


CMU SCS

15-826: Multimedia Databases and Data Mining

Lecture #19: SVD – part II – applications
C. Faloutsos



CMU SCS

Must-read Material

- [MM Textbook](#) Appendix D

15-826 Copyright: C. Faloutsos (2017) 2




CMU SCS

Outline

Goal: ‘Find **similar / interesting** things’

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

15-826 Copyright: C. Faloutsos (2017) 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 (2017) 4

CMU SCS

SVD - Detailed outline

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

15-826 Copyright: C. Faloutsos (2017) 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 (2017) 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 (2017) 7

CMU SCS

Case study - LSI

Q1: How to do queries with LSI?

Problem: Eg., find documents with 'data'

$$\begin{array}{c} \text{data} \\ \updownarrow \text{CS} \\ \text{MD} \end{array}
 \begin{array}{c} \text{retrieval} \\ \text{inf.} \downarrow \text{brain} \text{ lung} \end{array}
 \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}
 =
 \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}
 \times
 \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}
 \times
 \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

15-826 Copyright: C. Faloutsos (2017) 8

Case study - LSI

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

retrieval
data inf. ↓ brain lung

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

15-826 Copyright: C. Faloutsos (2017) 9

Case study - LSI

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

retrieval
data inf. ↓ brain lung

$q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$

term2

term1

15-826 Copyright: C. Faloutsos (2017) 10

Case study - LSI

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

retrieval
data inf. ↓ brain lung

$q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$

term2

term1

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

15-826 Copyright: C. Faloutsos (2017) 11

Case study - LSI

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

retrieval
data inf. ↓ brain lung

$q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$

term2

term1

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

15-826 Copyright: C. Faloutsos (2017) 12

Case study - LSI

compactly, we have:

$$q V = q_{\text{concept}}$$

Eg:

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

term-to-concept similarities

15-826 Copyright: C. Faloutsos (2017) 13

Case study - LSI

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

15-826 Copyright: C. Faloutsos (2017) 14

Case study - LSI

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

$$d_{\text{concept}} = d V$$

Eg:

$$d = \begin{bmatrix} \text{data} & \text{inf} & \text{retrieval} & \text{brain} & \text{lung} \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix} = \begin{bmatrix} 1.16 & 0 \end{bmatrix}$$

term-to-concept similarities

15-826 Copyright: C. Faloutsos (2017) 15

Case study - LSI

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

$$\begin{array}{ccc} \begin{array}{c} \text{data} \quad \text{inf} \quad \text{retrieval} \quad \text{brain} \quad \text{lung} \\ \text{d} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix} \end{array} & \xrightarrow{\quad} & \begin{array}{c} \text{CS-concept} \\ \downarrow \\ \begin{bmatrix} 1.16 & 0 \end{bmatrix} \end{array} \\ \begin{array}{c} \text{data} \quad \text{inf} \quad \text{retrieval} \quad \text{brain} \quad \text{lung} \\ \text{q} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} & \xrightarrow{\quad} & \begin{array}{c} \text{CS-concept} \\ \downarrow \\ \begin{bmatrix} 0.58 & 0 \end{bmatrix} \end{array} \end{array}$$

15-826 Copyright: C. Faloutsos (2017) 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 (2017) 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 (2017) 18

CMU SCS

Case study - LSI

- Solution: ~ LSI

Diagram illustrating the LSI solution for multi-lingual IR:

CS (Cross-Similarity) and MD (Matrix Decomposition) are shown as vertical arrows on the left, indicating the relationship between the two matrices.

data	retrieval	inf.	brain	lung	information	datos			
1	1	1	0	0	1	1	1	0	0
2	2	2	0	0	1	2	2	0	0
1	1	1	0	0	1	1	1	0	0
5	5	5	0	0	5	5	4	0	0
0	0	0	2	2	0	0	0	2	2
0	0	0	3	3	0	0	0	2	3
0	0	0	1	1	0	0	0	1	1

15-826 Copyright: C. Faloutsos (2017) 19

CMU SCS

SVD - Case studies

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

15-826 Copyright: C. Faloutsos (2017) 20

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 (2017) 21

CMU SCS

Problem - specs

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

day	Wc	Th	Fr	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 Ltd.	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 (2017) 22

