

CMU SCS

15-826: Multimedia Databases and Data Mining

Lecture #19: SVD - part II (case studies)
C. Faloutsos

CMU SCS

Must-read Material

- [Textbook](#) Appendix D

15-826 Copyright: C. Faloutsos (2009) 2

CMU SCS

Outline

Goal: 'Find **similar / interesting** things'

- Intro to DB
- ➔ • Indexing - similarity search
- Data Mining

15-826 Copyright: C. Faloutsos (2009) 3

CMU SCS

Indexing - Detailed outline

- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
- fractals
- text
- ➔ Singular Value Decomposition (SVD)
- multimedia
- ...

15-826 Copyright: C. Faloutsos (2009) 4

CMU SCS

SVD - Detailed outline

- Motivation
- Definition - properties
- Interpretation
- Complexity
- ➔ Case studies
- SVD properties
- Conclusions

15-826 Copyright: C. Faloutsos (2009) 5

CMU SCS

SVD - Case studies

- ➔ multi-lingual IR; LSI queries
- compression
- PCA - 'ratio rules'
- Karhunen-Lowe transform
- query feedbacks
- google/Kleinberg algorithms

15-826 Copyright: C. Faloutsos (2009) 6

CMU SCS

Case study - LSI

Q1: How to do queries with LSI?
 Q2: multi-lingual IR (english query, on spanish text?)

15-826 Copyright: C. Faloutsos (2009) 7

CMU SCS

Case study - LSI

Q1: How to do queries with LSI?
 Problem: Eg., find documents with 'data'

retrieval
 data inf. brain lung

$$\begin{matrix} \uparrow \\ \text{CS} \\ \downarrow \\ \uparrow \\ \text{MD} \\ \downarrow \end{matrix}
 \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}
 =
 \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}
 \times
 \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}
 \times
 \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-826 Copyright: C. Faloutsos (2009) 8

CMU SCS

Case study - LSI

Q1: How to do queries with LSI?
 A: map query vectors into 'concept space' – how?

retrieval
 data inf. brain lung

$$\begin{matrix} \uparrow \\ \text{CS} \\ \downarrow \\ \uparrow \\ \text{MD} \\ \downarrow \end{matrix}
 \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}
 =
 \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}
 \times
 \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}
 \times
 \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-826 Copyright: C. Faloutsos (2009) 9

CMU SCS

Case study - LSI

Q1: How to do queries with LSI?
 A: map query vectors into 'concept space' – how?

$q = \begin{bmatrix} \text{retrieval} & \text{inf.} & \text{brain} & \text{lung} \\ \text{data} & \downarrow & & \\ 1 & 0 & 0 & 0 \end{bmatrix}$

term2

term1

15-826 Copyright: C. Faloutsos (2009) 10

CMU SCS

Case study - LSI

Q1: How to do queries with LSI?
 A: map query vectors into 'concept space' – how?

$q = \begin{bmatrix} \text{retrieval} & \text{inf.} & \text{brain} & \text{lung} \\ \text{data} & \downarrow & & \\ 1 & 0 & 0 & 0 \end{bmatrix}$

term2

term1

A: inner product (cosine similarity) with each 'concept' vector v_i

15-826 Copyright: C. Faloutsos (2009) 11

CMU SCS

Case study - LSI

Q1: How to do queries with LSI?
 A: map query vectors into 'concept space' – how?

$q = \begin{bmatrix} \text{retrieval} & \text{inf.} & \text{brain} & \text{lung} \\ \text{data} & \downarrow & & \\ 1 & 0 & 0 & 0 \end{bmatrix}$

term2

term1

A: inner product (cosine similarity) with each 'concept' vector v_i

15-826 Copyright: C. Faloutsos (2009) 12

CMU SCS

Case study - LSI

compactly, we have:
 $q_{\text{concept}} = q \mathbf{V}$
 Eg:

$q =$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} \text{retrieval} & \text{lung} \\ \text{data} & \text{inf.} & \text{brain} & \text{lung} \\ \downarrow & & & \end{bmatrix}$	$\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}$	$=$	$\begin{bmatrix} \text{CS-concept} \\ \downarrow \\ 0.58 & 0 \end{bmatrix}$
-------	---	--	--	-----	---

term-to-concept similarities

15-826 Copyright: C. Faloutsos (2009) 13

CMU SCS

Case study - LSI

Drill: how would the document ('information', 'retrieval') be handled by LSI?

