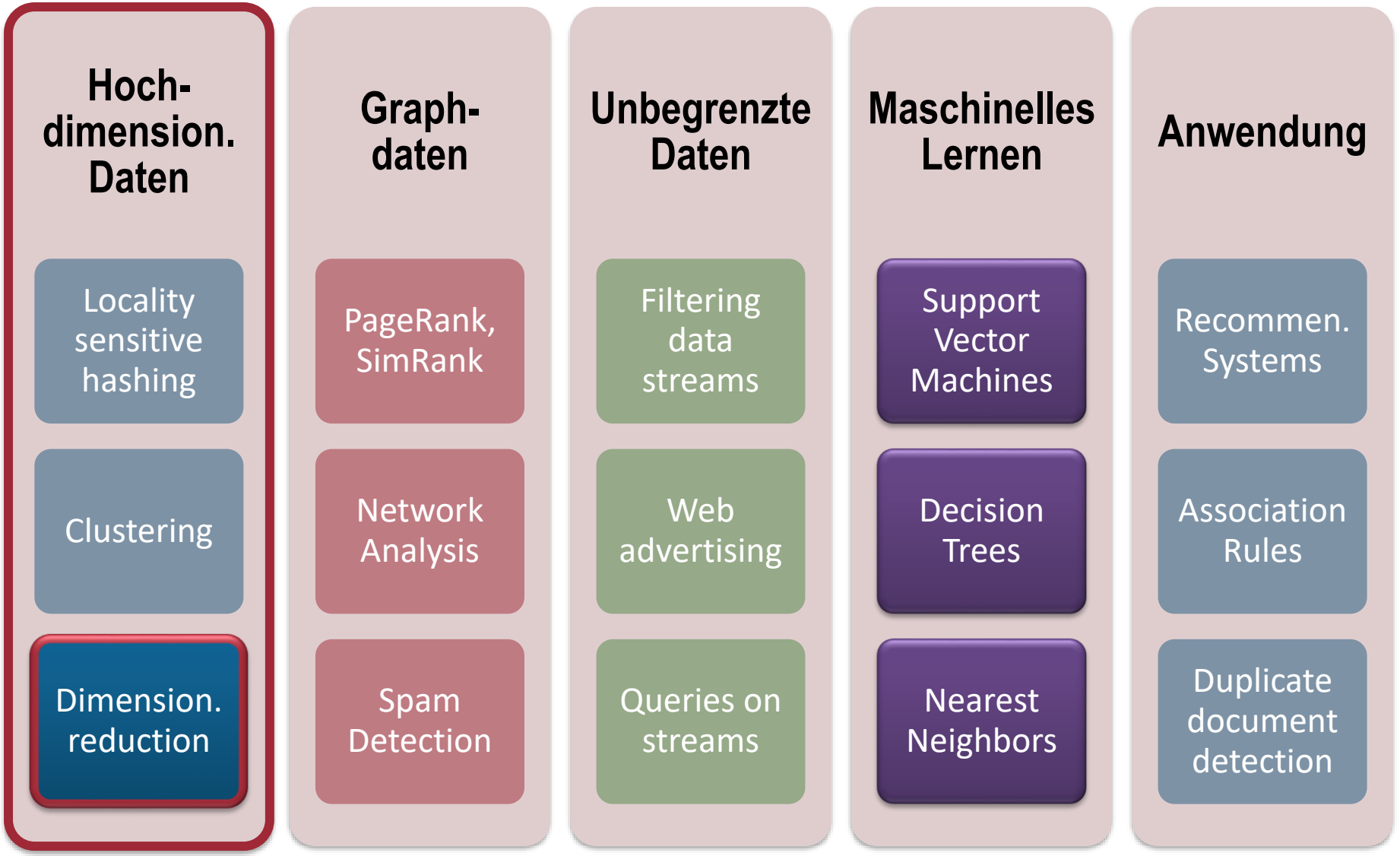
Data Mining

Kapitel 10: Dimensionality Reduction

Johannes Zschache
Wintersemester 2018/19

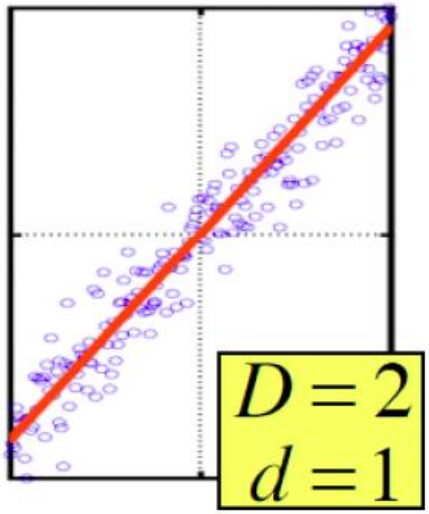
Abteilung Datenbanken, Universität Leipzig
http://dbs.uni-leipzig.de

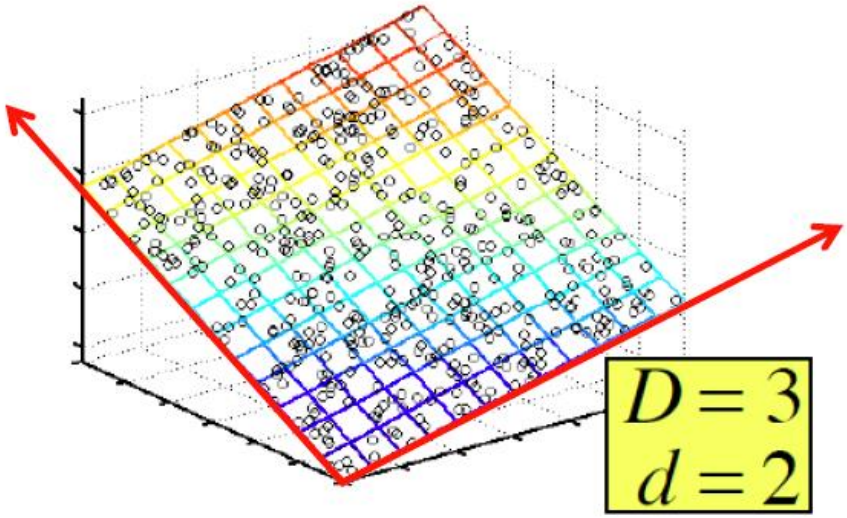
Übersicht


Hoch-dimension. Daten
Locality sensitive hashing
Clustering
Dimension. reduction
Graph-daten
PageRank, SimRank
Network Analysis
Spam Detection
Unbegrenzte Daten
Filtering data streams
Web advertising
Queries on streams
Maschinelles Lernen
Support Vector Machines
Decision Trees
Nearest Neighbors
Anwendung
Recommen. Systems
Association Rules
Duplicate document detection

Inhaltsverzeichnis

Dimensionsreduktion





- **Idee**: Falls Daten eines D-dimensionalen Raumes in der Nähe eines d-dimensionalen Unterraums liegen, können die Achsen des Unterraums als Faktoren dienen
- Wahl der Faktoren:
  - Erster Faktor zeigt in die Richtung, in welcher die Punkte ihre größte Streuung aufweisen
  - Zweiter Faktor ist orthogonal zum ersten Faktor und zeigt in die Richtung mit der größten Streuung unter den Punkten, usw..

