

CarnegieMellon

15-826: Multimedia Databases and Data Mining

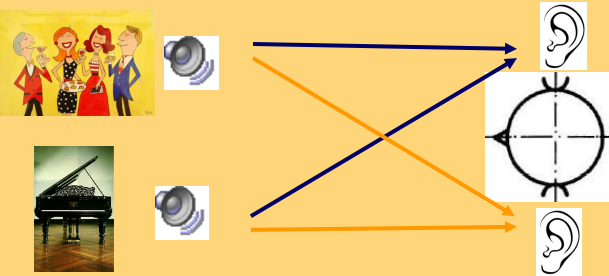
Lecture #20:
Independent Component Analysis (ICA)
Christos Faloutsos

1

CarnegieMellon

Problem: BSS

- two sound sources in a cocktail party – separate them




= “blind source separation”
(= unknown sources, unknown mixing)

15-826 Copyright (c) 2019 C. Faloutsos #2

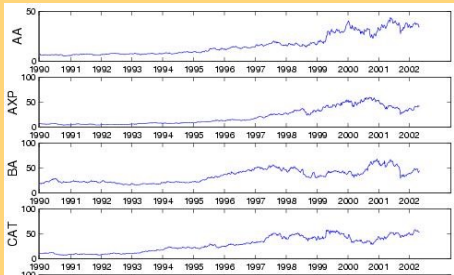
2

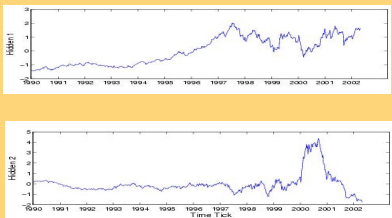
CarnegieMellon



Problem

Q: how to extract **sparse** hidden/latent variables?






15-826
Copyright (c) 2019 C. Faloutsos
3

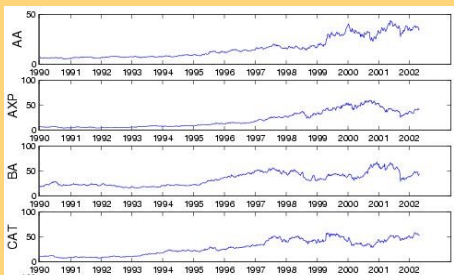
3

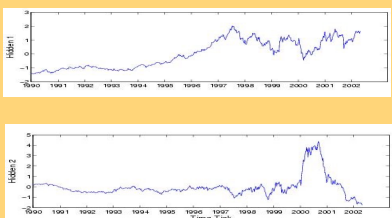
CarnegieMellon

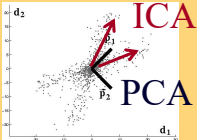


Answer

Q: how to extract **sparse** hidden/latent variables?
A: ~~SVD~~ ICA








15-826
Copyright (c) 2019 C. Faloutsos
4

4

CarnegieMellon

Must-read Material



- *AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases*, **Jia-Yu Pan**, Hiroyuki Kitagawa, Christos Faloutsos and Masafumi Hamamoto, PAKDD 2004, Sydney, Australia

15-826 Copyright (c) 2019 C. Faloutsos #5

5

CarnegieMellon

Outline

- Motivation
- Formulation
- PCA and ICA
- Example applications
- Conclusion

15-826 Copyright (c) 2019 C. Faloutsos #6

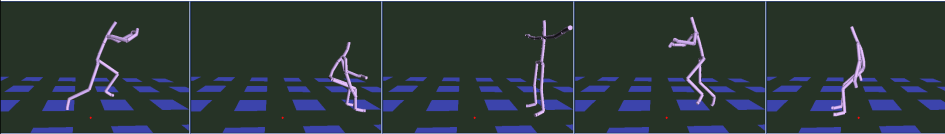
6

CarnegieMellon

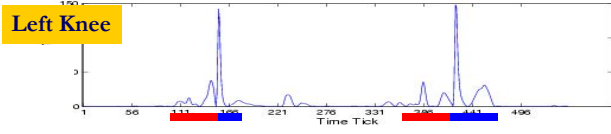
Motivation:

(Q1) Find patterns in data

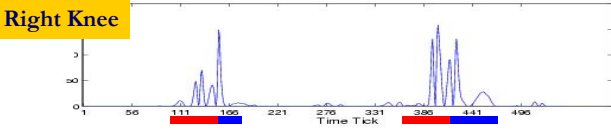
- Motion capture data: broad jumps



Left Knee



Right Knee



Take-off

Landing

15-826 Copyright (c) 2019 C. Faloutsos #7

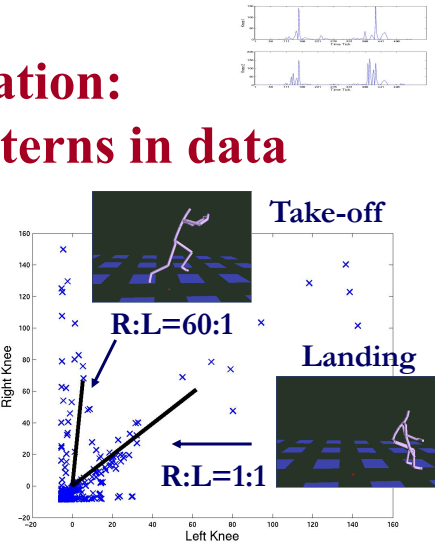
7

CarnegieMellon

Motivation:

(Q1) Find patterns in data

- Human would say
 - Pattern 1: along diagonal
 - Pattern 2: along vertical axis
- How to find these automatically?



