
Singular Value Decomposition

— Boston University CS 506 - Lance Galletti —

Recall

$$\begin{array}{c} \text{n data} \\ \text{points} \end{array} \left\{ \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nm} \end{pmatrix} \right.$$

$\underbrace{\hspace{10em}}$
m features

Goal

Examine this matrix and uncover its linear algebraic properties to:

1. Approximate A with a smaller matrix B that is easier to store but contains similar information as A
2. Dimensionality Reduction / Feature Extraction
3. Anomaly Detection & Denoising

Linear Algebra Review

Definition: The vectors in a set $\mathbf{V} = \{ \vec{\mathbf{v}}_1, \dots, \vec{\mathbf{v}}_n \}$ are **linearly independent** if

$$\mathbf{a}_1 \vec{\mathbf{v}}_1 + \dots + \mathbf{a}_n \vec{\mathbf{v}}_n = \vec{\mathbf{0}}$$

can only be satisfied by $\mathbf{a}_i = \mathbf{0}$

Note: this means no vector in that set can be expressed as a **linear combination** of other vectors in the set.

Linear Algebra Review

Definition:

The **determinant** of a square matrix A is a scalar value that encodes properties about the **linear mapping** described by A .

2x2:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$\det(A) = ad - bc$$

Linear Algebra Review

Definition:

The **determinant** of a square matrix A is a scalar value that encodes properties about the **linear mapping** described by A .

3x3:

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \quad \det(A) = a \cdot \det \begin{pmatrix} e & f \\ h & i \end{pmatrix} - b \cdot \det \begin{pmatrix} d & f \\ g & i \end{pmatrix} + c \cdot \det \begin{pmatrix} d & e \\ g & h \end{pmatrix}$$

Linear Algebra Review

Definition:

The **determinant** of a square matrix A is a scalar value that encodes properties about the **linear mapping** described by A .

$n \times n$:

Can recursively compute it. How?

Linear Algebra Review

Property:

n vectors $\{\vec{v}_1, \dots, \vec{v}_n\}$ in an **n**-dimensional space are **linearly independent** iff the matrix **A**:

$$\mathbf{A} = [\vec{v}_1, \dots, \vec{v}_n] \text{ (n x n)}$$

has non-zero determinant.

Q: Can **m** > **n** vectors in an **n**-dimensional space be linearly independent?

Linear Algebra Review

Definition:

A **basis** \mathbf{B} of a vector space (over a field \mathbf{F}) is a **linearly independent** subset of \mathbf{V} that **spans** \mathbf{V} . \mathbf{B} **spans** \mathbf{V} if for every vector \mathbf{v} in \mathbf{V} it is possible to choose $\mathbf{v}_1, \dots, \mathbf{v}_n$ in \mathbf{F} and $\vec{\mathbf{b}}_1, \dots, \vec{\mathbf{b}}_n$ in \mathbf{B} such that:

$$\mathbf{v} = \mathbf{v}_1 \vec{\mathbf{b}}_1 + \dots + \mathbf{v}_n \vec{\mathbf{b}}_n$$

Ex: North & East in 2d-plane

Linear Algebra Review

Definition:

The **rank** of a matrix **A** is the dimension of the vector space spanned by its column space. This is equivalent to the maximal number of linearly independent columns / rows of **A**.

Definition:

A matrix **A** is **full-rank** iff $\text{rank}(\mathbf{A}) = \min(m, n)$

Note: Get the rank of a matrix through the **Gram-Schmidt process**

Approximation

In practice, matrices describing our dataset contain a lot of redundant information.

It would be great to capture all the information of our dataset in the least amount of space possible.

Approximation

To store an $\mathbf{n} \times \mathbf{m}$ matrix \mathbf{A} requires storing $\mathbf{m} \cdot \mathbf{n}$ values.

However, if the rank of the matrix of \mathbf{A} is \mathbf{k} , \mathbf{A} can be factored as

$$\mathbf{A} = \mathbf{UV}$$

where

\mathbf{U} is $\mathbf{n} \times \mathbf{k}$

\mathbf{V} is $\mathbf{k} \times \mathbf{m}$

which requires storing $\mathbf{k}(\mathbf{m} + \mathbf{n})$ values.