Dimensionsreduktion

- Einfaches Beispiel:

Kunde	Montag	Dienstag	Mittwoch	Donnerstag	Freitag	Repräsentation
A	1	1	1	0	0	[1, 0]
B	2	2	2	0	0	[2, 0]
C	1	1	1	0	0	[1, 0]
D	5	5	5	0	0	[5, 0]
E	0	0	0	2	2	[0, 2]
F	0	0	0	3	3	[0, 3]

- **Rang** einer Matrix: Anzahl der *linear unabhängigen* Zeilen/Spalten
  - Im Beispiel: Zeilenrang ist 2
  - Aufspaltung:

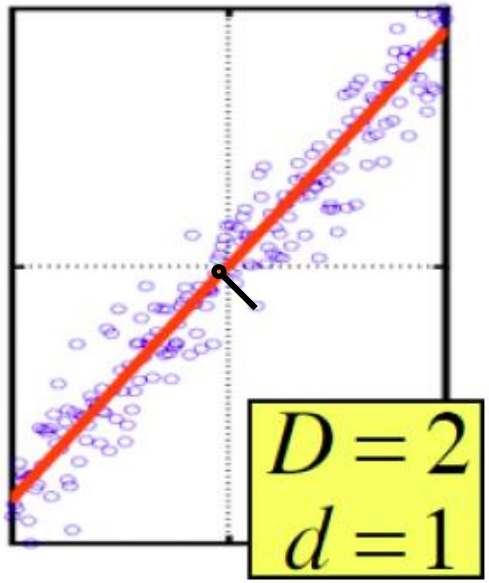
$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 1 & 0 \\ 5 & 0 \\ 0 & 2 \\ 0 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Neue Achsen/Faktoren

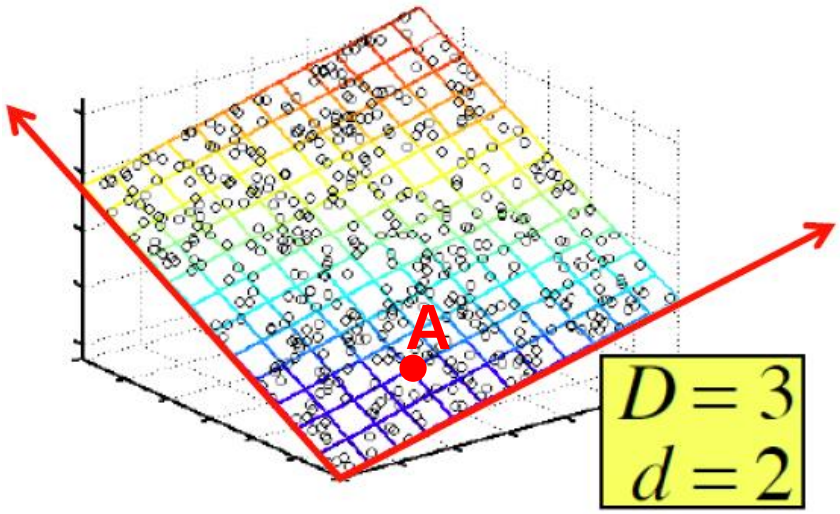
Neue Koordinaten

Dimensionsreduktion

Ziel: Aufdeckung der Datenachsen



Anstatt 2 Koordinaten, wird jeder Punkt nur über eine Koordinate repräsentieren: Position auf der roten Linie

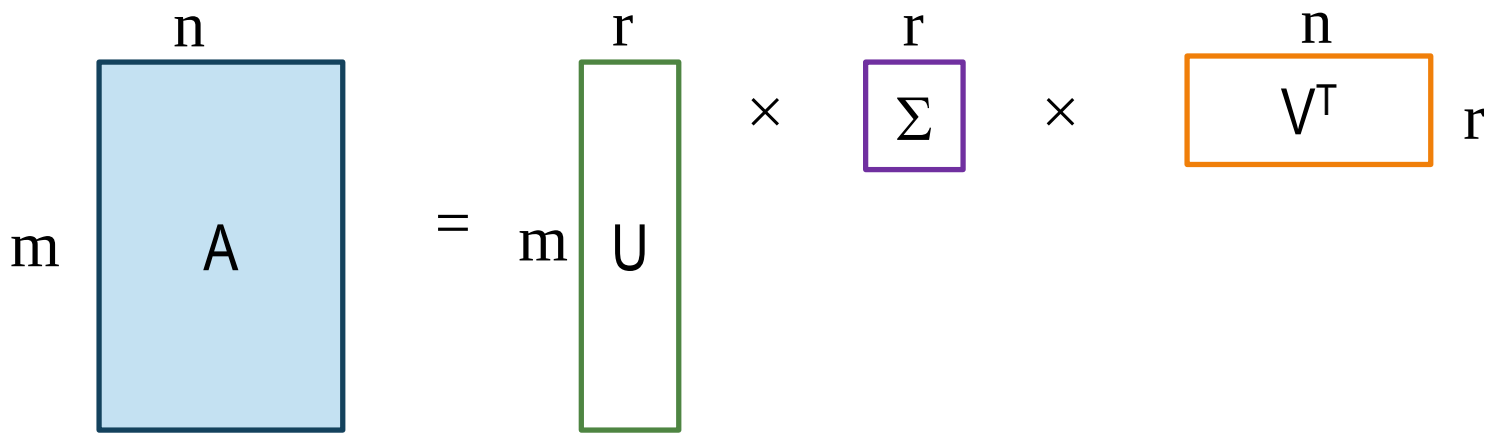


Punkt A wird z.B. anstatt durch [3,4,2] durch [2,1] repräsentiert (neue x-Achse ist Vektor (1,2,1))

Inhaltsverzeichnis

Singular Value Decomposition (SVD)

- Zerlegung eine Matrix A in das Produkt dreier Matrizen:

