

Matrix-vector multiplication

$$
\mathbf{A}\mathbf{x} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \sum_{j=1}^{n} a_{2j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}x_j \end{pmatrix} = \mathbf{y}
$$

Symbolically

$$
\begin{pmatrix} \times \\ \times \\ \times \\ \times \end{pmatrix} = \begin{pmatrix} \leftarrow & - & - & \rightarrow \\ \leftarrow & - & - & \rightarrow \\ \leftarrow & - & - & \rightarrow \\ \leftarrow & - & - & \rightarrow \end{pmatrix} \begin{pmatrix} \uparrow \\ | \\ | \\ \downarrow \end{pmatrix}
$$

In practice

$$\begin{pmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 5 \\ -2 \end{pmatrix} = \begin{pmatrix} ? \\ ? \\ ? \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 5 \\ -2 \end{pmatrix} = \begin{pmatrix} 2 \cdot 5 + 3 \cdot (-2) \\ 6 \cdot 5 + 4 \cdot (-2) \\ 1 \cdot 5 + 0 \cdot (-2) \end{pmatrix} = \begin{pmatrix} 4 \\ 22 \\ 5 \end{pmatrix}$$

Or

$$\begin{pmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 5 \\ -2 \end{pmatrix} = 5 \cdot \begin{pmatrix} 2 \\ 6 \\ 1 \end{pmatrix} - 2 \cdot \begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 22 \\ 5 \end{pmatrix}$$

Alternative presentation of matrix-vector multiplication:

Denote the column vectors of the matrix $\mathbf{A}$ by $\mathbf{a_j}$. Then

$$\mathbf{y} = \mathbf{A}\mathbf{x} = (\mathbf{a}_1 \mathbf{a}_2 \ldots \mathbf{a}_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \sum_{j=1}^{n} x_j \mathbf{a}_j$$

So the vector $\mathbf{y}$ is a **linear combination** of the columns of $\mathbf{A}$.

Often this is a useful way to consider matrix-vector multiplication:

Example

Let columns of $\mathbf{A}$ be different "topics":

	Topic1	Topic2	Topic3
Term1	1	0	0
Term2	1	0	0
Term3	0	1	0
Term4	0	0	1

$= \mathbf{A}.$

Then if we multipy $\mathbf{A}$ by the vector $w = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}$, we get...

$$
\mathbf{A}\mathbf{w} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}
$$

$$
= 2 \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + 0 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + 1 \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 0 \\ 1 \end{pmatrix} = \mathbf{y},
$$

which represents a document dealing primarily with topic 1 and secondarily with topic 3.

Note on computational aspects

- The column oriented approach is also good when considering computational efficiency.
- Modern computing devices are able to exploit the fact that a vector operation is a very regular sequence of scalar operations.
- This approach is embedded in packages like Matlab and LAPACK (and others).
- SAXPY, GAXPY

Matrix-matrix multiplication

Let $\mathbf{A} \in \mathbb{R}^{m \times s}$ and $\mathbf{B} \in \mathbb{R}^{s \times n}$. Then, by definition,

$$\mathbb{R}^{m \times n} \ni \mathbf{C} = \mathbf{AB} = (c_{ij}),$$

$$c_{ij} = \sum_{k=1}^{s} a_{ik} b_{kj}, \quad i = 1...m, \quad j = 1...n.$$

Note: each column vector in $\mathbf{B}$ is multiplied by $\mathbf{A}$.

In practice

$$\begin{pmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 5 & 1 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} ? & ? \\ ? & ? \\ ? & ? \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 5 & 1 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 2\cdot 5 - 3\cdot 2 & 2\cdot 1 + 3\cdot 1 \\ 6\cdot 5 - 4\cdot 2 & 6\cdot 1 + 4\cdot 1 \\ 1\cdot 5 - 0\cdot 2 & 1\cdot 1 + 0\cdot 1 \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 22 & 10 \\ 5 & 1 \end{pmatrix}$$

$$= \left(\quad 5\cdot\begin{pmatrix} 2 \\ 6 \\ 1 \end{pmatrix} - 2\cdot\begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix} \quad 1\cdot\begin{pmatrix} 2 \\ 6 \\ 1 \end{pmatrix} + 1\cdot\begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix} \right)$$

Matrix multiplication code

```
for i=1:m,
    for j=1:n,
        for k=1:s,
            c(i,j)=c(i,j)+a(i,s)*b(s,j);
        end
    end
end
```

Note: loops may be permuted in 6 different ways!

