Carnegie Mellon

15-826: Multimedia Databases and Data Mining

Lecture #20:

Independent Component Analysis (ICA)

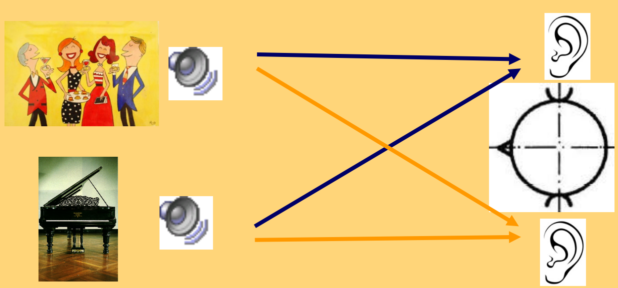
Christos Faloutsos

1

Carnegie Mellon

Problem: BSS

- two sound sources in a cocktail party – separate them



="blind source separation"
(= unknown sources, unknown mixing)

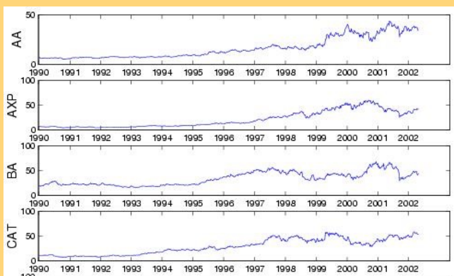
15-826 Copyright (c) 2019 C. Faloutsos #2

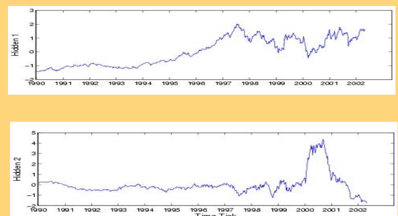
2

Carnegie Mellon

Problem

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





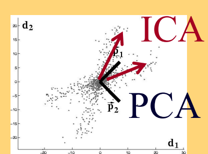
15-826 Copyright (c) 2019 C. Faloutsos 3

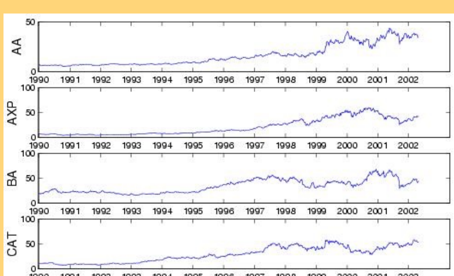
3

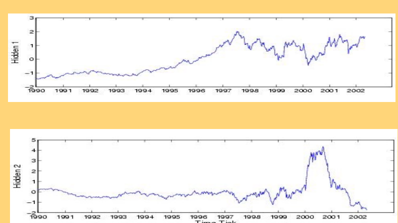
Carnegie Mellon

Answer

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







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

4

Carnegie Mellon

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

5

Carnegie Mellon

Outline

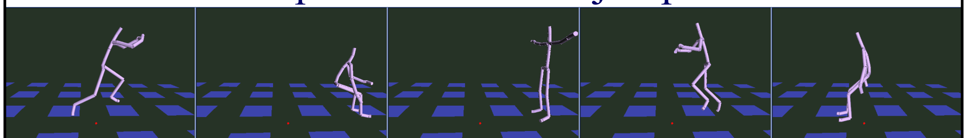
- Motivation
- Formulation
- PCA and ICA
- Example applications
- Conclusion