CMU SCS

Idea

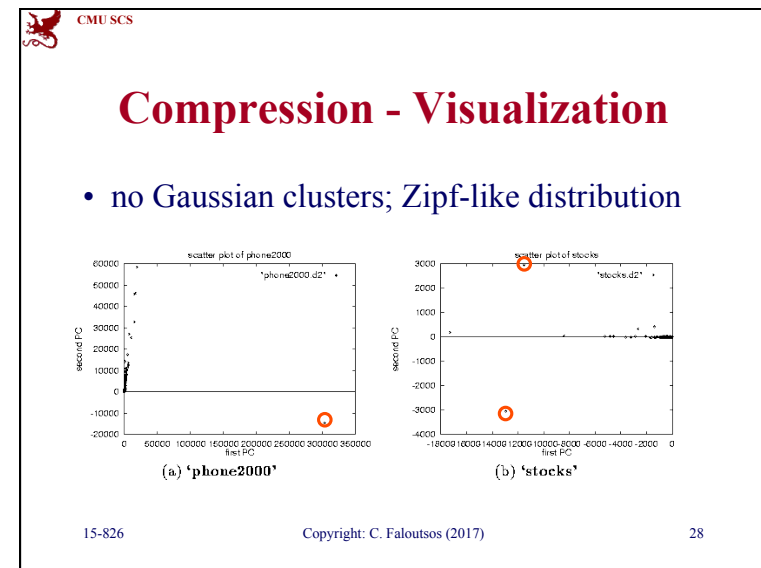
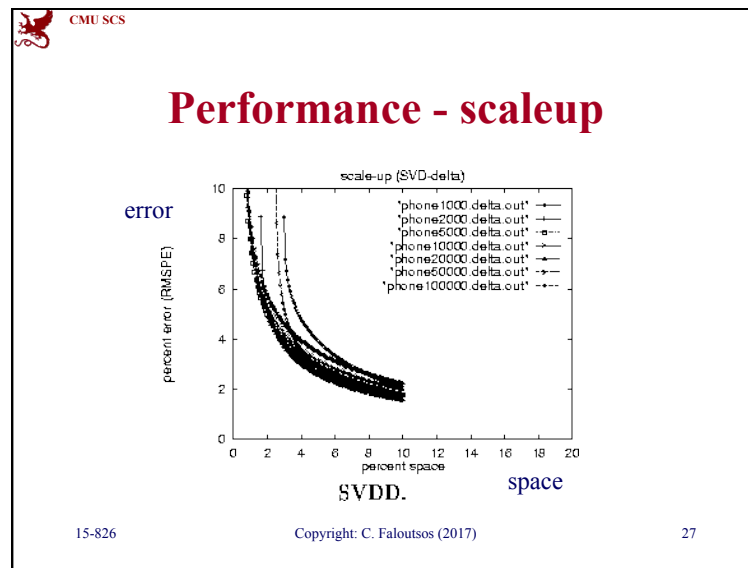
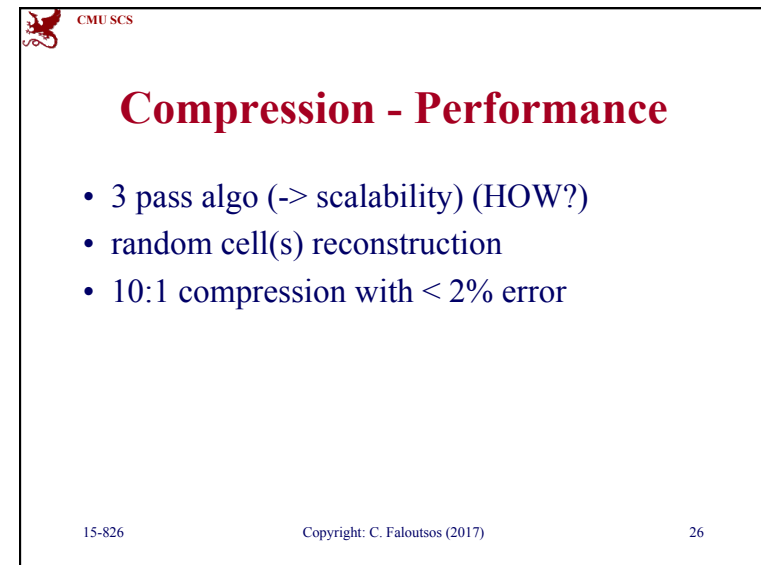
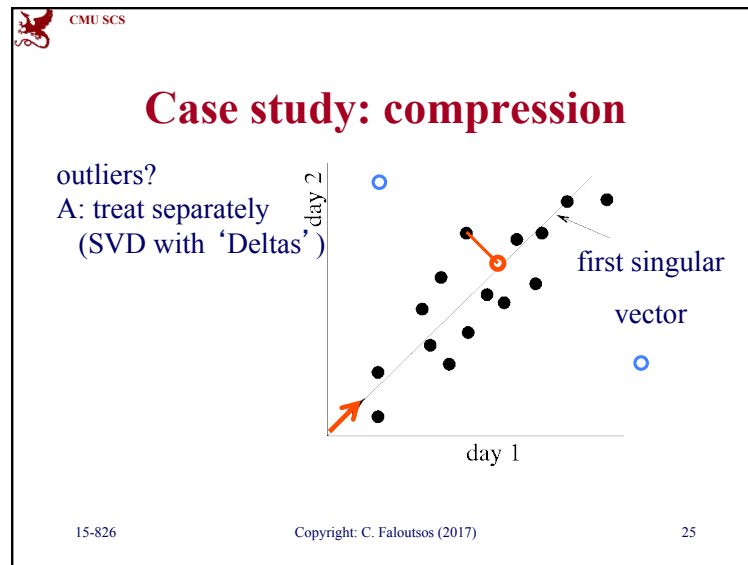
15-826 Copyright: C. Faloutsos (2017) 23