15-826 Copyright: C. Faloutsos (2009) 14

CMU SCS

Case study - LSI

Drill: how would the document ('information', 'retrieval') be handled by LSI? A: SAME:

$d_{\text{concept}} = d \mathbf{V}$ Eg:	$d =$	$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} \text{retrieval} & \text{lung} \\ \text{data} & \text{inf.} & \text{brain} & \text{lung} \\ \downarrow & & & \end{bmatrix}$	$\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}$	$=$	$\begin{bmatrix} \text{CS-concept} \\ \downarrow \\ 1.16 & 0 \end{bmatrix}$
--	-------	---	--	--	-----	---

term-to-concept similarities

15-826 Copyright: C. Faloutsos (2009) 15

CMU SCS

Case study - LSI

Observation: document ('information', 'retrieval') will be retrieved by query ('data'), although it does not contain 'data'!!

CS-concept

		retrieval				
	data	inf.	brain	lung		
$d =$	$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix}$	----->	$\begin{bmatrix} 1.16 & 0 \end{bmatrix}$			
$q =$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$	----->	$\begin{bmatrix} 0.58 & 0 \end{bmatrix}$			

15-826 Copyright: C. Faloutsos (2009) 16

CMU SCS

Case study - LSI

Q1: How to do queries with LSI?
 → Q2: multi-lingual IR (english query, on spanish text?)

15-826 Copyright: C. Faloutsos (2009) 17

CMU SCS

Case study - LSI

- Problem:
 - given many documents, translated to both languages (eg., English and Spanish)
 - answer queries across languages

15-826 Copyright: C. Faloutsos (2009) 18

CMU SCS

Case study - LSI

- Solution: ~ LSI

			retrieval		informacion	
		inf.	brain	lung	datos	
	data					
↑	CS	↓	↓	↓	↓	↓
↓						
↑						
↓						
↑						
↓						
↑	MD	↓	↓	↓	↓	↓
↓						
↑						
↓						
↑						
↓						

15-826 Copyright: C. Faloutsos (2009) 19

CMU SCS

Case study - LSI

- Solution: ~ LSI

			retrieval		informacion	
		inf.	brain	lung	datos	
	data					
↑	CS	↓	↓	↓	↓	↓
↓						
↑						
↓						
↑						
↓						
↑	MD	↓	↓	↓	↓	↓
↓						
↑						
↓						
↑						
↓						

15-826 Copyright: C. Faloutsos (2009) 20

CMU SCS

SVD - Case studies

- multi-lingual IR; LSI queries
- ➔ • compression
- PCA - 'ratio rules'
- Karhunen-Lowe transform
- query feedbacks
- google/Kleinberg algorithms

15-826 Copyright: C. Faloutsos (2009) 21

CMU SCS

Case study: compression

[Korn+97]

Problem:

- given a matrix
- compress it, but maintain 'random access'

(surprisingly, its solution leads to data mining and visualization...)

15-826 Copyright: C. Faloutsos (2009) 22

CMU SCS

Problem - specs

- ~10**6 rows; ~10**3 columns; no updates;
- random access to any cell(s) ; small error: OK

day	Mo	Tu	We	Th	Fri	Sa	Su
customer	7/10/96	7/11/96	7/12/96	7/13/96	7/14/96		
ABC Inc.	1	1	1	0	0		
DEF Inc.	2	2	2	0	0		
GHI Inc.	1	1	1	0	0		
KLM Co.	5	5	5	0	0		
Smith	0	0	0	2	2		
Johnson	0	0	0	3	3		
Thompson	0	0	0	1	1		

15-826 Copyright: C. Faloutsos (2009) 23

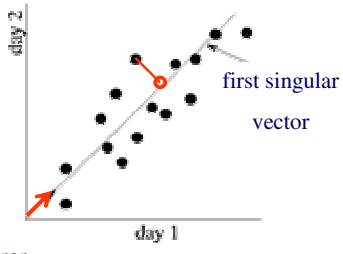
CMU SCS

Idea

15-826 Copyright: C. Faloutsos (2009) 24

CMU SCS

SVD - reminder



• space savings: 2:1
• minimum RMS error

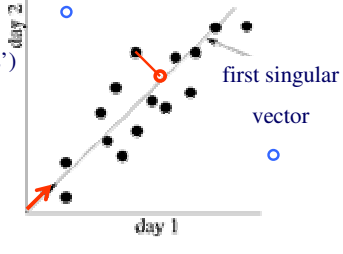
15-826 Copyright: C. Faloutsos (2009) 25

The figure shows a scatter plot with 'day 1' on the x-axis and 'day 2' on the y-axis. A series of black dots represents data points, showing a positive linear correlation. A grey line of best fit is drawn through the points. A red arrow points from the origin towards the line, labeled 'first singular vector'. A red circle is placed on the line, and a dashed line connects it to the origin, illustrating the orthogonal distance from the origin to the line.