6

Carnegie Mellon

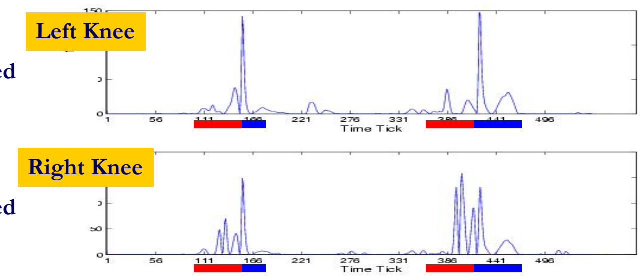
Motivation: (Q1) Find patterns in data

- Motion capture data: broad jumps



Energy exerted

Energy exerted

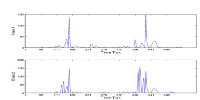


Take-off
Landing

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

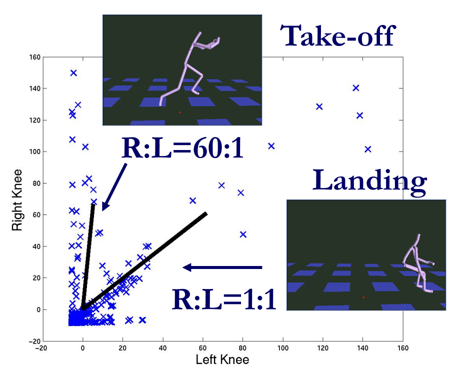
7

Carnegie Mellon



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

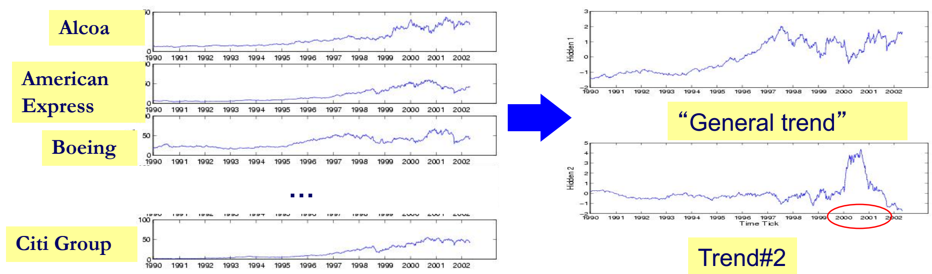
8

Carnegie Mellon

Motivation: (Q2) Find hidden variables

Stock prices

Hidden variables (= 'topics' = concepts)

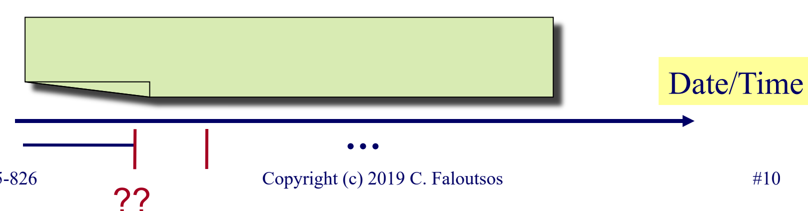


9

Carnegie Mellon

(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



10

Carnegie Mellon

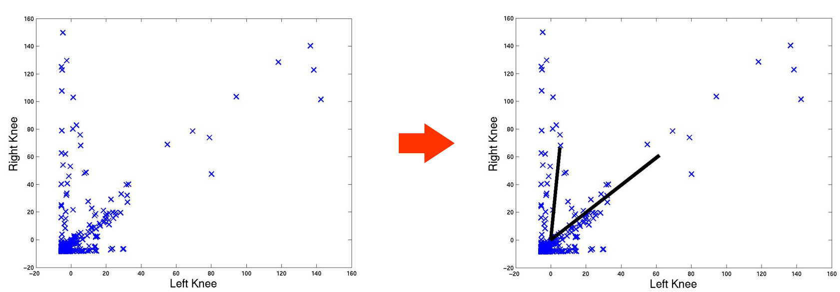
Outline

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

11

Carnegie Mellon

Formulation: Finding patterns



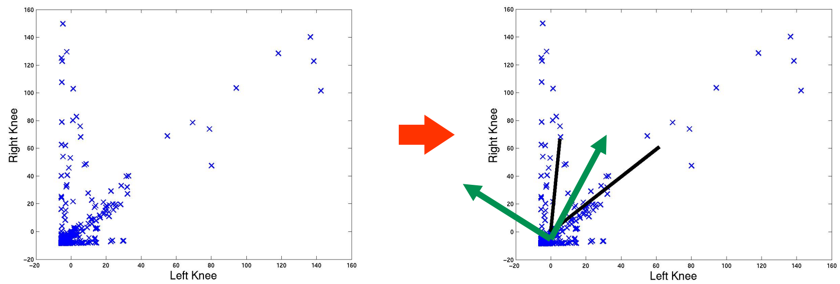
Given n data points, each with m attributes.