CMU SCS

SVD - reminder

- space savings: 2:1
- minimum RMS error

15-826 Copyright: C. Faloutsos (2017) 24



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 (2017) 29

CMU SCS

PCA - ‘Ratio Rules’

[Korn+00]
Typically: ‘*Association Rules*’ (eg.,
{bread, milk} -> {butter})
But, can we discover more details? like:
\$-bread : \$-milk : \$-butter ~ \$2 : \$4 : \$3

15-826 Copyright: C. Faloutsos (2017) 30

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 (2017) 31

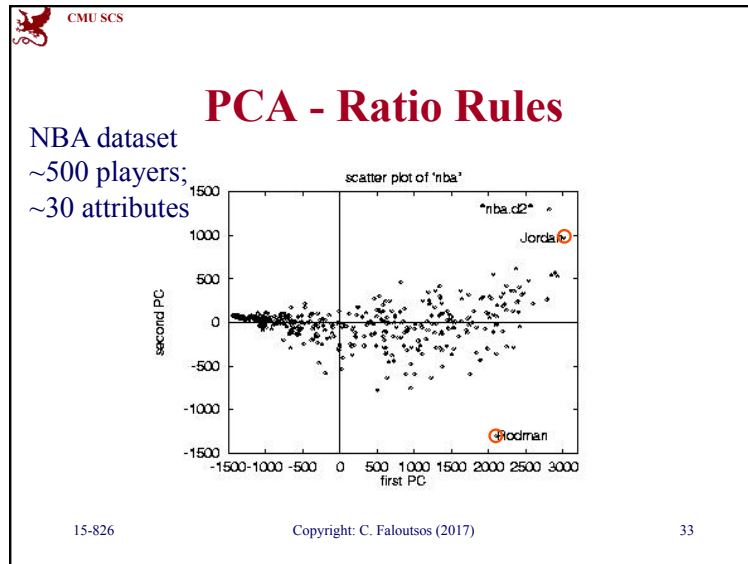
CMU SCS

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 (2017) 32



CMU SCS

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 (2017) 34

CMU SCS

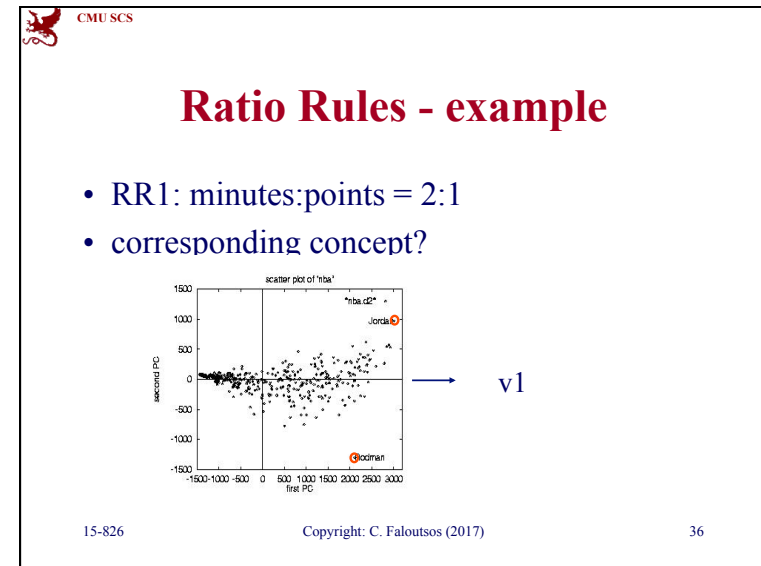
PCA - Ratio Rules

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

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

v_1

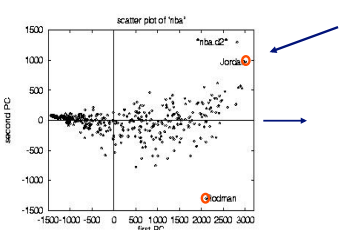
15-826 Copyright: C. Faloutsos (2017) 35



CMU SCS

Ratio Rules - example

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

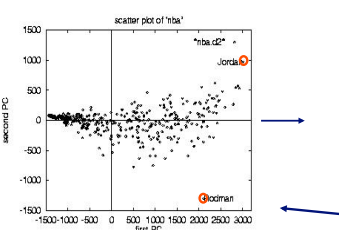


15-826 Copyright: C. Faloutsos (2017) 37

CMU SCS

Ratio Rules - example

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



15-826 Copyright: C. Faloutsos (2017)

CMU SCS

Ratio Rules - example

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

15-826 Copyright: C. Faloutsos (2017) 39

CMU SCS

Ratio Rules - example

- RR2: points:rebounds negatively correlated(!)

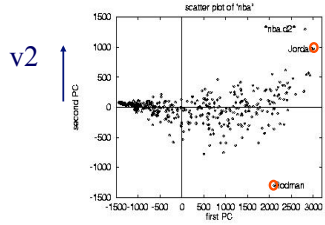
field	RR ₁	RR ₂	RR ₃
minutes played	.808	-.4	
field goals			
goal attempts			
points	.406	.199	