Each point is the measurement at a time tick (total 550 points).

15-826 Copyright (c) 2019 C. Faloutsos #8

8

CarnegieMellon

Motivation: (Q2) Find hidden variables

Stock prices

Hidden variables (= 'topics' = concepts)

15-826 Copyright (c) 2019 C. Faloutsos #9

9

CarnegieMellon

(Q3): Topic discovery on text streams

- Data: CNN headline news (Jan.-Jun. 1998)
- Documents of 10 topics in one single text stream
 - FIND: the document boundaries
 - AND: the terms of each topic

15-826 Copyright (c) 2019 C. Faloutsos #10

10

CarnegieMellon

Outline

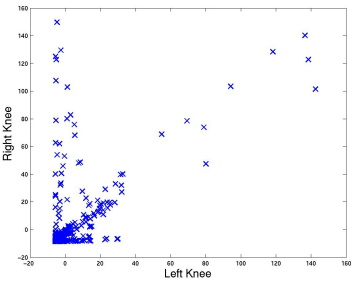
- Motivation
- ➔ • Formulation
- PCA and ICA
- Example applications
- Conclusion

15-826 Copyright (c) 2019 C. Faloutsos #11

11

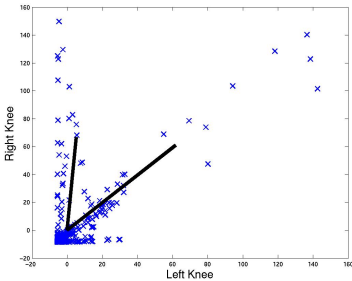
CarnegieMellon

Formulation: Finding patterns



Given n data points,
each with m attributes.

➔



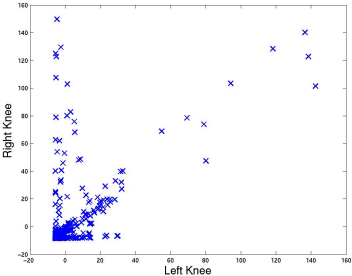
Find patterns that describe
data properties the best.

15-826 Copyright (c) 2019 C. Faloutsos #12

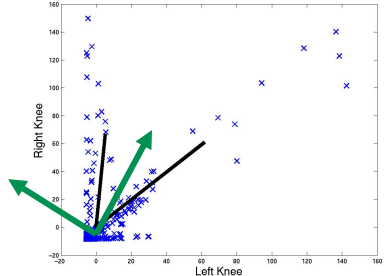
12

CarnegieMellon

Formulation: Finding patterns



Given n data points,
each with m attributes.



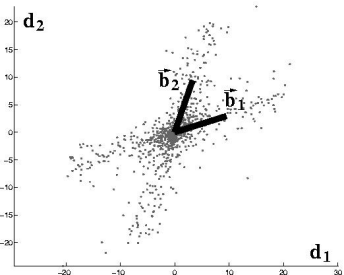
**SVD/PCA: ORTHOGONAL
vectors**

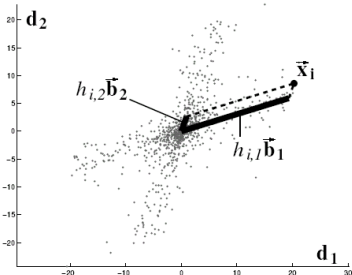
15-826
Copyright (c) 2019 C. Faloutsos
#13

13

CarnegieMellon

Linear representation





- Find vectors that describe the data set the best.
- Each point: linear combination of the vectors (patterns):

$$\vec{x}_i = h_{i,1} \vec{b}_1 + h_{i,2} \vec{b}_2$$

15-826
Copyright (c) 2019 C. Faloutsos
#14

14

CarnegieMellon

Patterns as data “vocabulary”

Good pattern
≈ sparse coding

b_1 alone, can
describe x_i .

(a) ICA representation of \vec{x}_i

$\vec{x}_i = h_{i,1}\vec{b}_1 + h_{i,2}\vec{b}_2$

15-826
Copyright (c) 2019 C. Faloutsos
#15

15

CarnegieMellon

PCA: first step of ICA

PCA finds the hyperplane.

ICA finds the correct patterns.