Find patterns that describe data properties the best.

12

Carnegie Mellon

Formulation: Finding patterns



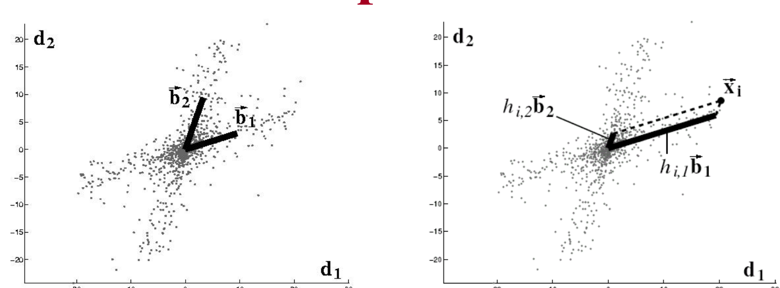
Given n data points,
each with m attributes.

SVD/PCA: ORTHOGONAL vectors

13

Carnegie Mellon

Linear representation



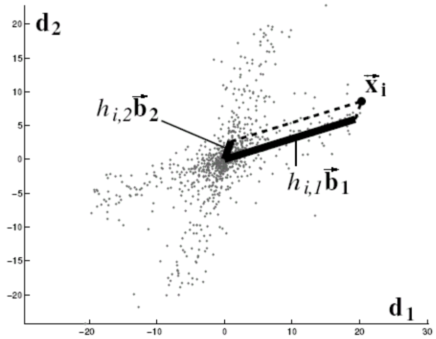
- Find vectors that describe the data set the best.
- Each point: linear combination of the vectors (patterns):

$$\vec{\mathbf{x}}_{\mathbf{i}} = h_{i,1}\vec{\mathbf{b}}_{\mathbf{1}} + h_{i,2}\vec{\mathbf{b}}_2$$

14

Carnegie Mellon

Patterns as data "vocabulary"



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

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

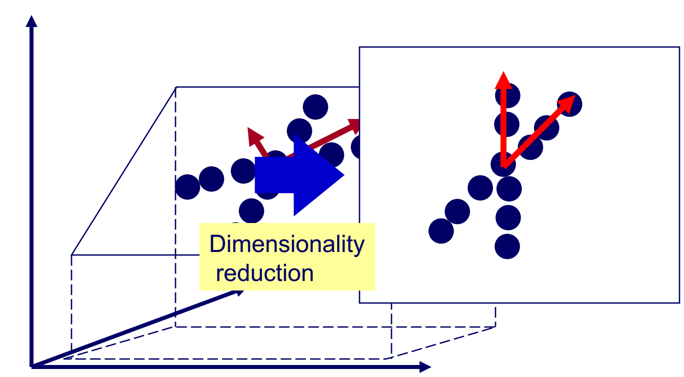
Good pattern
≈ sparse coding

b_1 alone, can describe x_i.

15

Carnegie Mellon

PCA: first step of ICA



PCA finds the hyperplane.

ICA finds the correct patterns.

16

Carnegie Mellon

Software

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

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

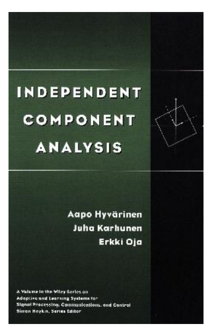
15-826 Copyright (c) 2019 C. Faloutsos #17