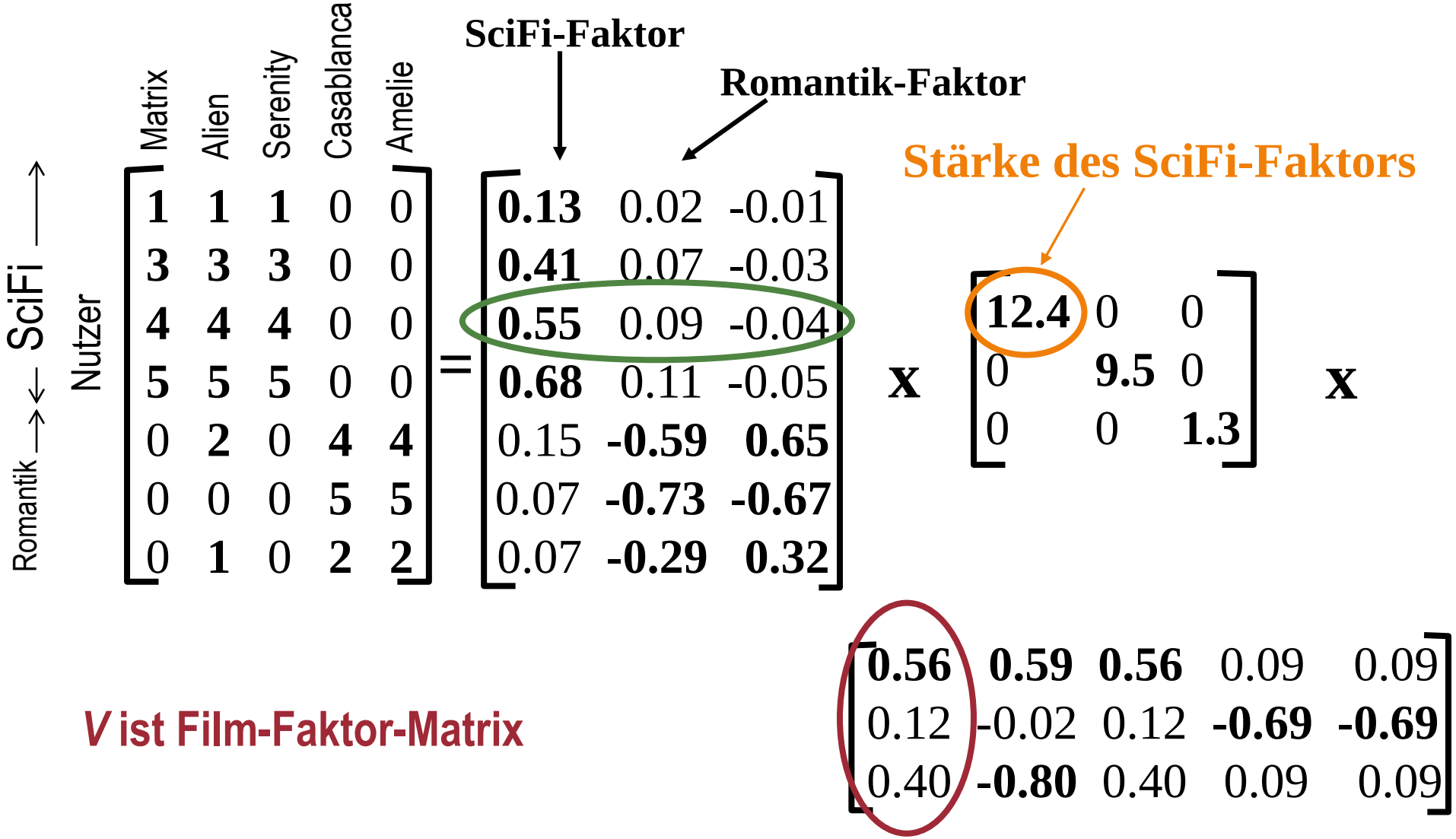


$$\boldsymbol{A}_{[\boldsymbol{m}\times\boldsymbol{n}]} = U_{[m\times r]} \cdot \Sigma_{[r\times r]} \cdot \left(V_{[n\times r]}\right)^T$$

- Matrix Σ ist Diagonalmatrix deren Einträge ($\boldsymbol{\sigma_1} \geq \boldsymbol{\sigma_2} \geq \ldots \geq \boldsymbol{\sigma_r} \geq 0$) nicht-negativ sind und als **Singulärwerte** bezeichnet werden
- Spalten von U und V sind orthonormal: $U^T \cdot U = V^T \cdot V = I$ (Einheitsmatrix)
- Anzahl der Faktoren = r = Rang von A

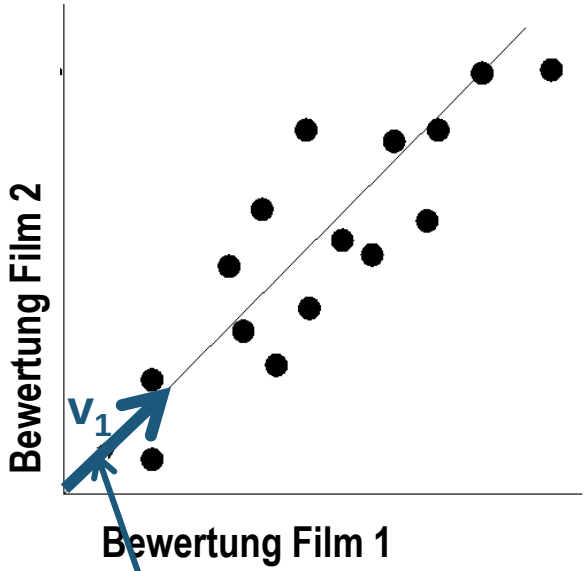
Beispiel: Nutzer und Filme

U ist Nutzer-Faktor-Matrix



V ist Film-Faktor-Matrix

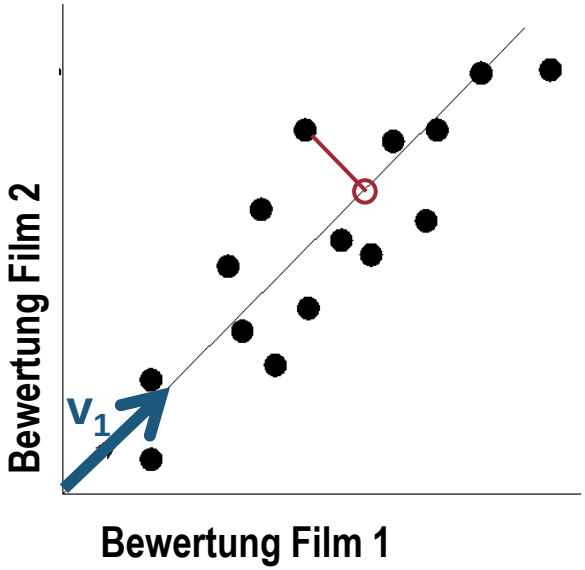
SVD: Dimensionsreduktion



$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \mathbf{x} \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix} \mathbf{x}$$

$$\begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

SVD: Dimensionsreduktion



$$\begin{bmatrix} \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix} \end{bmatrix}^T = \begin{bmatrix} 1.61 & 0.19 & -0.01 \\ 5.08 & 0.66 & -0.03 \\ 6.82 & 0.85 & -0.05 \\ 8.43 & 1.04 & -0.06 \\ 1.86 & -5.60 & 0.84 \\ 0.86 & -6.93 & -0.87 \\ 0.86 & -2.75 & 0.41 \end{bmatrix}$$

Projektionen auf die Achse des SciFi-Faktors $(U \Sigma)^T$

SVD: Dimensionsreduktion

Reduktion der Dimensionen: Setze kleinsten Singulärwerte auf Null

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

Rang-2-Approximation von A (je größer der Rang desto genauer die Approximation)

SVD: Dimensionsreduktion

$$A = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} \approx \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \mathbf{x} \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{x} \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

$$A - B = \begin{bmatrix} \mathbf{0.92} & \mathbf{0.95} & \mathbf{0.92} & 0.01 & 0.01 \\ \mathbf{2.91} & \mathbf{3.01} & \mathbf{2.91} & -0.01 & -0.01 \\ \mathbf{3.90} & \mathbf{4.04} & \mathbf{3.90} & 0.01 & 0.01 \\ \mathbf{4.82} & \mathbf{5.00} & \mathbf{4.82} & 0.03 & 0.03 \\ 0.70 & \mathbf{0.53} & 0.70 & \mathbf{4.11} & \mathbf{4.11} \\ -0.69 & 1.34 & -0.69 & \mathbf{4.78} & \mathbf{4.78} \\ 0.32 & \mathbf{0.23} & 0.32 & \mathbf{2.01} & \mathbf{2.01} \end{bmatrix}$$

Genauigkeit über Frobeniusnorm:

$$\|A - B\|_F = \sqrt{\Sigma_{ij} (A_{ij} - B_{ij})^2}$$

SVD

- **Satz**: Sei k mit $0 \leq k \leq r$ die Anzahl der gewünschten Faktoren, $\boldsymbol{A} = \boldsymbol{U}\,\Sigma\,\boldsymbol{V}^T$ und $\boldsymbol{B} = \boldsymbol{U}\,\boldsymbol{S}\,\boldsymbol{V}^T$ wobei $\boldsymbol{S}$ aus Σ konstruiert wurde, indem die letzten $r - k$ Diagonalelemente auf Null gesetzt wurden. Dann gilt:

$$\mathrm{B} = \min_{\mathrm{C}} \|A - C\|_F$$

- Für eine gegebene Anzahl an Faktoren k minimiert SVD den Fehler $\|A - B\|_F$, so dass B die *beste* Rang-k-Approximation für A darstellt