CMU SCS

Case study: compression

outliers?
A: treat separately
(SVD with 'Deltas')



first singular vector

15-826 Copyright: C. Faloutsos (2009) 26

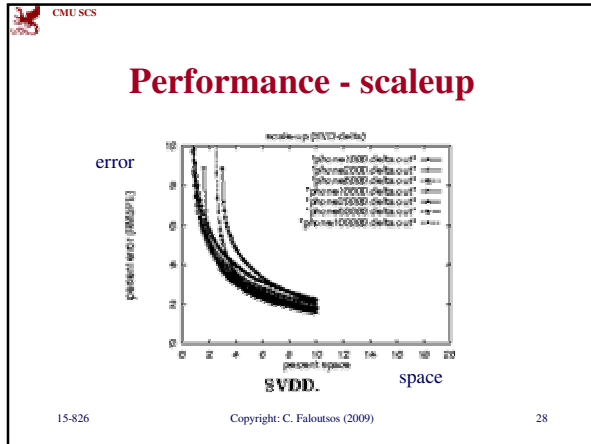
The figure is similar to the previous one, but includes two blue circles representing outliers. One outlier is located at a high 'day 2' value and low 'day 1' value, while the other is at a low 'day 2' value and high 'day 1' value. The line of best fit and the first singular vector arrow are still present.

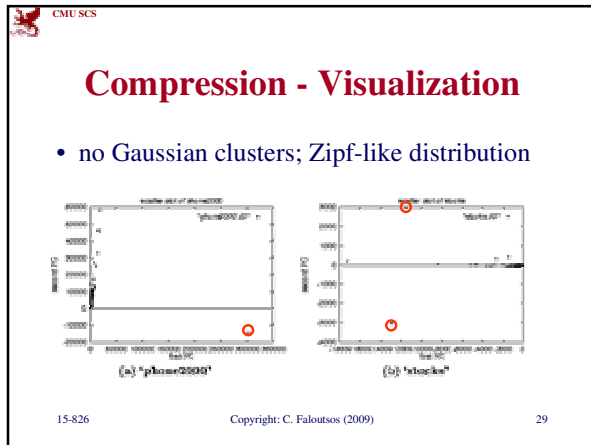
CMU SCS

Compression - Performance

- 3 pass algo (-> scalability) (HOW?)
- random cell(s) reconstruction
- 10:1 compression with < 2% error

15-826 Copyright: C. Faloutsos (2009) 27





- ### SVD - Case studies
- multi-lingual IR; LSI queries
 - compression
 - ➔ PCA - 'ratio rules'
 - Karhunen-Lowe transform
 - query feedbacks
 - google/Kleinberg algorithms

CMU SCS

PCA - 'Ratio Rules'

[Korn+00]
 Typically: 'Association Rules' (eg.,
 {bread, milk} -> {butter})

But:

- which set of rules is 'better'?
- how to reconstruct missing/corrupted values?
- need binary/bucketized values

15-826 Copyright: C. Faloutsos (2009) 31

CMU SCS

PCA - 'Ratio Rules'

Idea: try to find 'concepts':

- singular vectors dictate rules about ratios:
 bread:milk:butter = 2:4:3

15-826 Copyright: C. Faloutsos (2009) 32

CMU SCS

PCA - 'Ratio Rules'

Identical to PCA = Principal Components Analysis

- Q1: which set of rules is 'better'?
- ➔ - Q2: how to reconstruct missing/corrupted values?
- Q3: is there need for binary/bucketized values?
- Q4: how to interpret the rules (= 'principal components')?

15-826 Copyright: C. Faloutsos (2009) 33

CMU SCS

PCA - 'Ratio Rules'

Q2: how to reconstruct missing/corrupted values?

Eg:

- rule: bread:milk = 3:4
- a customer spent \$6 on bread - how about milk?