17

Carnegie Mellon

References

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



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

18

Carnegie Mellon

Outline

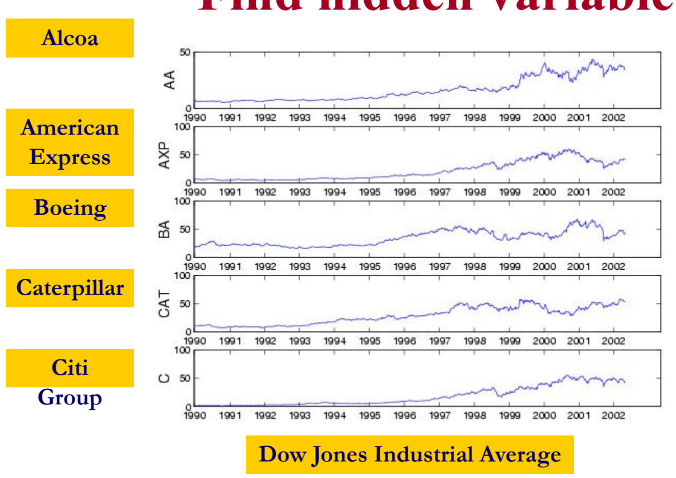
- Motivation
- Formulation
- PCA and ICA
- Example applications
  - Hidden variables in stock prices
  - Find topics in documents
- Conclusion

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

19

Carnegie Mellon

Motivation: Find hidden variables



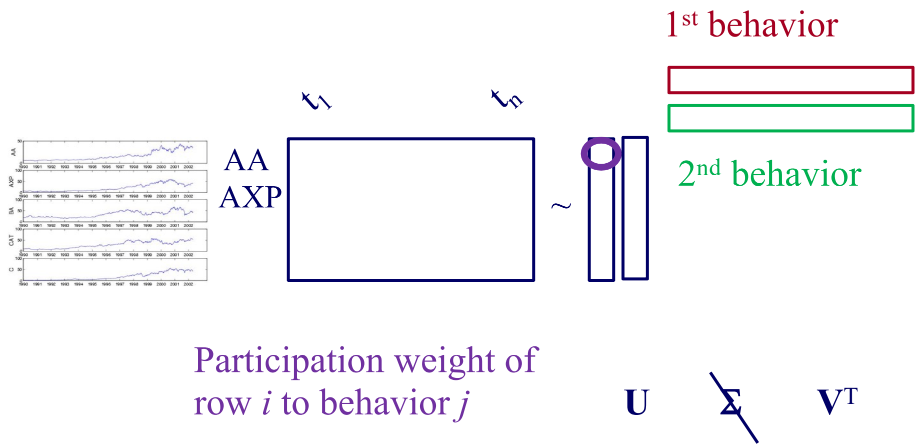
Find common hidden variables, and weights.