- Wie klein sollte man k wählen?
- Behalte 80-90% der „Energie“ $\sum_i \sigma_i^2$ (Summe über die quadrierten Diagonalelemente von Σ)
- Beispiel: Singulärwerte 12.4, 9.5, und 1.3 $\rightarrow$ Energie: 245.7
  - Entfernen des letzten Singulärwertes setzt Energie auf 244 (99%)
  - Mit nur dem größten Singulärwert wäre die Energie auf 63% reduziert

SVD: Berechnung

- SVD für eine Matrix $\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^{\boldsymbol{T}}$
- Es gilt: $A^{\mathrm{T}} = \left(U\Sigma V^{\mathrm{T}}\right)^{\mathrm{T}} = \left(V^{\mathrm{T}}\right)^{\mathrm{T}}\Sigma^{\mathrm{T}}U^{\mathrm{T}} = V\Sigma U^{\mathrm{T}}$
  - Regel für die Transponierte eines Produkts von Matrizen
  - Zweifache Transposition löst sich auf
  - Transposition einer Diagonalmatrix ergibt die selbe Diagonalmatrix
- Somit gilt: $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{V}\Sigma\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}\Sigma^{2}\boldsymbol{V}^{\mathrm{T}}$
  - Da Spalten von U orthonormal: $U^{\mathrm{T}}U = \mathrm{I}$ (Identitätsmatrix)
  - Σ^2 ist eine Diagonalmatrix deren i-tes Diagonalelement das Quadrat des i-ten Diagonalelements von Σ ist
- Da auch die Spalten von V orthonormal: $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{V}\Sigma^{2}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{V}\Sigma^{2}$

D.h. V ist die Matrix aus Eigenvektoren von $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ und die Diagonalelemente von Σ^2 sind die dazugehörigen Eigenwerte.

- Analog: $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{U}\Sigma^{2}$

Eigenwerte und -vektoren

- Sei M eine quadratische Matrix. Eine reelle Zahl λ heißt Eigenwert von M und ein Vektor $e \neq 0$ der dazugehörige Eigenvektor, falls $Me = \lambda e$.
- Beispiel:

$$\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} = 7 \begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix}$$

- Somit ist $\begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} \approx \begin{bmatrix} 0.447 \\ 0.894 \end{bmatrix}$ ein Eigenvektor der Matrix $\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$ und 7 der dazugehörige Eigenwert

- Berechnung der Eigenvektoren und –werte z.B. über **die Power-Iteration-Methode** (siehe auch Folie 4-14 aus „Link Analyis“)

Eigenwerte und -vektoren

- Power-Iteration-Methode:
  - Zu Beginn: beliebiger Vektor $\boldsymbol{x}_0 \neq 0$
  - Iteration: $\boldsymbol{x}_{k+1} = \frac{M\boldsymbol{x}_k}{||M\boldsymbol{x}_k||}$
  - || ... || bezeichnet die Frobeniusnorm
  - Stopp, falls Änderungen in $\boldsymbol{x}_k$ vernachlässigbar klein
- Beispiel: $\mathrm{M} = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$ und $x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
  - $Mx_0 = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$ und $\|Mx_0\| = \sqrt{5^2 + 8^2} = 9.434$
  - $x_1 = \begin{bmatrix} 0.530 \\ 0.848 \end{bmatrix}$
  - $Mx_1 = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} 0.530 \\ 0.848 \end{bmatrix} = \begin{bmatrix} 3.286 \\ 6.148 \end{bmatrix}$ und $\|Mx_1\| = 6.971$
  - $x_2 = \begin{bmatrix} 0.471 \\ 0.882 \end{bmatrix}$
  - …

Eigenwerte und -vektoren

- Die Power-Iteration-Methode berechnet den ersten Eigenvektor $\boldsymbol{x}$ (mit dem größten Eigenwert)
- Dazugehörige Eigenwert: $\lambda = \boldsymbol{x}^{\mathrm{T}} M \boldsymbol{x}$
- Beispiel:

$$\begin{bmatrix} 0.447 \\ 0.894 \end{bmatrix}^T \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} 0.447 \\ 0.894 \end{bmatrix} = 6.993$$

- Reduzierung der Matrix M um den Anteil, der durch den ersten Eigenwert und –vektor generiert wird:

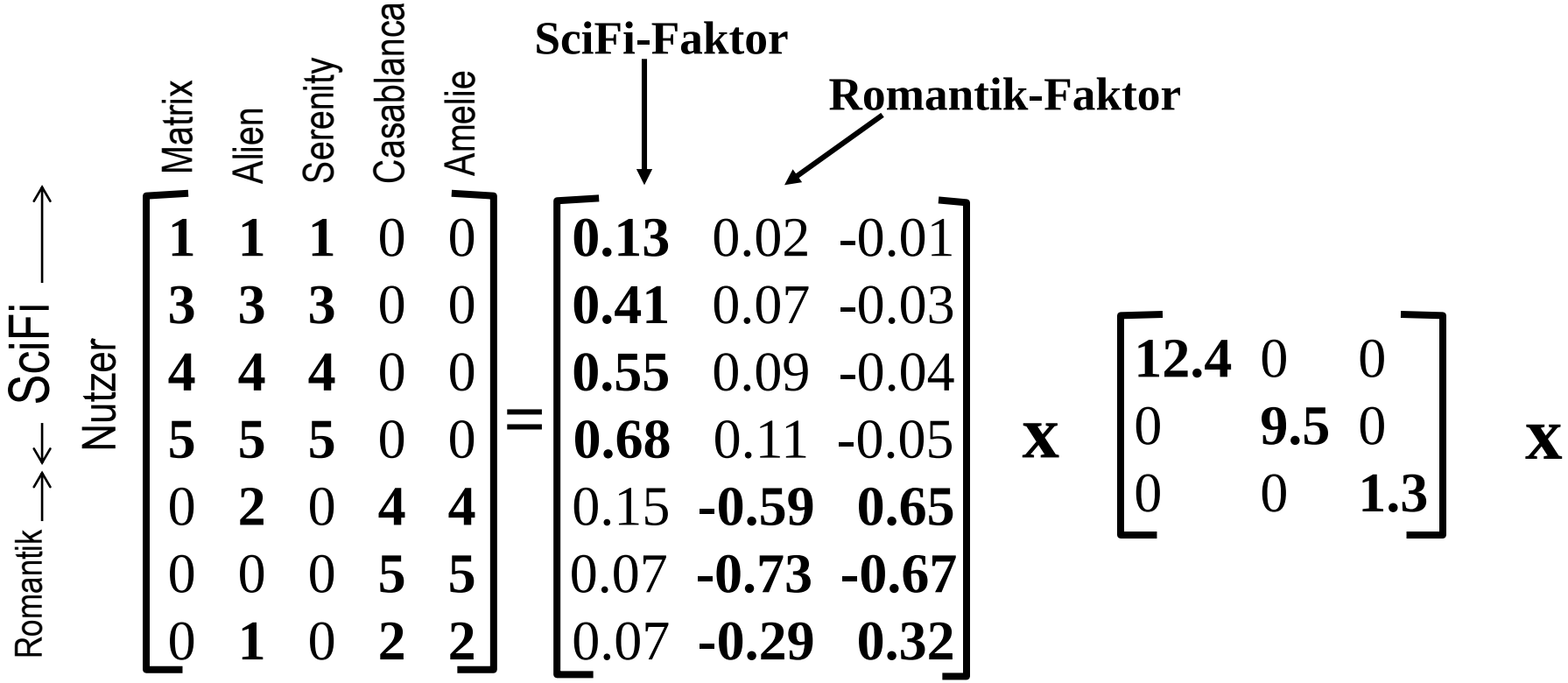
$$M^* := M - \lambda\, x\, x^T$$

- Power-Iteration-Method auf M^* berechnet den zweiten Eigenvektor von M (mit dem zweitgrößten Eigenwert von M)
- Beispiel:

$$\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} - 6.993 \begin{bmatrix} 0.447 \\ 0.894 \end{bmatrix} \begin{bmatrix} 0.447 \\ 0.894 \end{bmatrix}^T = \begin{bmatrix} 2.601 & -0.797 \\ -0.797 & 0.413 \end{bmatrix}$$

Beispiel: Nutzer und Filme

U ist Nutzer-Faktor-Matrix



V ist Film-Faktor-Matrix

$$\begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

Beispiel: Nutzer und Filme

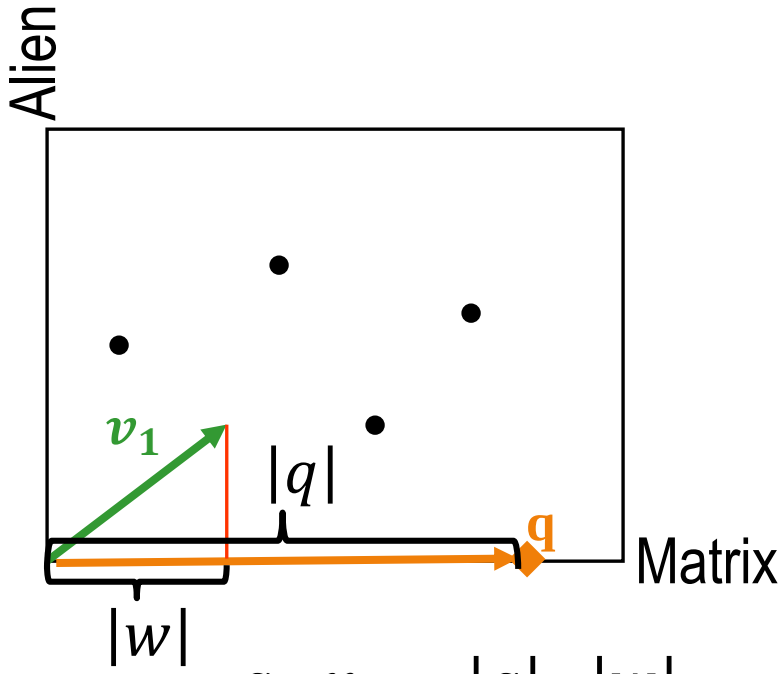
- Neuer Nutzer q mit Bewertung für Film “Matrix”:

	Matrix	Alien	Serenity	Casablanca	Amelie
q =	5	0	0	0	0

- Projektion in den Film-Faktor-Raum

	v_1	v_2
	0.56	0.12
	0.59	-0.02
	0.56	0.12
	0.09	-0.69
	0.09	-0.69

Ersten beiden Spalten der Film-Faktor-MatrixV



$$q \cdot v_1 = |q| \cdot |w|$$

SciFi-Faktor $q \cdot v_1 = 2.8$

Romantik-Faktor $q \cdot v_2 = 0.6$

Beispiel: Nutzer und Filme

- Neuer Nutzer q mit Bewertung für Film “Matrix”:

$$\mathbf{q} = \begin{bmatrix} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\ 5 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- Faktor-Repräsentation:

$$\begin{matrix} q \cdot v_1 & q \cdot v_2 \\ \begin{bmatrix} 2.8 & 0.6 \end{bmatrix} \end{matrix}$$

- Rückführung in den Film-Raum:

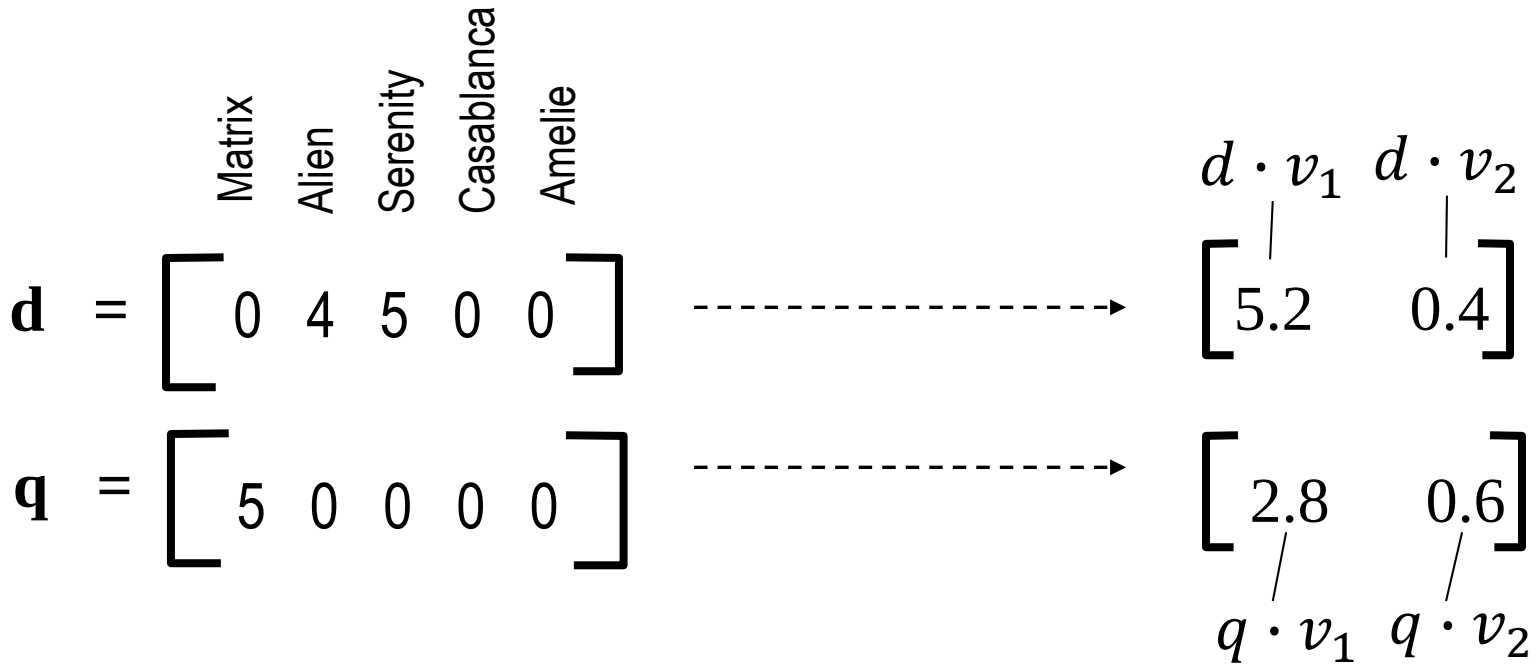
$$\begin{bmatrix} 2.8 & 0.6 \end{bmatrix} \mathbf{x} \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \end{bmatrix} =$$

$$\underbrace{\begin{bmatrix} \mathbf{1.64} & \mathbf{1.64} & \mathbf{1.64} & -1.62 & -1.62 \end{bmatrix}}$$