15-826
Copyright (c) 2019 C. Faloutsos
#16

16

CarnegieMellon

Software

- Open source software: ‘fastICA’
<http://research.ics.aalto.fi/ica/fastica/>
- Or ‘autosplit’ :
www.cs.cmu.edu/~jypan/software/autosplit_cmu.tar.gz

15-826 Copyright (c) 2019 C. Faloutsos #17

17

CarnegieMellon

References

- Aapo Hyvärinen, Juha Karhunen, Erkki Oja: *Independent Component Analysis*, John Wiley & Sons, 2001



15-826 Copyright (c) 2019 C. Faloutsos #18

18

CarnegieMellon

Outline

- Motivation
- Formulation
- PCA and ICA
- Example applications
 - ➔ – Hidden variables in stock prices
 - Find topics in documents
- Conclusion

15-826 Copyright (c) 2019 C. Faloutsos #19

19

CarnegieMellon

Motivation: Find hidden variables

Alcoa

American Express

Boeing

Caterpillar

Citi Group

Find common hidden variables, and weights.

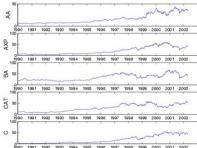
Dow Jones Industrial Average

15-826 Copyright (c) 2019 C. Faloutsos 20

20

CarnegieMellon

ICA: Like SVD, but sparse U



\mathcal{X}

 $\mathcal{X} \sim \mathcal{U} \Sigma \mathcal{V}^T$

1st behavior

2nd behavior

Participation weight of row i to behavior j

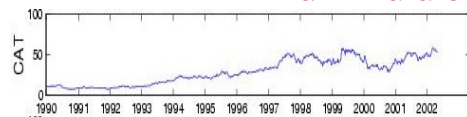
$\mathcal{U} \quad \Sigma \quad \mathcal{V}^T$

15-826
Copyright (c) 2019 C. Faloutsos
21

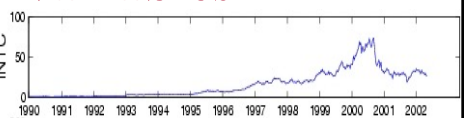
21

CarnegieMellon

Motivation: Find hidden variables = behaviors



CATERPILLAR



INTEL


B_{1,CAT}

B_{1,INTC}

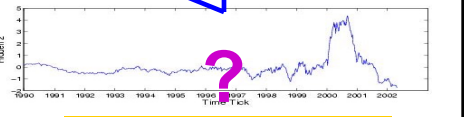
?