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

20

Carnegie Mellon

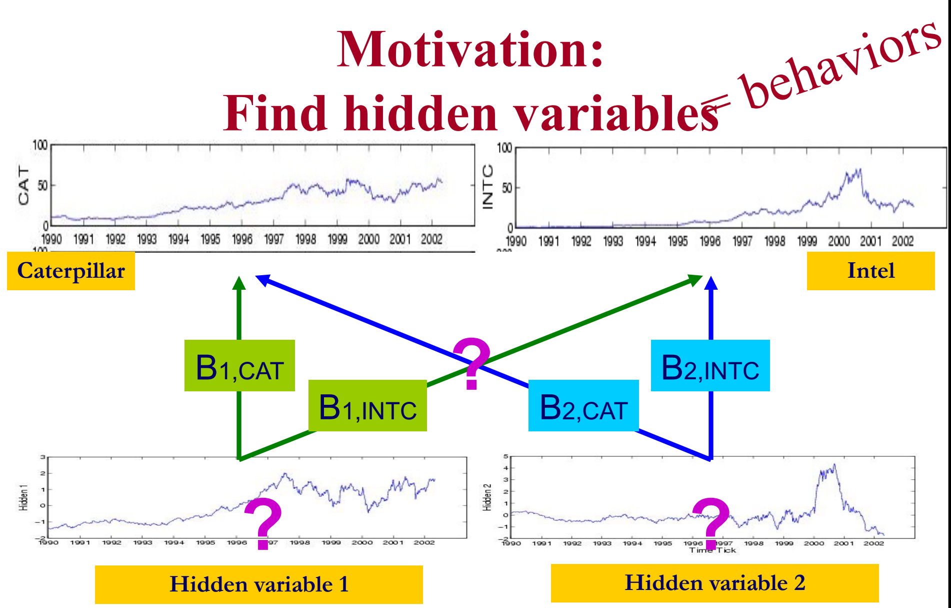
ICA: Like SVD, but sparse U



21

Carnegie Mellon

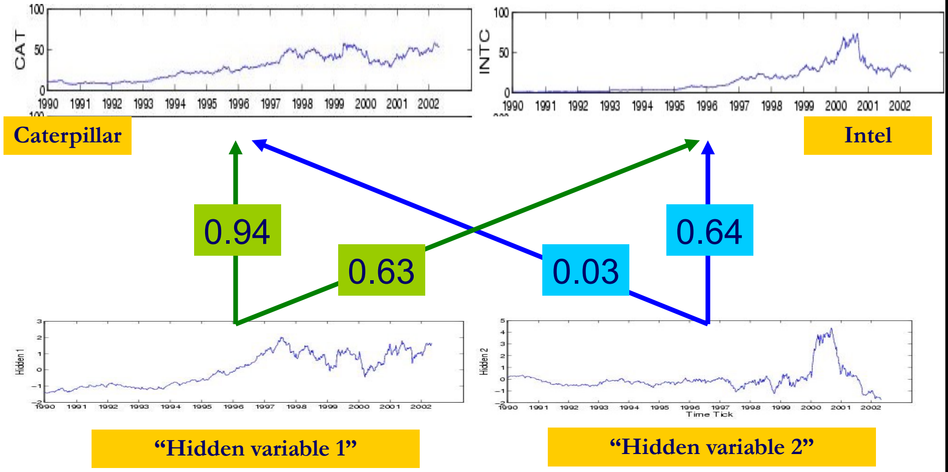
Motivation: Find hidden variables



22

Carnegie Mellon

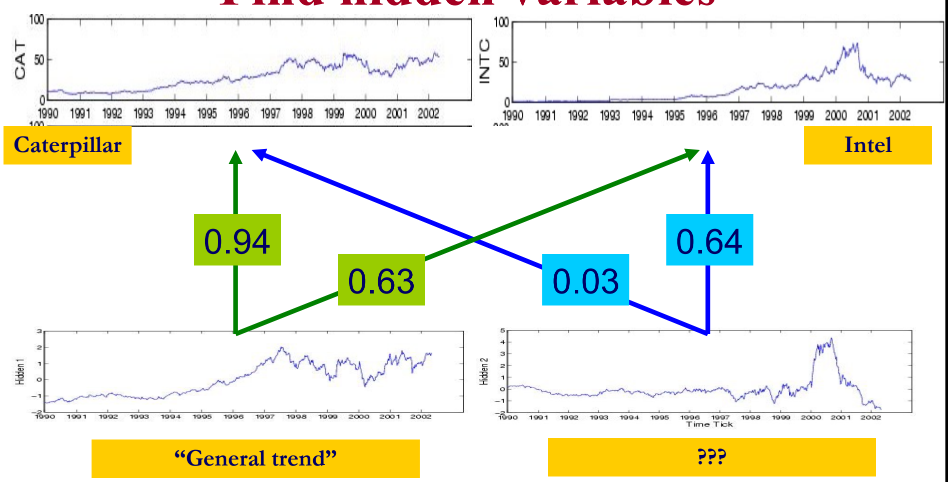