Empfehlungen für Filme „Alien“ und „Serenity“

Beispiel: Nutzer und Filme

- **Beobachtung**: Ein weiterer Nutzer **d**, der nur die Filme „Alien“ und „Serenity“ bewertet hat, kann, im Vergleich zum Nutzer q, einen ähnlichen Faktor-Vektor haben; *obwohl es keinen Film gibt, der von beiden Nutzern bewertet wurde*



Inhaltsverzeichnis

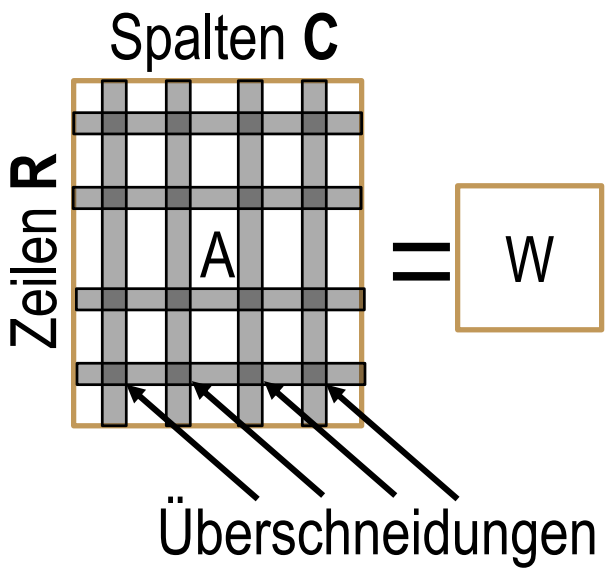
CUR Decomposition

- Die zu zerlegende Matrix A ist oft spärlich besetzt
- Aber die Matrizen der Zerlegung U und V sind nicht spärlich besetzt, d.h. die Berechnung dieser Matrizen ist sehr aufwendig
- Die CUR-Zerlegung löst dieses Problem
  - Die Matrizen der Zerlegung C und R sind spärlich besetzt, falls A spärlich besetzt
  - Matrix C besteht aus zufällig ausgewählten Spalten von A
  - Matrix R besteht aus zufällig ausgewählten Zeilen von A
  - Wahl von U, so dass $||A - C \cdot U \cdot R||_F$ klein

$$\begin{pmatrix} \\ A \\ \\ \end{pmatrix} \approx \begin{pmatrix} \\ C \\ \\ \end{pmatrix} \cdot \begin{pmatrix} \\ U \\ \\ \end{pmatrix} \cdot \begin{pmatrix} \\ R \\ \\ \end{pmatrix}$$

Berechnung der Matrix U

- $\boldsymbol{W}$ sei die Matrix aller Überschneidungen der Spalten C und der Reihen R
- Berechnung der SVD von $\boldsymbol{W}$: $\boldsymbol{W} = \boldsymbol{X}\,\boldsymbol{Z}\,\boldsymbol{Y}^{\mathrm{T}}$
- Dann ist $\boldsymbol{U} := \boldsymbol{Y}\,(\boldsymbol{Z}^{+})^{2}\,\boldsymbol{X}^{\mathrm{T}}$
  - Z^{+}: Diagonalmatrix mit Diagonalelementen: $Z_{ii}^{+} = \frac{1}{Z_{ii}}$ falls $Z_{ii} \neq 0$ und 0 sonst
  - Z^{+} nennt man auch "Moore-Penrose-Pseudoinverse": Anstatt $\mathrm{Z}\mathrm{Z}^{-1} = I$ gilt $\mathrm{Z}\mathrm{Z}^{+}\mathrm{Z} = Z$ und $\mathrm{Z}^{+}\mathrm{Z}\mathrm{Z}^{+} = Z^{+}$



- Beispiel:

$$W = \begin{bmatrix} 0 & 5 \\ 5 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Dann ist:

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1/5 & 0 \\ 0 & 1/5 \end{bmatrix}^{2} \cdot \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1/25 \\ 1/25 & 0 \end{bmatrix}$$

Auswahl der Spalten und Zeilen

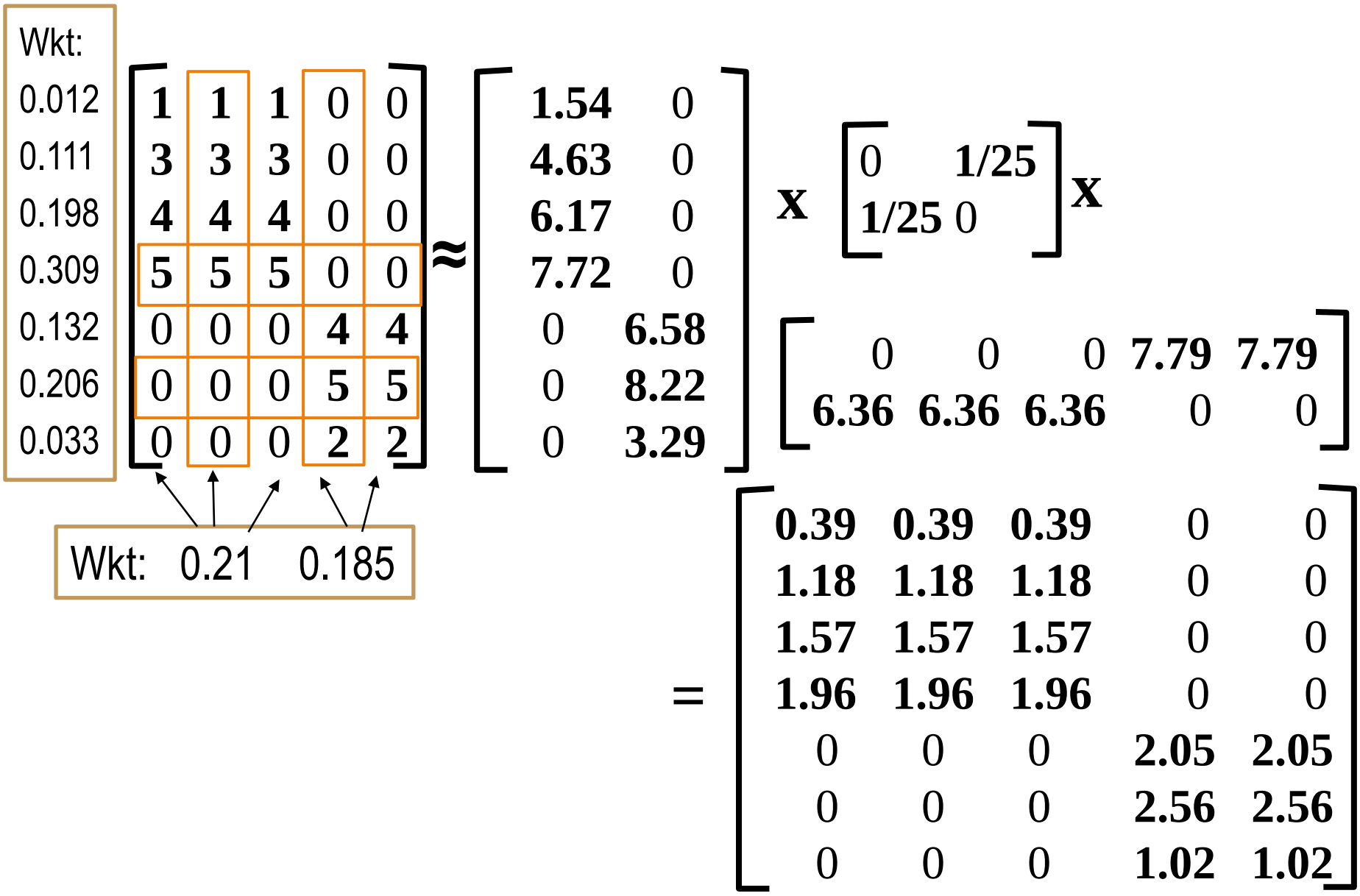
- Um den Fehler $||A - C \cdot U \cdot R||_F$ deutlich zu verringern sollten die Zeilen und Spalten nach *Wichtigkeit* ausgewählt werden
- Die Wichtigkeit einer Zeile/Spalte: quadrierte Frobeniusnorm
- Wahrscheinlichkeiten der Auswahl sind proportional zu deren Wichtigkeit
- Beispiel: Spalte [3,4,5] hat Wichtigkeit 50 und die Spalte [3,0,1] hat Wichtigkeit $10 \rightarrow$ Wahrscheinlichkeit für erste Zeile ist fünfmal so groß wie Wahrscheinlichkeit der zweiten Zeile
- Algorithmus für Spalten:

Input: matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, sample size c
Output: $\mathbf{C}_d \in \mathbb{R}^{m \times c}$

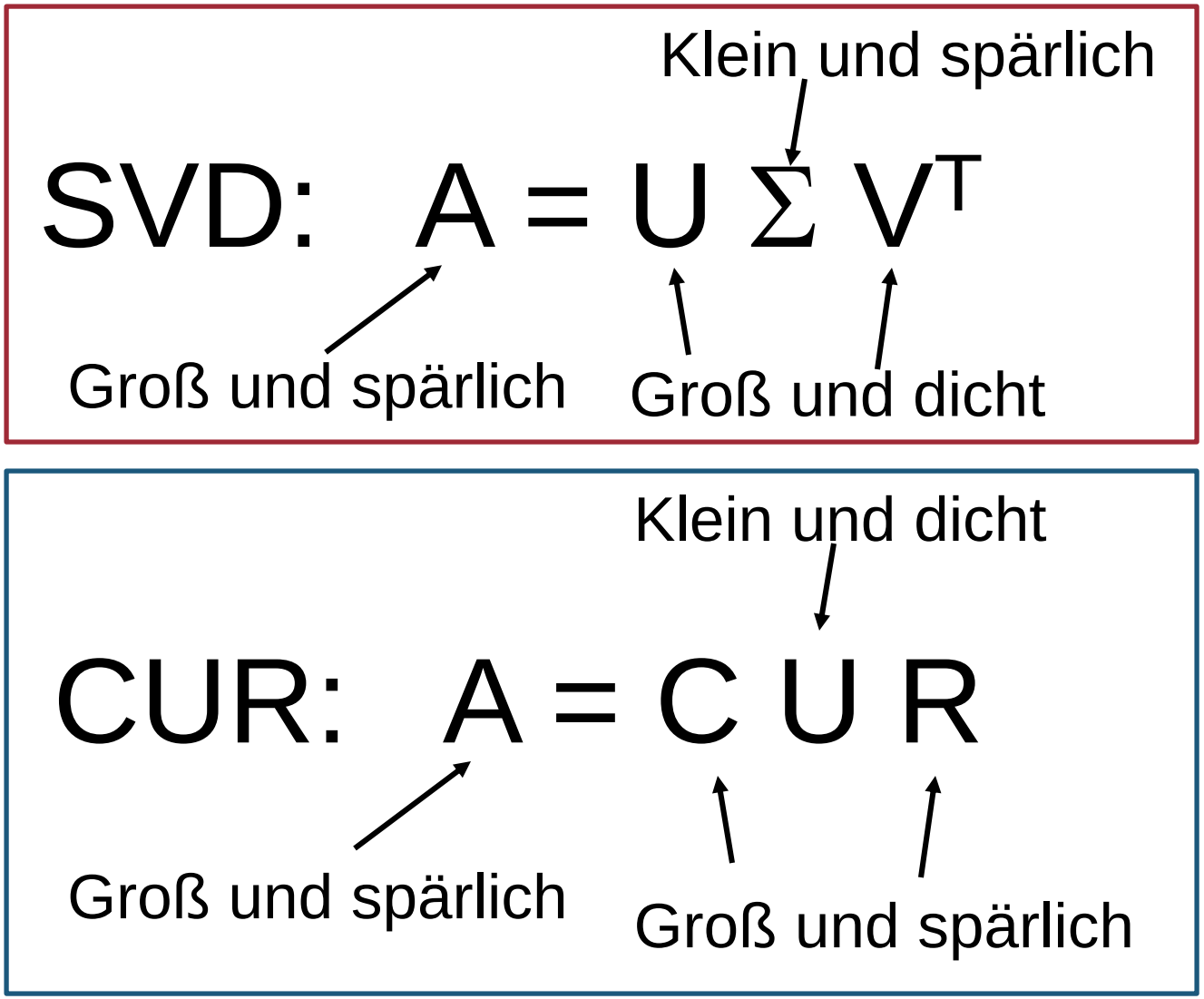
1. for $x = 1 : n$ [column distribution]
2. $P(x) = \sum_i \mathbf{A}(i,x)^2 / \sum_{i,j} \mathbf{A}(i,j)^2$
3. for $i = 1 : c$ [sample columns]
4. Pick $j \in 1 : n$ based on distribution $P(x)$
5. Compute $\mathbf{C}_d(:,i) = \mathbf{A}(:,j) / \sqrt{cP(j)}$

- Eine Spalte kann mehrmals ausgewählt werden
- Skalierung der Spalten, um dies zu korrigieren

Beispiel


Wkt:
0.012
0.111
0.198
0.309
0.132
0.206
0.033
1 1 1 0 0
3 3 3 0 0
4 4 4 0 0
5 5 5 0 0
0 0 0 4 4
0 0 0 5 5
0 0 0 2 2
≈
1.54 0
4.63 0
6.17 0
7.72 0
0 6.58
0 8.22
0 3.29
x
0 1/25
1/25 0
x
0 0 0 7.79 7.79
6.36 6.36 6.36 0 0
=
0.39 0.39 0.39 0 0
1.18 1.18 1.18 0 0
1.57 1.57 1.57 0 0
1.96 1.96 1.96 0 0
0 0 0 2.05 2.05
0 0 0 2.56 2.56
0 0 0 1.02 1.02
Wkt: 0.21 0.185