B_{2,CAT}

B_{2,INTC}



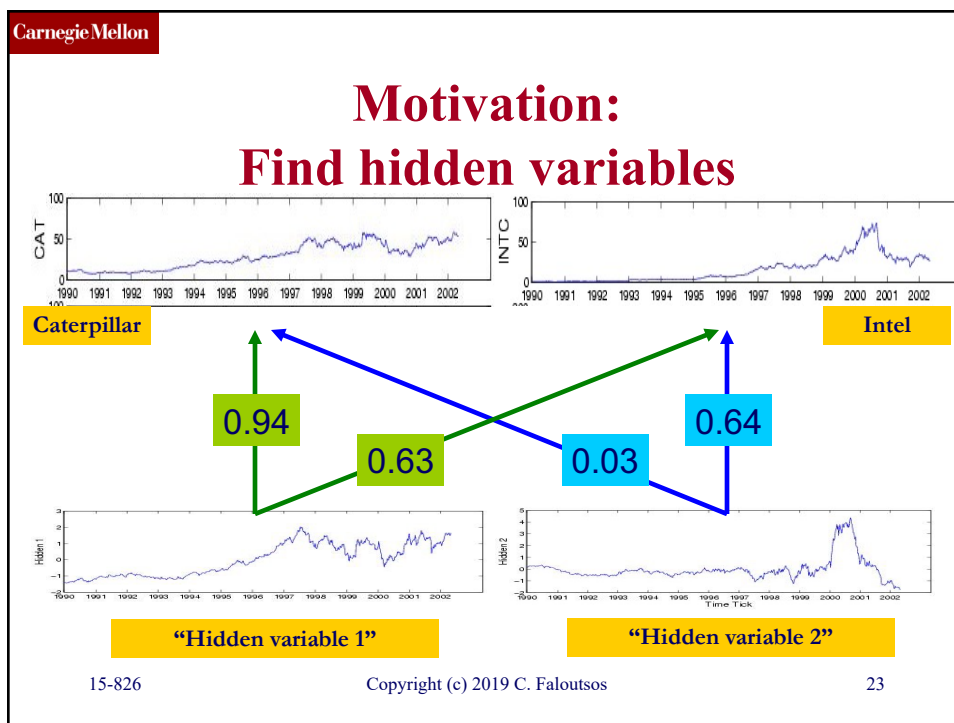
Hidden variable 1



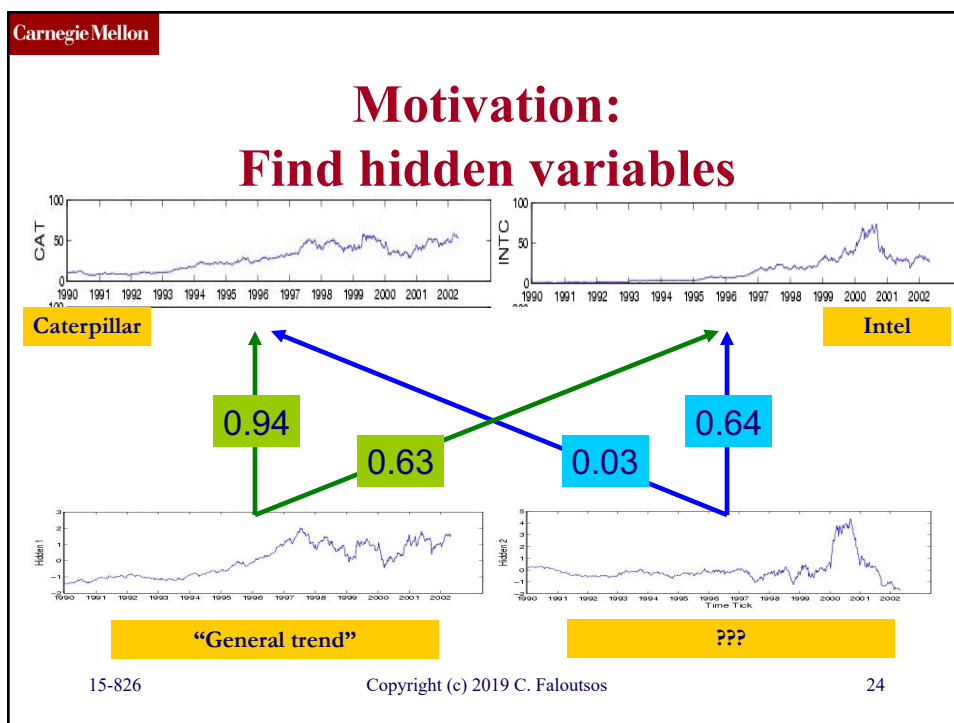
Hidden variable 2

15-826
Copyright (c) 2019 C. Faloutsos
22

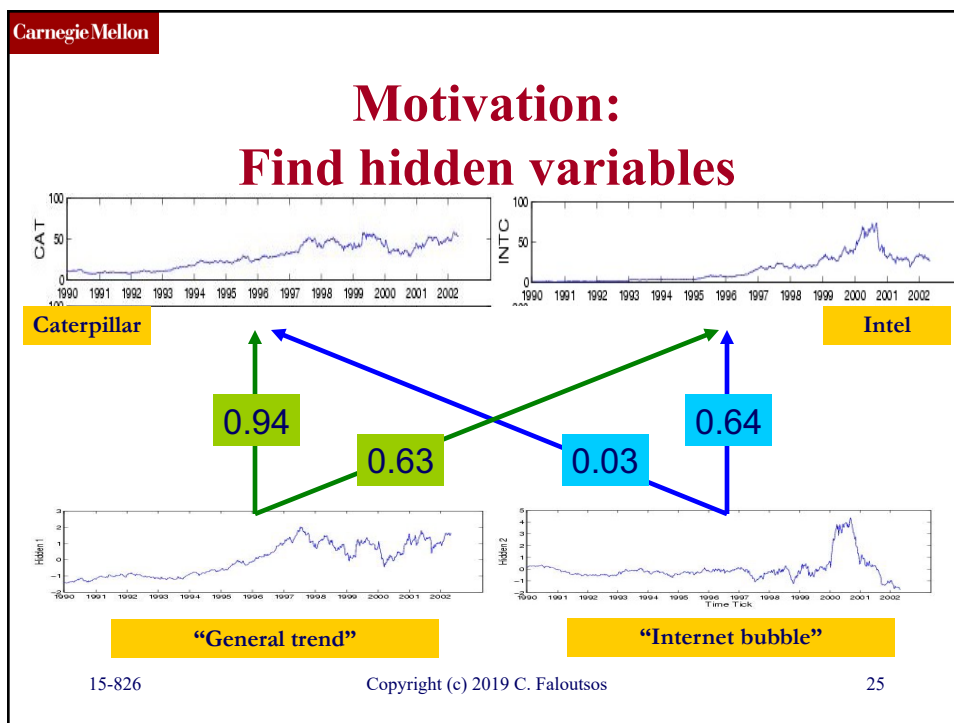
22



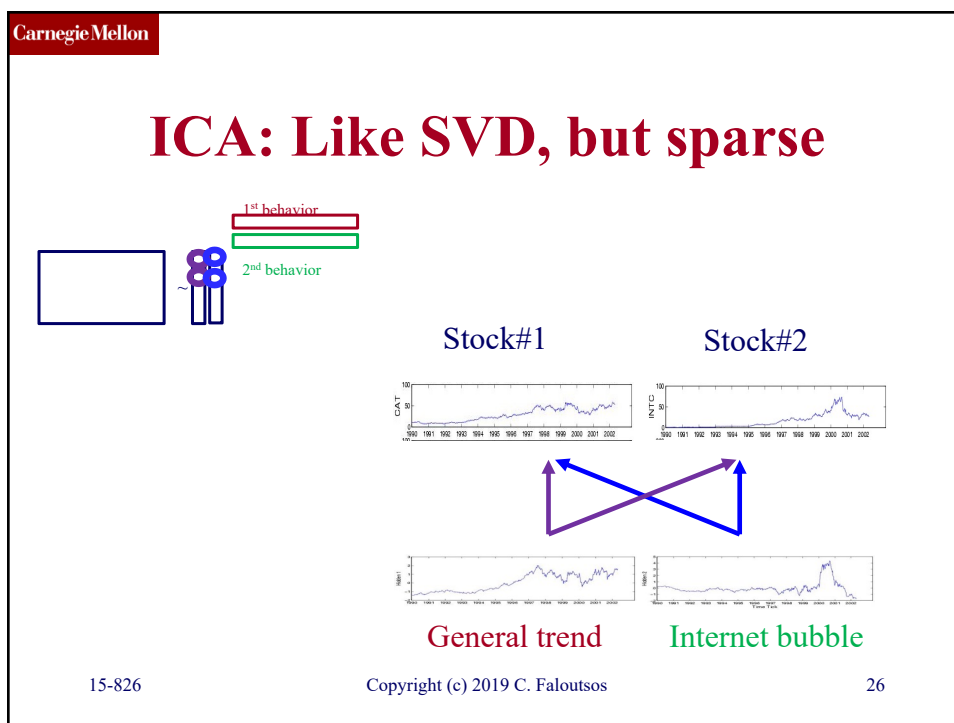
23



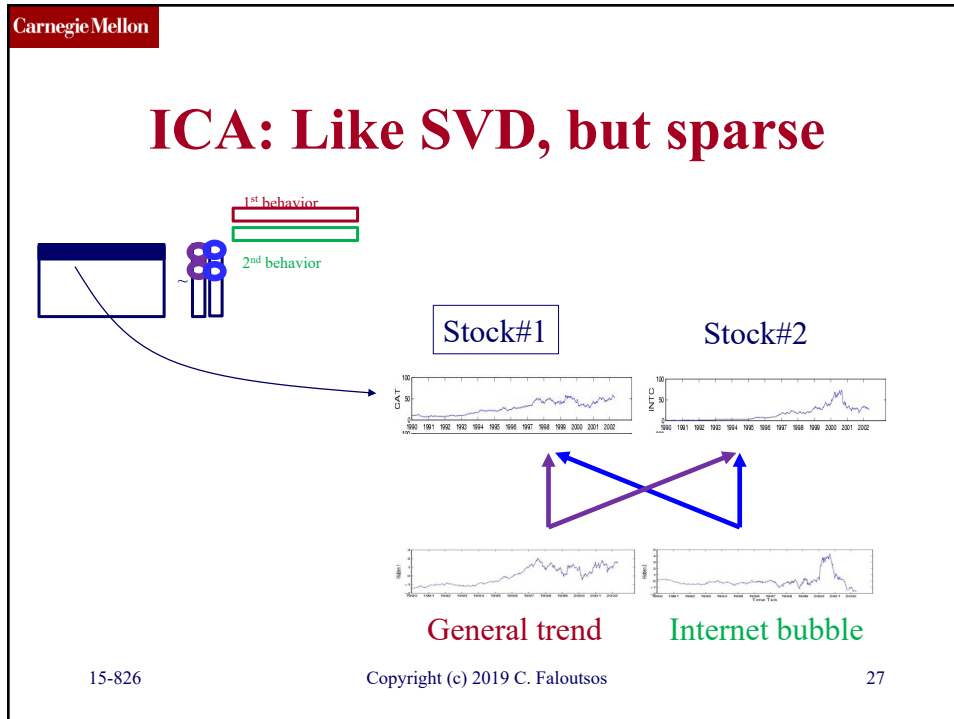
24



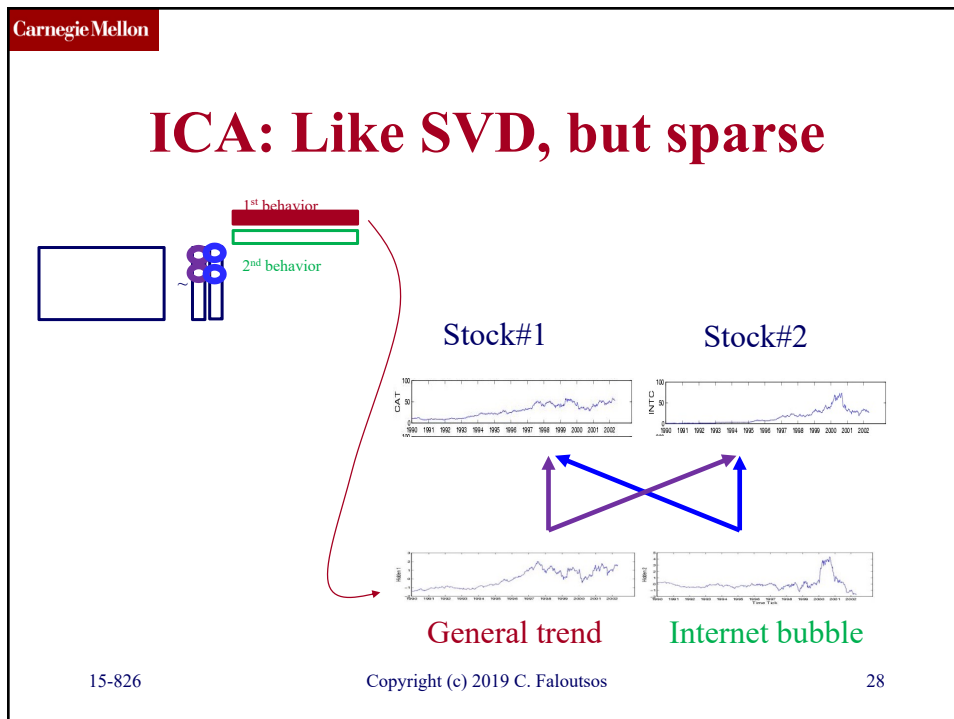
25



26



27



28

CarnegieMellon

ICA: Like SVD, but sparse

The diagram shows a matrix on the left being decomposed into two independent components: '1st behavior' (red) and '2nd behavior' (green). These components are used to explain the behavior of 'Stock#1' and 'Stock#2'. The '1st behavior' is associated with a 'General trend' (red), and the '2nd behavior' is associated with an 'Internet bubble' (green). Four line graphs show the time series for Stock#1 and Stock#2, with arrows indicating the contribution of each behavior to the overall stock price movement.

15-826 Copyright (c) 2019 C. Faloutsos 29

29

CarnegieMellon