Motivation: Find hidden variables



23

Carnegie Mellon

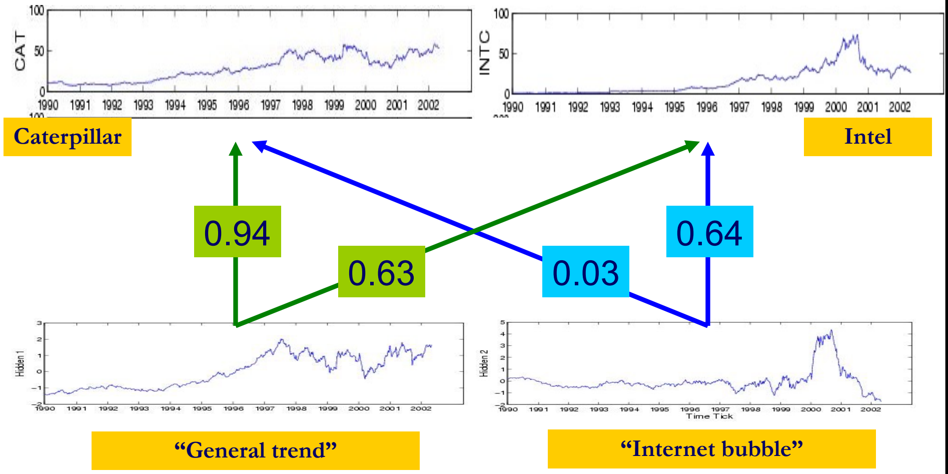
Motivation: Find hidden variables



24

Carnegie Mellon

Motivation: Find hidden variables

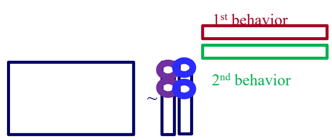


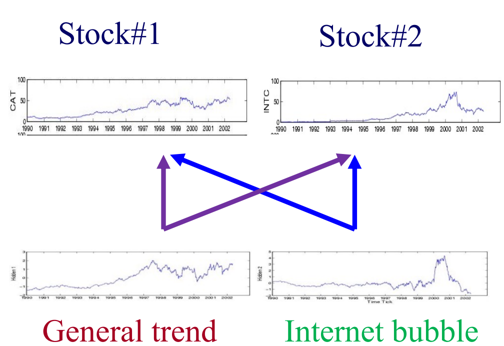
15-826 Copyright (c) 2019 C. Faloutsos 25