15-826 Copyright: C. Faloutsos (2009) 34

CMU SCS

PCA - 'Ratio Rules'

pictorially:

15-826 Copyright: C. Faloutsos (2009) 35

CMU SCS

PCA - 'Ratio Rules'

harder cases: overspecified/underspecified

over-specified:

- milk:bread:butter = 1:2:3
- a customer got
 - \$2 bread and \$4 milk
- how much milk?

Answer: minimize distance between 'feasible' and 'expected' values (using SVD...)

15-826 Copyright: C. Faloutsos (2009) 36

CMU SCS

PCA - 'Ratio Rules'

harder cases: underspecified

15-826 Copyright: C. Faloutsos (2009) 37

CMU SCS

PCA - 'Ratio Rules'

bottom line: we can reconstruct any count of missing values

This is very useful:

- can spot outliers (how?)
- can measure the 'goodness' of a set of rules (how?)

15-826 Copyright: C. Faloutsos (2009) 38

CMU SCS

PCA - 'Ratio Rules'

Identical to PCA = Principal Components Analysis

- ➔ - Q1: which set of rules is 'better'?
- ✓ - Q2: how to reconstruct missing/corrupted values?
- Q3: is there need for binary/bucketized values?
- Q4: how to interpret the rules (= 'principal components')?

15-826 Copyright: C. Faloutsos (2009) 39

CMU SCS

PCA - 'Ratio Rules'

- Q1: which set of rules is 'better'?
- A: the ones that needs the fewest outliers:
 - pretend we don't know a value (eg., \$ of 'Smith' on 'bread')
 - reconstruct it
 - and sum up the squared errors, for all our entries
- (other answers are also reasonable)

15-826 Copyright: C. Faloutsos (2009) 40

CMU SCS

PCA - 'Ratio Rules'

Identical to PCA = Principal Components Analysis

- ✓ – Q1: which set of rules is 'better'?
- ✓ – Q2: how to reconstruct missing/corrupted values?
- ➔ – Q3: is there need for binary/bucketized values?
- Q4: how to interpret the rules (= 'principal components')?

15-826 Copyright: C. Faloutsos (2009) 41

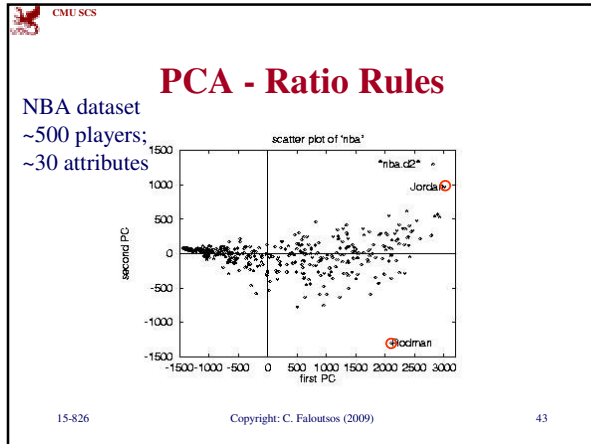
CMU SCS

PCA - 'Ratio Rules'

Identical to PCA = Principal Components Analysis

- ✓ – Q1: which set of rules is 'better'?
- ✓ – Q2: how to reconstruct missing/corrupted values?
- ✓ – Q3: is there need for binary/bucketized values? **NO**
- ➔ – Q4: how to interpret the rules (= 'principal components')?

15-826 Copyright: C. Faloutsos (2009) 42



-
- ### PCA - Ratio Rules
- PCA: get singular vectors v_1, v_2, \dots
 - ignore entries with small abs. value
 - try to interpret the rest
- 15-826 Copyright: C. Faloutsos (2009) 44

PCA - Ratio Rules

NBA dataset - V matrix (term to 'concept' similarities)

field	RR_1	RR_2	RR_3
minutes played	.308	-.4	
field goals			
goal attempts	.406	.199	
points		-.489	.602
total rebounds			-.486
assists			-.07
steals			

v_1

15-826 Copyright: C. Faloutsos (2009) 45

CMU SCS

Ratio Rules - example

- RR1: minutes:points = 2:1
- corresponding concept?

v1

15-826 Copyright: C. Faloutsos (2009) 46

CMU SCS

Ratio Rules - example

- RR1: minutes:points = 2:1
- CO
- CO

15-826 Copyright: C. Faloutsos (2009) 47

CMU SCS

Ratio Rules - example

- RR1: minutes:points = 2:1
- CO
- CO

v1

15-826 Copyright: C. Faloutsos (2009)

CMU SCS

Ratio Rules - example

- RR1: minutes:points = 2:1
- corresponding concept?
- A: 'goodness' of player

15-826 Copyright: C. Faloutsos (2009) 49

CMU SCS

Ratio Rules - example

- RR2: points:rebounds negatively correlated(!)

field	RR ₁	RR ₂	RR ₃
minutes played	.808	-.4	
field goals			
goal attempts			
points	.406	.199	
total rebounds		-.489	.602
assists			-.486
steals			-.07

15-826 Copyright: C. Faloutsos (2009) 50

CMU SCS

Ratio Rules - example

- RR2: points:rebounds negatively correlated(!) - concept?