What else can ICA tell us?

15-826 Copyright (c) 2019 C. Faloutsos 30

30

CarnegieMellon

Companies related to hidden variable 1

B _{1,j}			
Highest		Lowest	
Caterpillar	0.938512	AT&T	0.021885
Boeing	0.911120	WalMart	0.624570
MMM	0.906542	Intel	0.638010
Coca Cola	0.903858	Home Depot	0.647774
Du Pont	0.900317	Hewlett-Packard	0.658768

All companies are affected by the “general trend” variable (with weights 0.6~0.9), except AT&T.

15-826
Copyright (c) 2019 C. Faloutsos
#31

31

CarnegieMellon

General trend (and outlier)

AT&T

United Technologies

Walmart

Exxon Mobil

15-826
Copyright (c) 2019 C. Faloutsos
#32

32

CarnegieMellon

Companies related to hidden variable 2

B _{2,j}			
Highest		Lowest	
Intel	0.641102	Philip Morris	-0.194843
Hewlett-Packard	0.621159	International Paper	-0.089569
GE	0.509164	Caterpillar	0.031678
American Express	0.504871	Procter and Gamble	0.109576
Disney	0.490529	Du Pont	0.133337

Tech company

2000-2001 "Internet bubble"

15-826
Copyright (c) 2019 C. Faloutsos
#33

33

CarnegieMellon