SVD vs. CUR

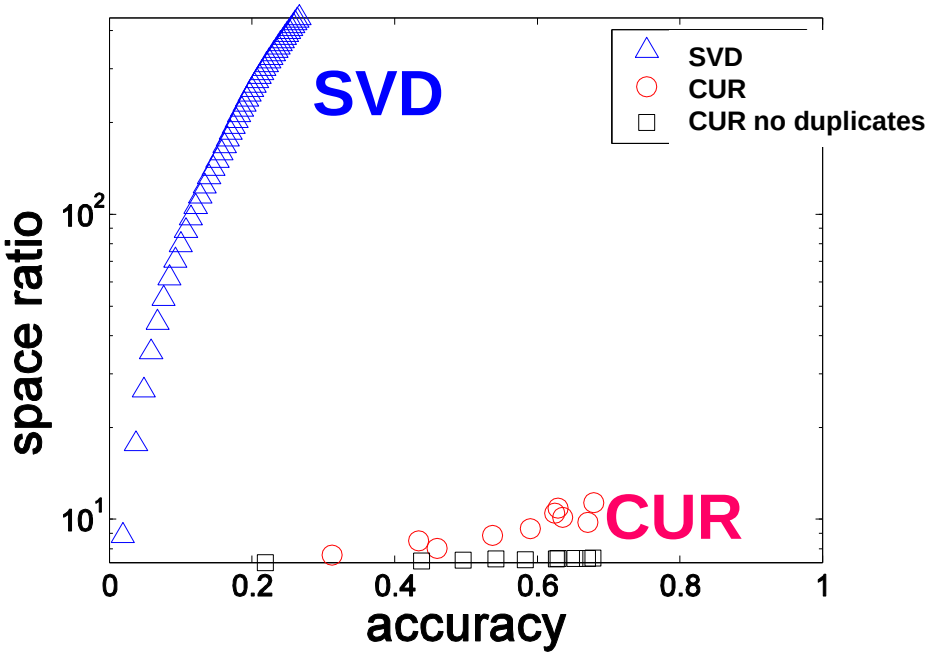

Klein und spärlich
SVD: A = U Σ V^T
Groß und spärlich
Groß und dicht
Klein und dicht
CUR: A = C U R
Groß und spärlich
Groß und spärlich

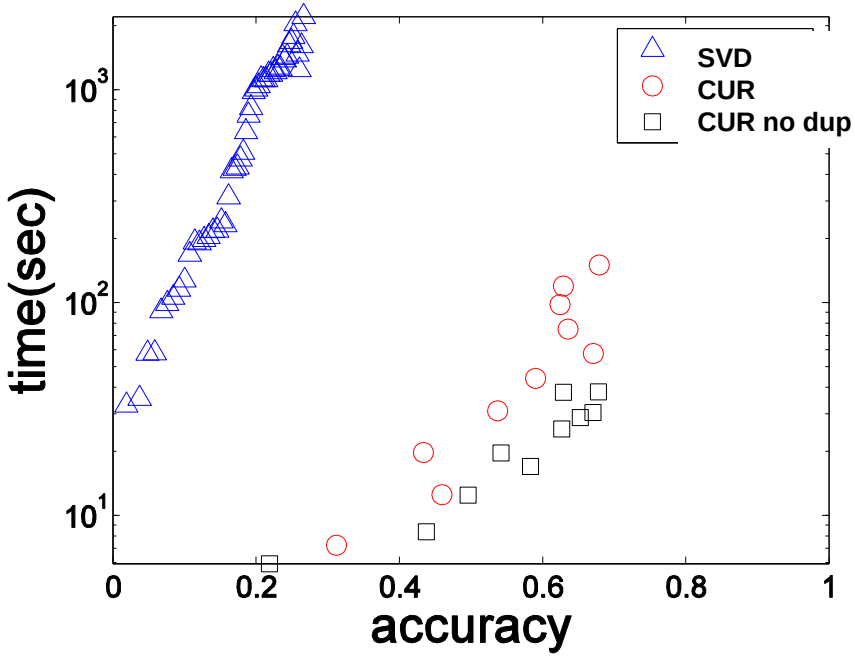
SVD vs. CUR: Experiment

- DBLP Daten
  - Bibliographische Sammlung wissenschaftlicher Publikationen im Bereich Informatik
  - Autor-Konferenz-Matrix $\boldsymbol{A}$
    - Einträge A_{ij}: Anzahl der Publikationen des Autors i zur Konferenz j
    - 428 000 Autoren (Zeilen), 3 659 Konferenzen (Spalten)
    - Matrix ist sehr groß und spärlich besetzt

- Dimensionsreduktion über SVD und CUR
  - Wie lange laufen die Algorithmen?
  - Wie groß ist der Fehler zwischen approximierter und tatsächlicher Matrix
  - Wie viel Speicherplatz wird benötigt?

SVD vs. CUR: Experiment





- **Accuracy:** 1 – relative Summe der quadrierten Fehler
- **Space ratio:** Benötigter Speicherplatz
- **Time: CPU Zeit**

Sun, Faloutsos: *Less is More: Compact Matrix Decomposition for Large Sparse Graphs*, SDM '07.
http://www.cs.cmu.edu/~jimeng/papers/SunSDM07.pdf

Referenzen, Beispiele, Übungen

Kapitel 11 aus „Mining of Massive Datasets“:
http://www.mmds.org/

Übungen:

- SVD: 11.3.1

Übung 11.3.1

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 4 & 3 \\ 0 & 2 & 4 \\ 1 & 3 & 5 \end{bmatrix}, M^T M = \begin{bmatrix} 36 & 37 & 38 \\ 37 & 49 & 61 \\ 38 & 61 & 84 \end{bmatrix}, MM^T = \begin{bmatrix} 14 & 26 & 22 & 16 & 22 \\ 26 & 50 & 46 & 28 & 40 \\ 22 & 46 & 50 & 20 & 32 \\ 16 & 28 & 20 & 20 & 26 \\ 22 & 40 & 32 & 26 & 35 \end{bmatrix}$$

Eigenvektoren $M^T M$:

$$\begin{bmatrix} -0.41 \\ -0.56 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.82 \\ 0.13 \\ -0.56 \end{bmatrix}, \ldots$$

Eigenvektoren MM^T:

$$\begin{bmatrix} -0.30 \\ -0.57 \\ -0.52 \\ -0.32 \\ -0.46 \end{bmatrix}, \begin{bmatrix} 0.16 \\ -0.03 \\ -0.74 \\ 0.51 \\ 0.41 \end{bmatrix}, \ldots$$

Eigenwerte:

$$153.6, \qquad 15.4, \qquad 0.0, \ldots$$

Übung 11.3.1

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 4 & 3 \\ 0 & 2 & 4 \\ 1 & 3 & 5 \end{bmatrix} = \underbrace{\begin{bmatrix} -0.30 & 0.16 \\ -0.57 & -0.03 \\ -0.52 & -0.74 \\ -0.32 & 0.51 \\ -0.46 & 0.41 \end{bmatrix}}_{U} \cdot \underbrace{\begin{bmatrix} 12.4 & 0 \\ 0 & 3.9 \end{bmatrix}}_{\Sigma} \cdot \underbrace{\begin{bmatrix} -0.41 & -0.56 & -0.72 \\ 0.82 & 0.13 & -0.56 \end{bmatrix}}_{V^T}$$

$$\begin{bmatrix} -0.30 & 0.16 \\ -0.57 & -0.03 \\ -0.52 & -0.74 \\ -0.32 & 0.51 \\ -0.46 & 0.41 \end{bmatrix} \cdot \begin{bmatrix} 12.4 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} -0.41 & -0.56 & -0.72 \\ 0.82 & 0.13 & -0.56 \end{bmatrix} = \begin{bmatrix} 1.5 & 2.1 & 2.6 \\ 2.9 & 4.0 & 5.1 \\ 2.6 & 3.6 & 4.6 \\ 1.6 & 2.3 & 2.9 \\ 2.3 & 3.2 & 4.1 \end{bmatrix}$$

Energie: $12.4^2 \approx 153.6$ und $12.4^2 + 3.9^2 = 169$, d.h. ca 91%