total rebounds		-.489	.602
assists			-.486
steals			-.07

15-826 Copyright: C. Faloutsos (2017) 40

CMU SCS

Ratio Rules - example

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



15-826 Copyright: C. Faloutsos (2017) 41

CMU SCS

Ratio Rules - example

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

15-826 Copyright: C. Faloutsos (2017) 42

CMU SCS

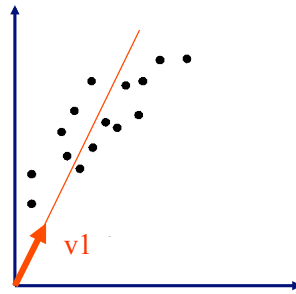
SVD - Case studies

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

15-826 Copyright: C. Faloutsos (2017) 43

CMU SCS

K-L transform



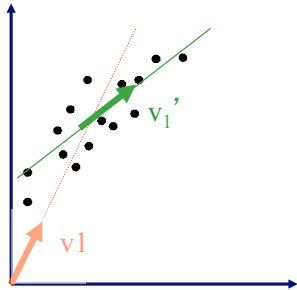
[Duda & Hart]; [Fukunaga]

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

15-826 Copyright: C. Faloutsos (2017) 44

CMU SCS

K-L transform

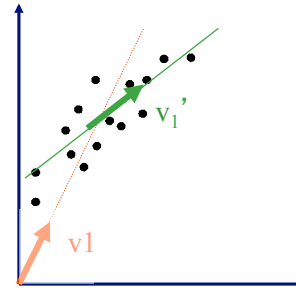


A subtle point:
SVD will give vectors that go through the origin
Q: how to find v_1' ?

15-826 Copyright: C. Faloutsos (2017) 45

CMU SCS

K-L transform



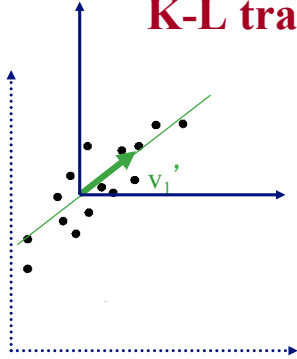
A subtle point:
SVD will give vectors that go through the origin
Q: how to find v_1' ?

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

15-826 Copyright: C. Faloutsos (2017) 46

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 (2017) 47

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 (2017) 48

 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 (2017) 49

 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 (2017) 50

 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 (2017) 51

 CMU SCS

References

- [Korn+, '00] 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 (2017) 52