Companies related to hidden variable 2

B _{2,j}			
Highest		Lowest	
Intel	0.641102	Philip Morris	-0.194843
Hewlett-Packard	0.621159	International Paper	-0.089569
GE	0.509164	Caterpillar	0.031678
American Express	0.504871	Procter and Gamble	0.109576
Disney	0.490529	Du Pont	0.133337

Tech company

Companies affected by the "internet bubble" variable (with weights 0.5~0.6) are tech-related. Other companies are un-related (weights < 0.15).

15-826
Copyright (c) 2019 C. Faloutsos
#34

34

CarnegieMellon

Outline

- Motivation
- Formulation
- PCA and ICA
- Example applications
 - Hidden variables in stock prices
 - ➔ – Find topics in documents
- Conclusion

15-826 Copyright (c) 2019 C. Faloutsos #35

35

CarnegieMellon

Topic discovery on text streams

- Data: CNN headline news (Jan.-Jun. 1998)
- Documents of 10 topics in one single text stream
 - Documents are sorted by date/time
 - Subsequent documents may have different topics

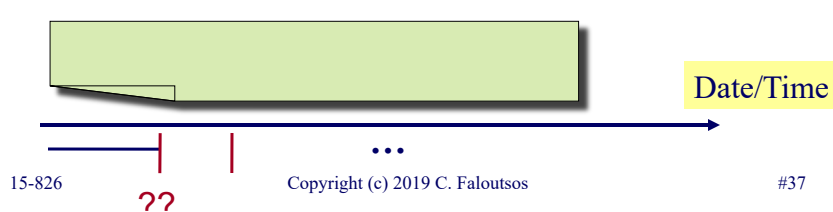
15-826 Copyright (c) 2019 C. Faloutsos #36