Approximation

Goal:

Approximate \mathbf{A} with $\mathbf{A}^{(k)}$ (low-rank matrix) such that

1. $\mathbf{d}(\mathbf{A}, \mathbf{A}^{(k)})$ is small
2. k is small compared to m & n

Frobenius Distance

$$d_F(A, B) = \|A - B\|_F = \sqrt{\sum_{i,j} (a_{ij} - b_{ij})^2}$$

i.e. the pairwise differences in values of A and B

Approximation

Definition:

When $k < \text{rank}(\mathbf{A})$, the **rank-k approximation** of \mathbf{A} (in the least squares sense) is

$$A^{(k)} = \arg \min_{\{B | \text{rank}(B)=k\}} d_F(A, B)$$

Approximation

Definition:

The **Singular Value Decomposition** of a rank- r matrix A has the form

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

where

U is $n \times r$

The columns of U are orthogonal & unit length ($U^T U = I$)

V is $m \times r$

The columns of V are orthogonal & unit length ($V^T V = I$)

Approximation

Definition:

The **Singular Value Decomposition** of a rank-r matrix A has the form

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_r \end{pmatrix}$$

with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$

σ_i is the square root of the eigenvalues of $\mathbf{A}^T\mathbf{A}$ and are called **singular values**

Approximation

Property:

$$d_F(A, A^{(k)})^2 = \sum_{i=k+1}^r \sigma_i^2$$

Note: the larger **k** is, the smaller the distance.

Approximation

Find $\mathbf{A}^{(k)}$ by decomposing \mathbf{A} :

$$A = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1 & V_2 \end{pmatrix}$$

$$\mathbf{A}^{(k)} = \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^T$$

Where

\mathbf{U}_1 is $\mathbf{n} \times \mathbf{k}$

$\mathbf{\Sigma}_1$ is $\mathbf{k} \times \mathbf{k}$

\mathbf{V}_1 is $\mathbf{m} \times \mathbf{k}$

Approximation

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1

=

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

x

9.64	0
0	5.29

x

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Approximation

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1

~

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

X

9.64	0
0	0

X

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Approximation

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1

~

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

x

9.64	0
0	0

x

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Approximation

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1

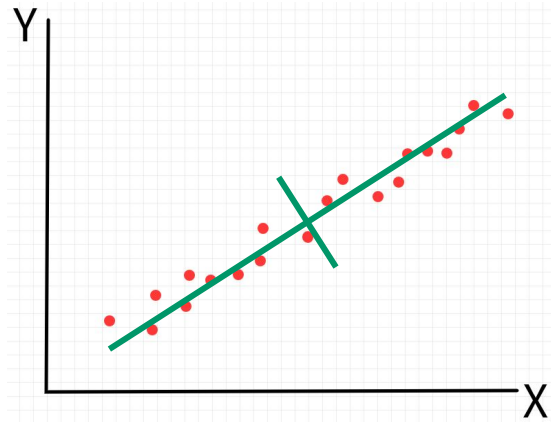
~

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Demo

Approximation

The i^{th} **singular vector** represents the direction of the i^{th} most variance.



Singular Values express the importance / significance of a singular vector

Approximation

To find the right **k** you can:

1. Look at the singular value plot to find the elbow point
2. Look at the residual error of choosing different **k**

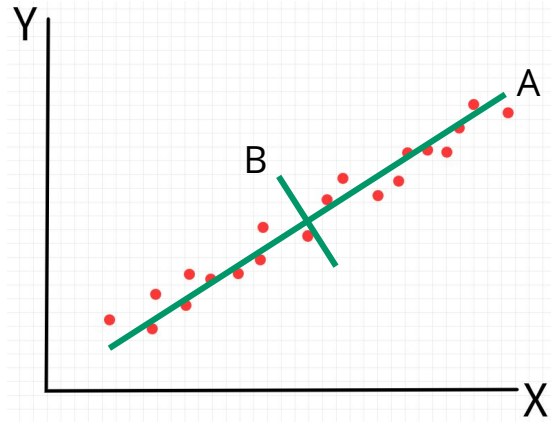
Demo

Principal Component Analysis

Idea: project the data onto a subspace generated from a subset of singular vectors / principal components.