25

Carnegie Mellon

ICA: Like SVD, but sparse



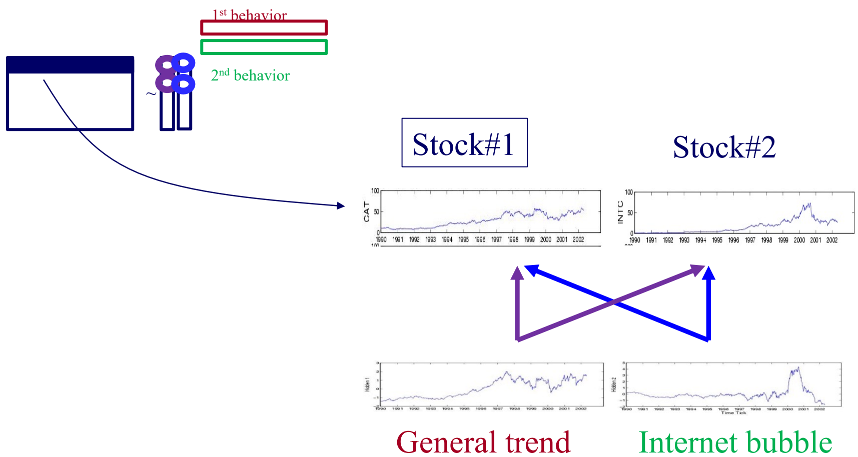


15-826 Copyright (c) 2019 C. Faloutsos 26

26

Carnegie Mellon

ICA: Like SVD, but sparse

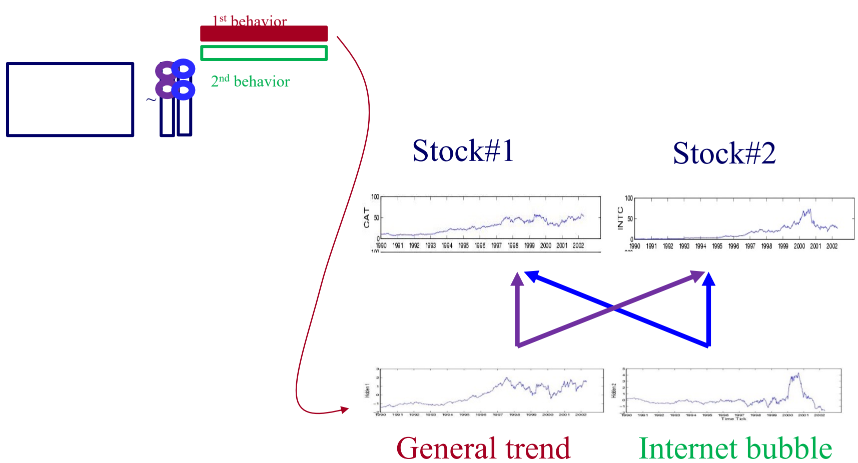


15-826 Copyright (c) 2019 C. Faloutsos 27

27

Carnegie Mellon

ICA: Like SVD, but sparse

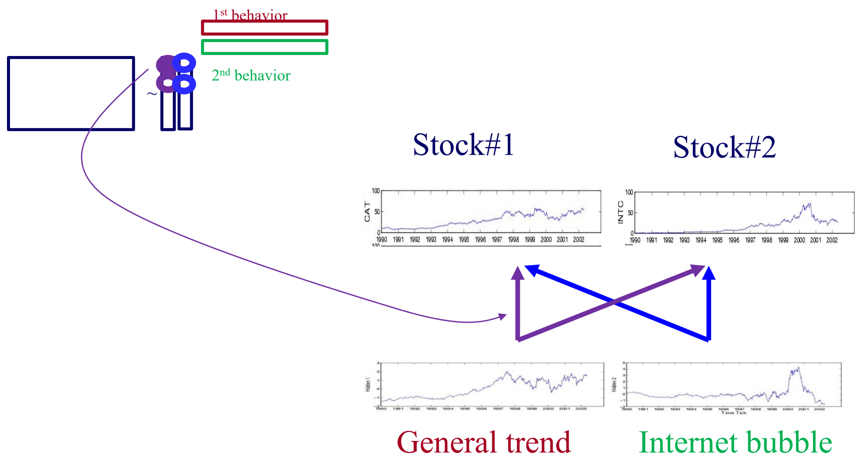


15-826 Copyright (c) 2019 C. Faloutsos 28

28

Carnegie Mellon

ICA: Like SVD, but sparse



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

29

Carnegie Mellon

What else can ICA tell us?

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

30

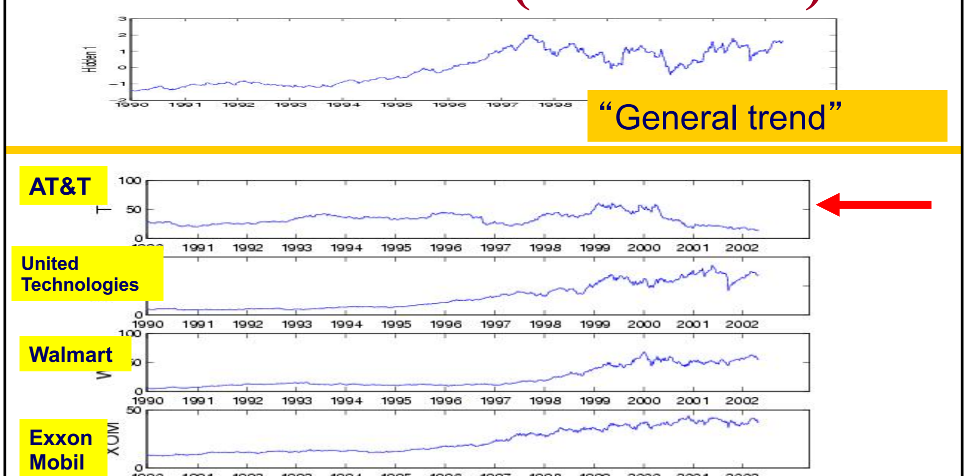
Companies related to hidden variable 1

B1,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.