36

CarnegieMellon

Topic discovery on text streams

- Data: CNN headline news (Jan.-Jun. 1998)
- Documents of 10 topics in one single text stream
 - FIND: the document boundaries
 - AND: the terms of each topic



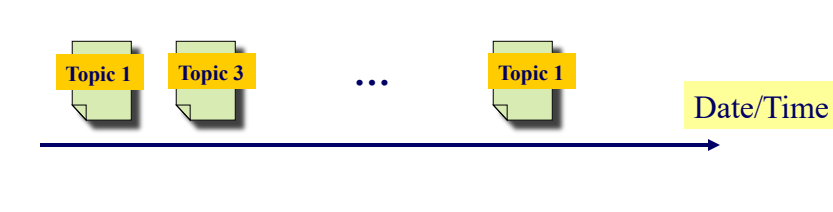
The diagram shows a horizontal timeline labeled 'Date/Time' with an arrow pointing right. A long green bar represents the text stream. Below the bar, a blue line with vertical tick marks indicates document boundaries. The first tick mark is labeled '15-826'. A red vertical line is labeled '??'. There are three dots between the red line and the next tick mark. The text 'Copyright (c) 2019 C. Faloutsos' is centered below the timeline, and '#37' is at the bottom right.

37

CarnegieMellon

Topic discovery on text streams

- Known: number of topics = 10
- Unknown: (1) topic of each document (2) topic description



The diagram shows a horizontal timeline labeled 'Date/Time' with an arrow pointing right. Three green document icons are shown above the timeline. The first icon is labeled 'Topic 1', the second 'Topic 3', and the third 'Topic 1'. There are three dots between the second and third icons. The text 'Copyright (c) 2019 C. Faloutsos' is centered below the timeline, and '#38' is at the bottom right.

38

CarnegieMellon

How to proceed?

The diagram shows a horizontal timeline with an arrow pointing to the right. A yellow box labeled "Date/Time" is positioned at the right end of the timeline. A green bar is positioned above the timeline, starting from the left and extending to the right. The bar has a small notch on its left side. Below the timeline, there are two vertical red lines, followed by an ellipsis "..." and then the number "39".

15-826 Copyright (c) 2019 C. Faloutsos 39

39

CarnegieMellon

How to proceed?

- A: Sliding windows

The diagram shows a horizontal timeline with an arrow pointing to the right. A yellow box labeled "Date/Time" is positioned at the right end of the timeline. A green bar is positioned above the timeline, starting from the left and extending to the right. The bar has a small notch on its left side. Above the green bar, there are several blue horizontal lines of varying lengths, representing sliding windows. Below the timeline, there are two vertical red lines, followed by an ellipsis "..." and then the number "40".

15-826 Copyright (c) 2019 C. Faloutsos 40

40

CarnegieMellon

Topic discovery in documents

Step 1

New stories

Windowing
(n=1659)
(30 words)

$$X_{[n \times m]} = \begin{bmatrix} -\bar{x}_1 \\ -\bar{x}_2 \\ \vdots \\ -\bar{x}_n \end{bmatrix}$$

$x_i = [1, 5, \dots, 0]$

m=3887 (dictionary size)

1st behavior

2nd behavior

15-826 Copyright (c) 2019 C. Faloutsos #41

41

CarnegieMellon

Step 3: Interpret the patterns

$\begin{bmatrix} -b'_1 \\ -b'_2 \\ \vdots \\ -b'_{10} \end{bmatrix}$

$b'_i = [0, 0.7, \dots, 0.6]$

m=3887 (dictionary size)

Top words: "animal", "zoo", ...

A hidden topic!

Topics found

ID	Sorted word list				
A	Mckinne	Sergeant	sexual	Major	Armi
B	bomb	Rudolph	Clinic	Atlanta	Birmingham
C	Winfrei	Beef	Texa	Oprah	Cattl
D	Viagra	Drug	Impot	Pill	Doctor
E	Zamora	Graham	Kill	Former	Jone
F	Medal	Olymp	Gold	Women	Game
G	Pope	Cube	Castro	Cuban	Visit
H	Asia	Economi	Japan	Econom	Asian
I	Super	Bowl	Game	Team	Re
J	Peopl	Tornado	Florida	Re	bomb

15-826

42

CarnegieMellon

Step 3: Evaluate the patterns

ID	True Topic
1	Sgt. Gene Mckinney is on trial for alleged sexual misconduct
2	A bomb explodes in a Birmingham, AL abortion clinic
3	The Cattle Industry in Texas sues Oprah Winfrey for defaming beef
4	New impotency drug Viagra is approved for use
5	Diane Zamora is convicted of helping to murder her lover's girlfriend

ID	Sorted word list				
A	mckinne	sergeant	sexual	major	armi
B	bomb	rudolph	clinic	atlanta	birmingham
C	winfrei	beef	texa	oprah	cattl
D	viagra	drug	Impot	pill	doctor
E	zamora	graham	kill	former	jone

AutoSplit finds correct topics.