How to measure the "size" of a vector?

Vector norms

The most common vector norms are

- 1-norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$
- Euclidean norm: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$
- max-norm: $\|\mathbf{x}\|_\infty = \max_{1 \le i \le n} |x_i|$
- all of the above are special cases of the L_p-norm (or p-norm): $\|\mathbf{x}\|_p = (\sum_{i=1}^{n} x_i^p)^{1/p}$

General definition of a vector norm

Generally, a vector norm is a mapping $\mathbb{R}^n \to \mathbb{R}$, with the properties

- $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x}$,
- $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$,
- $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$, for all $\alpha \in \mathbb{R}$,
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, the triangular equality.

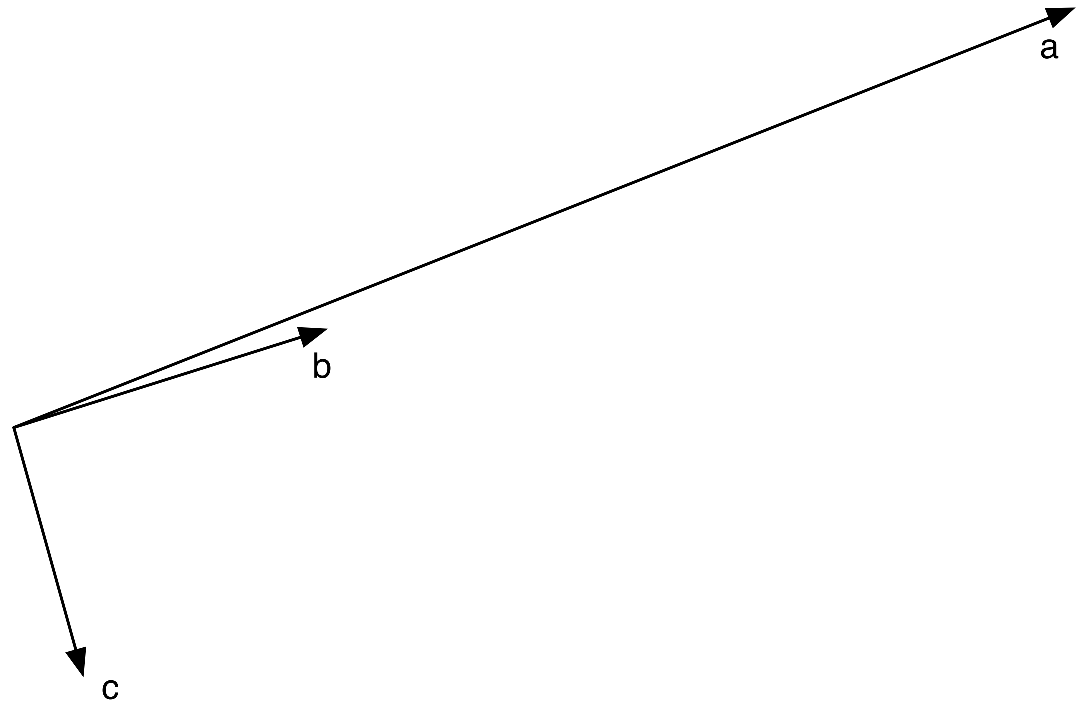
How to measure distance between vectors?

- Obvious answer: the distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ is $\|\mathbf{x} - \mathbf{y}\|$, where $\|\cdot\|$ is some vector norm.
- Frequently one measures the distance by the Euclidean norm $\|\mathbf{x} - \mathbf{y}\|_2$. So usually, if the index is dropped, this is what is meant.
- Alternative: use the angle between two vectors $\mathbf{x}$ and $\mathbf{y}$ to measure the distance between them.
- How to calculate the angle between two vectors?