We want to project onto the components that capture most of the variance / information in the data.

Principal Component Analysis



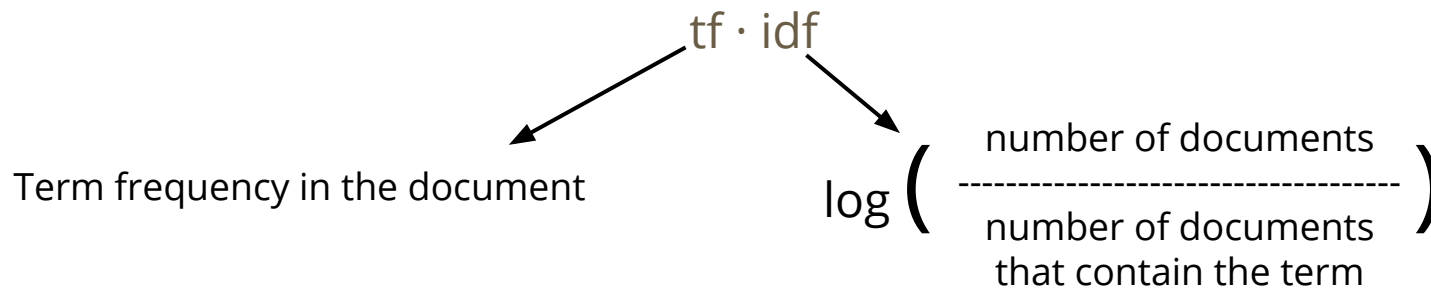
Which principal component should we project on?

Demo

Latent Semantic Analysis

Inputs are documents. Each word is a feature. We can represent each document by:

- The presence of the word (0 / 1)
- Count of the word (0, 1, ...)
- Frequency of the word ($n_i / \sum n_i$)
- TfiDf



Latent Semantic Analysis

	data	information	retrival	brain	lung
CS-paper-1	1	1	1	0	0
CS-paper-2	2	2	2	0	0
CS-paper-3	1	1	1	0	0
CS-paper-4	5	5	5	0	0
Med-paper-1	0	0	0	2	2
Med-paper-2	0	0	0	3	3
Med-paper-3	0	0	0	1	1

Latent Semantic Analysis

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1

 $=$

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

 \times

9.64	0
0	5.29

 \times

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Latent Semantic Analysis

CS concept

MD concept

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

X

9.64	0
0	5.29

X

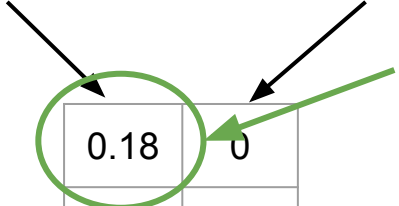
0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Latent Semantic Analysis

CS concept

MD concept

doc-to-concept similarity



0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

X

9.64	0
0	5.29

X

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Latent Semantic Analysis

doc-to-concept
similarity matrix

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

X

9.64	0
0	5.29

X

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Latent Semantic Analysis

doc-to-concept
similarity matrix

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

X

9.64	0
0	5.29

X

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

"strength" of the CS concept



Latent Semantic Analysis

doc-to-concept
similarity matrix

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

X

“strength” of the
each concept

9.64	0
0	5.29

X

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Latent Semantic Analysis

doc-to-concept
similarity matrix

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

X

"strength" of the
each concept

9.64	0
0	5.29

X

term-to-concept similarity

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Latent Semantic Analysis

doc-to-concept
similarity matrix

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53
0	0.80
0	0.27

X

"strength" of the
each concept

9.64	0
0	5.29

X

term-to-concept similarity
matrix

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Demo

Anomaly Detection

Define $\mathbf{O} = \mathbf{A} - \mathbf{A}^{(k)}$

The largest rows of \mathbf{O} could be considered anomalies

Demo