15-826 Copyright: C. Faloutsos (2009) 51

CMU SCS

Ratio Rules - example

- RR2: points: rebounds negatively correlated(!) - concept?
- A: position: offensive/defensive

15-826 Copyright: C. Faloutsos (2009) 52

CMU SCS

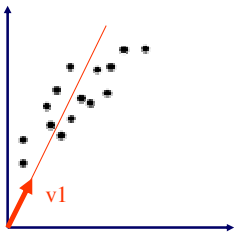
SVD - Case studies

- multi-lingual IR; LSI queries
- compression
- PCA - 'ratio rules'
- ➔ • Karhunen-Lowe transform
- query feedbacks
- google/Kleinberg algorithms

15-826 Copyright: C. Faloutsos (2009) 53

CMU SCS

K-L transform



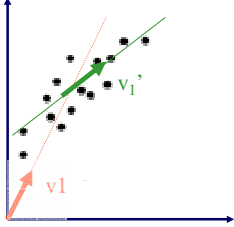
[Duda & Hart]; [Fukunaga]

A subtle point:
SVD will give vectors that go through the origin

15-826 Copyright: C. Faloutsos (2009) 54

CMU SCS

K-L transform

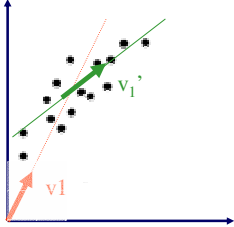


A subtle point:
SVD will give vectors that go through the origin
Q: how to find v_1' ?

15-826 Copyright: C. Faloutsos (2009) 55

CMU SCS

K-L transform



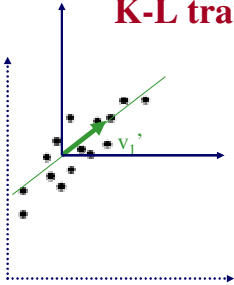
A subtle point:
SVD will give vectors that go through the origin
Q: how to find v_1' ?

A: 'centered' PCA, ie.,
move the origin to center of gravity

15-826 Copyright: C. Faloutsos (2009) 56

CMU SCS


K-L transform



A subtle point:
SVD will give vectors that go through the origin
Q: how to find v_1' ?

A: 'centered' PCA, ie.,
move the origin to center of gravity
and THEN do SVD


15-826 Copyright: C. Faloutsos (2009) 57

 CMU SCS

K-L transform

- How to 'center' a set of vectors (= data matrix)?
- What is the covariance matrix?
- A: see textbook
- ('whitening transformation')


15-826 Copyright: C. Faloutsos (2009) 58

 CMU SCS

Conclusions

- SVD: popular for dimensionality reduction / compression
- SVD is the 'engine under the hood' for PCA (principal component analysis)
- ... as well as the Karhunen-Lowe transform
- (and there is more to come ...)


15-826 Copyright: C. Faloutsos (2009) 59

 CMU SCS

References

- Duda, R. O. and P. E. Hart (1973). Pattern Classification and Scene Analysis. New York, Wiley.
- Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition, Academic Press.
- Jolliffe, I. T. (1986). Principal Component Analysis, Springer Verlag.


15-826 Copyright: C. Faloutsos (2009) 60

 CMU SCS

References

- Korn, F., H. V. Jagadish, et al. (May 13-15, 1997). Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences. ACM SIGMOD, Tucson, AZ.
- Korn, F., A. Labrinidis, et al. (1998). Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining. VLDB, New York, NY.

15-826 Copyright: C. Faloutsos (2009) 61

 CMU SCS

References

- Korn, F., A. Labrinidis, et al. (2000). "Quantifiable Data Mining Using Ratio Rules." VLDB Journal 8(3-4): 254-266.
- Press, W. H., S. A. Teukolsky, et al. (1992). Numerical Recipes in C, Cambridge University Press.

15-826 Copyright: C. Faloutsos (2009) 62