Angle between vectors

- The **inner product** between two vectors is defined by $(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$.
- This is associated with the Euclidean norm: $\|\mathbf{x}\|_2 = (\mathbf{x}^T\mathbf{x})^{1/2}$.
- The angle θ between two vectors $\mathbf{x}$ and $\mathbf{y}$ is $\cos\theta = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}$.
- The cosine of the angle between two vectors $\mathbf{x}$ and $\mathbf{y}$ can be used to measure the **similarity** between the two vectors:
  - if $\mathbf{x}$ and $\mathbf{y}$ are close, the angle between them is small, and $\cos\theta \approx 1$.
  - $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal**, if $\theta = \frac{\pi}{2}$, i.e. $\mathbf{x}^T\mathbf{y} = 0$.

Why not just use the Euclidean distance?



Example: term-document matrix

Each entry tells how many times a term appears in the document:

	Doc1	Doc2	Doc3
Term1	10	1	0
Term2	10	1	0
Term3	0	0	1

- Using the Euclidean distance Documents 1 and 2 look dissimilar, and Documents 2 and 3 look similar. This is just due to the length of the documents!

- Using the cosine of the angle between document vectors Documents 1 and 2 are similar to each other and dissimilar to Document 3.

Eigenvalues and eigenvectors

- Let $\mathbf{A}$ be a $n \times n$ matrix. The vector $\mathbf{v}$ that satisfies

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$$

  for some scalar λ is called the **eigenvector** of $\mathbf{A}$ and λ is the **eigenvalue** corresponding to the eigenvector $\mathbf{v}$.