15-826 #43

43

CarnegieMellon

Step 3: Evaluate the patterns

ID	AutoSplit				
A	mckinne	sergeant	sexual	major	armi
B	bomb	rudolph	clinic	atlanta	birmingham
C	winfrei	beef	texa	oprah	cattl
D	viagra	drug	Impot	pill	doctor
E	zamora	graham	kill	former	jone

ID	PCA				
A'	mckinne	bomb	women	sexual	sergeant
B'	bomb	mckinne	rudolph	clinic	atlanta
C'	winfrei	viagra	texa	beef	oprah
D'	viagra	winfrei	drug	texa	beef
E'	zamora	viagra	winfrei	graham	olymp

AutoSplit's topics are better than PCA.

15-826 #44

44

CarnegieMellon

Step 3: Evaluate the patterns

	AutoSplit				
A	pink	pink	pink	pink	pink
B	cyan	cyan	cyan	cyan	cyan
C	green	green	green	green	green
D	purple	purple	purple	purple	purple
E	yellow	yellow	yellow	yellow	yellow

	PCA				
A'	pink	cyan	white	pink	pink
B'	cyan	pink	cyan	cyan	cyan
C'	green	purple	green	green	green
D'	purple	green	purple	green	green
E'	yellow	purple	green	yellow	white

PCA vectors mix the topics.

15-826 **AutoSplit's topics are better than PCA.** #45

45

CarnegieMellon


Conclusion

- ICA: more flexible than PCA in finding patterns.
- Many applications
 - Find hidden variables in time series (e.g., stock prices)
 - Blind source separation
- Rule of thumb: plot after PCA;
 - if 'chicken-feet', try ICA

15-826 Copyright (c) 2019 C. Faloutsos #46

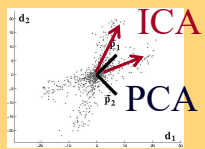
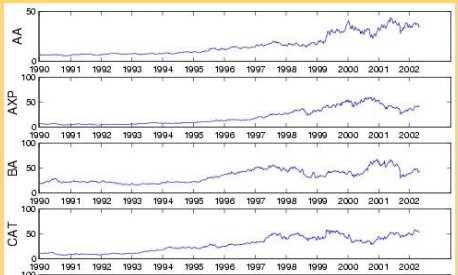
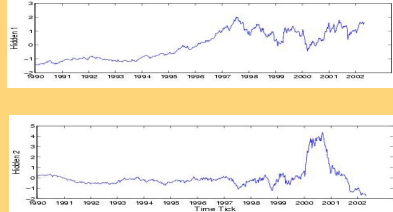
46

CarnegieMellon



Answer

Q: how to extract **sparse** hidden/latent variables?
A: ~~SVD~~ ICA

15-826 Copyright (c) 2019 C. Faloutsos 47


47

CarnegieMellon

Citation

- *AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases*, **Jia-Yu Pan**, Hiroyuki Kitagawa, Christos Faloutsos and Masafumi Hamamoto

PAKDD 2004, Sydney, Australia



15-826 Copyright (c) 2019 C. Faloutsos #48

48

CarnegieMellon

References

- Jia-Yu Pan, Andre Guilherme Ribeiro Balan, Eric P. Xing, Agma Juci Machado Traina, and Christos Faloutsos. *Automatic Mining of Fruit Fly Embryo Images. KDD, 2006.*
- Arnab Bhattacharya, Vebjorn Ljosa, Jia-Yu Pan, Mark R. Verardo, Hyungjeong Yang, Christos Faloutsos, and Ambuj K. Singh. *ViVo: Visual Vocabulary Construction for Mining Biomedical Images. ICDM, 2005.*
- Jia-Yu Pan, Hiroyuki Kitagawa, Christos Faloutsos, and Masafumi Hamamoto. *AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases. PAKDD, 2004.*


15-826 Copyright (c) 2019 C. Faloutsos #49

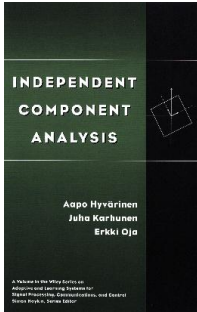
49

CarnegieMellon

References

- Aapo Hyvärinen, Juha Karhunen, Erkki Oja: [*Independent Component Analysis*](#), John Wiley & Sons, 2001





15-826 Copyright (c) 2019 C. Faloutsos #50

50

CarnegieMellon

Software

- Open source software: 'fastICA'
<http://research.ics.aalto.fi/ica/fastica/>

- Or 'autosplit' :

www.cs.cmu.edu/~jypan/software/autosplit_cmu.tar.gz

15-826

Copyright (c) 2019 C. Faloutsos

#51