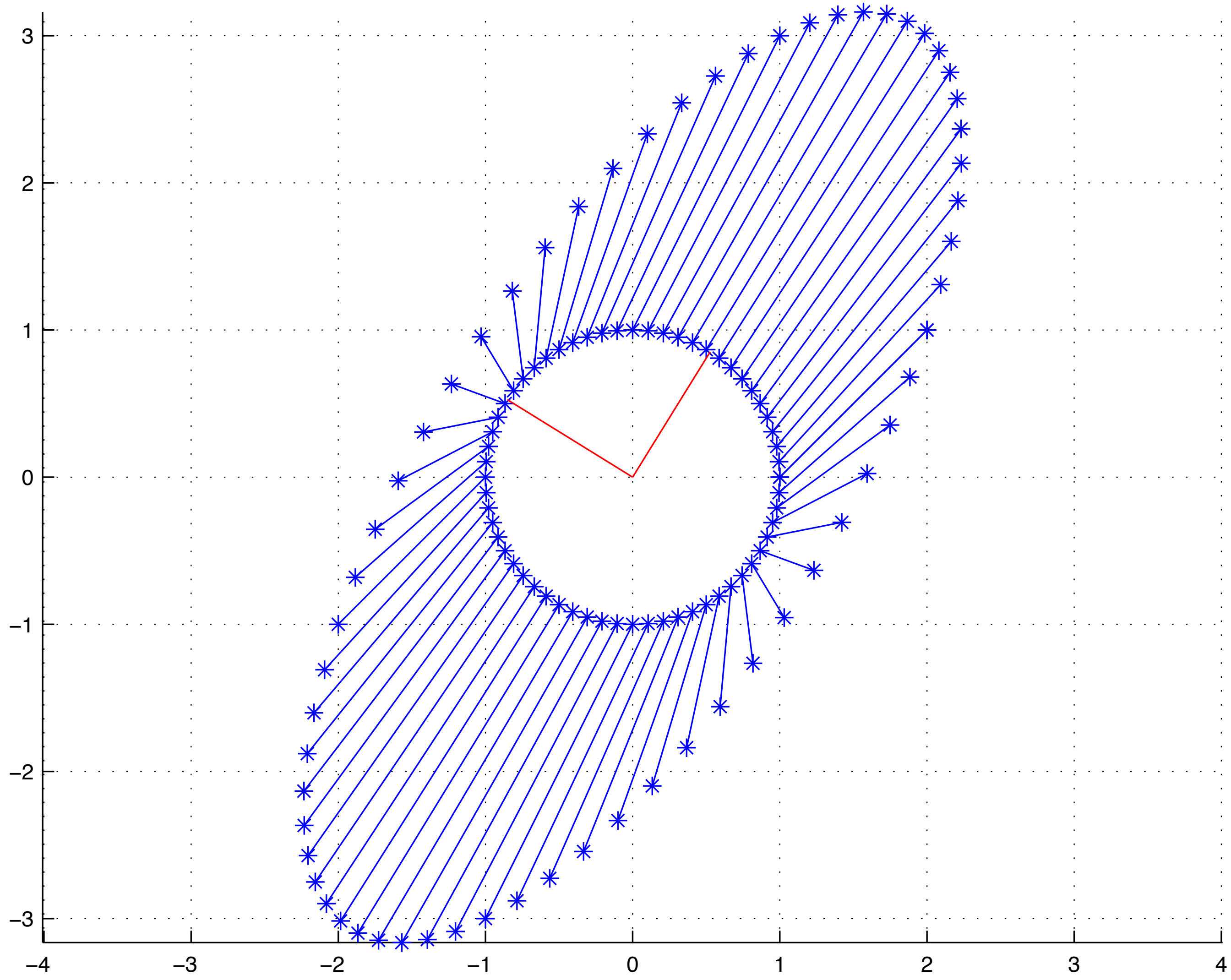
In practice

$$\mathbf{A}\mathbf{v} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \mathbf{v} = \lambda \mathbf{v}.$$

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix} = (2-\lambda)(3-\lambda) - 1 = 0 \quad \Rightarrow$$

$$\lambda_1 = 3.62 \qquad\qquad \lambda_2 = 1.38$$

$$\mathbf{v}_1 = \begin{pmatrix} 0.52 \\ 0.85 \end{pmatrix} \qquad\qquad \mathbf{v}_2 = \begin{pmatrix} 0.85 \\ -0.52 \end{pmatrix}$$


3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4

Matrix norms

- Let $\|.\|$ be a vector norm and $\mathbf{A} \in \mathbb{R}^{m \times n}$.
  The corresponding matrix norm is $\|\mathbf{A}\| = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|}$.

- $\|\mathbf{A}\|_2 = (\max_{1 \leq i \leq n} \lambda_i(\mathbf{A}^T\mathbf{A}))^{1/2}$ = square root of the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$. Heavy to compute!

- $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|$ (maximum over rows)

- $\|\mathbf{A}\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|$ (maximum over columns)

- $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$ Frobenius norm: does not correspond to any vector norm. Still, related to Euclidean vector norm.

Linear Independence

- Given a set of vectors $(\mathbf{v}_j)_{j=1}^n$ in $\mathbb{R}^m$, $m \geq n$, consider the set of linear combinations $y = \sum_{j=1}^n \alpha_j \mathbf{v}_j$ for arbitrary coefficients α_j.

- The vectors $(\mathbf{v}_j)_{j=1}^n$ are **linearly independent**, if $\sum_{j=1}^n \alpha_j \mathbf{v}_j = 0$ if and only if $\alpha_j = 0$ for all $j = 1, ..., n$.

- A set of m linearly independent vectors of $\mathbb{R}^m$ is called a **basis** in $\mathbb{R}^m$: any vector in $\mathbb{R}^m$ can be expressed as a linear combination of the basis vectors.

Example

The column vectors of the matrix

$$[\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3\ \mathbf{v}_4] = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

are not linearly independent, as

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \alpha_3 \mathbf{v}_3 + \alpha_4 \mathbf{v}_4 = 0$$

holds for $\alpha_1 = \alpha_3 = 1$, $\alpha_2 = \alpha_4 = -1$.

Rank of a matrix

- The **rank** of a matrix is the maximum number of linearly independent column vectors.

- A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with rank n is called **nonsingular**, and it has an **inverse** $\mathbf{A}^{-1}$ satisfying $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

- The (outer product) matrix $\mathbf{x}\mathbf{y}^T$ has rank 1: All columns of

$$\mathbf{x}\mathbf{y}^T = (y_1\mathbf{x}\ y_2\mathbf{x}\ \ldots\ y_n\mathbf{x})$$

are linearly dependent (and so are all the rows).

Example

The 4×4 matrix

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

has rank 3.

Example

- Consider a $m \times n$ term-document matrix $\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2\ \dots\ \mathbf{a}_n]$, where $\mathbf{a}_j \in \mathbb{R}^m$ are the documents.

- If $\mathbf{A}$ has rank 3, then all the documents can be expressed as a linear combination of only three vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3 \in \mathbb{R}^m$:

$$\mathbf{a}_j = w_{1j} \cdot \mathbf{v}_1 + w_{2j} \cdot \mathbf{v}_2 + w_{3j} \cdot \mathbf{v}_3, \quad j = 1, ..., n.$$

- The term-document matrix can be written as

$$\mathbf{A} = \mathbf{V}\mathbf{W}$$

where $\mathbf{V} = (\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3) \in \mathbb{R}^{m \times 3}$ and $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{3 \times n}$.

Condition number

- For any $a \neq 1$ the matrix $\mathbf{A} = \begin{pmatrix} a & 1 \\ 1 & 1 \end{pmatrix}$ is nonsingular and has the inverse $\mathbf{A}^{-1} = \frac{1}{a-1} \begin{pmatrix} 1 & -1 \\ -1 & a \end{pmatrix}$.
- As $a \to 1$, the norm of $\mathbf{A}^{-1}$ tends to infinity.
- Nonsingularity is not always enough!
- Define the **condition number** of a matrix to be $\kappa(A) = \|\mathbf{A}\| \|\mathbf{A}^{-1}\|$.
- Large condition number means trouble!

Orthogonality

- Two vectors $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal**, if $\mathbf{x}^T\mathbf{y} = 0$.

- Let $\mathbf{q}_j$, $j = 1, \ldots, n$ be orthogonal, i.e. $\mathbf{q}_i^T\mathbf{q}_j = 0$, $i \neq j$. Then they are linearly independent. (Proof?)

- Let the set of orthogonal vectors $\mathbf{q}_j$, $j = 1, \ldots, m$ in $\mathbb{R}^m$ be normalized, $\|\mathbf{q}\| = 1$. Then they are **orthonormal**, and constitute an **orthonormal basis** in $\mathbb{R}^m$.

- A matrix $\mathbb{R}^{m\times m} \ni \mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2\ \ldots\ \mathbf{q}_m]$ with orthonormal columns is called an **orthogonal matrix**.

Why we like orthogonal matrices

- An orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{m \times m}$ has rank m (since its columns are linearly independent).

- $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$. (Proofs?)

- The inverse of an orthogonal matrix $\mathbf{Q}$ is $\mathbf{Q}^{-1} = \mathbf{Q}^T$.

- The Euclidean length of a vector is invariant under an orthogonal transformation $\mathbf{Q}$: $\|\mathbf{Qx}\|^2 = (\mathbf{Qx})^T\mathbf{Qx} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2$.

- The product of two orthogonal matrices $\mathbf{Q}$ and $\mathbf{P}$ is orthogonal:

$$\mathbf{X}^T\mathbf{X} = (\mathbf{PQ})^T\mathbf{PQ} = \mathbf{Q}^T\mathbf{P}^T\mathbf{PQ} = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}.$$

References

[1] Lars Eldén: Matrix Methods in Data Mining and Pattern Recognition, SIAM 2007.

[2] G. H. Golub and C. F. Van Loan. Matrix Computations. 3rd ed. Johns Hopkins Press, Baltimore, MD., 1996.