31

General trend (and outlier)

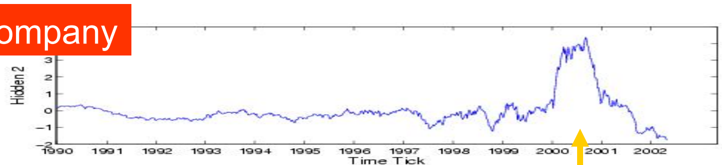


32

Companies related to hidden variable 2

B2,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"

33

Companies related to hidden variable 2

B2,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).

34

Carnegie Mellon

Outline

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

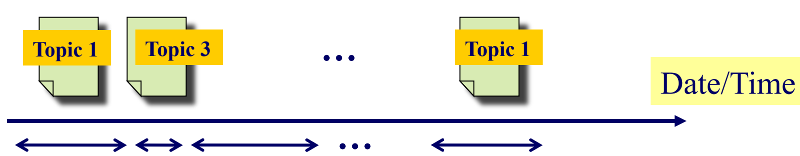
15-826 #35

35

Carnegie Mellon

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



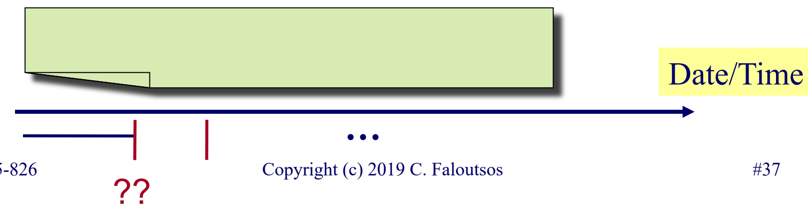
15-826 #36

36

Carnegie Mellon

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

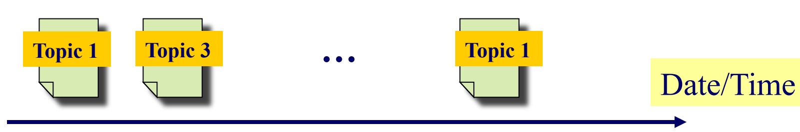


37

Carnegie Mellon

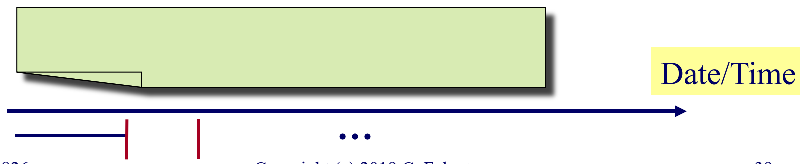
Topic discovery on text streams

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



38

How to proceed?

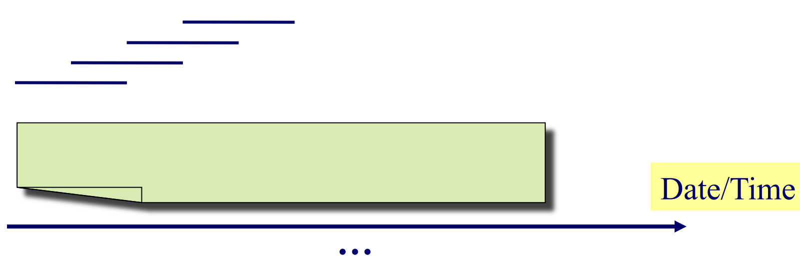


15-826 Copyright (c) 2019 C. Faloutsos 39

39

How to proceed?

- A: Sliding windows



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

40

Carnegie Mellon

Topic discovery in documents

Step 1

New stories

Windowing

(n=1659)

(30 words)

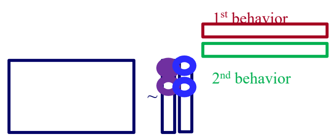
$$\begin{bmatrix} -\vec{x}_1- \\ -\vec{x}_2- \\ \vdots \\ -\vec{x}_n- \end{bmatrix}$$

$\mathbf{X}_{[n \times m]}$

aaron zoo

x_i = [1, 5, …, 0]

m=3887 (dictionary size)



41

Carnegie Mellon

Step 3: Interpret the patterns

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

aaron animal zoo

b'_i = [0, 0.7, …, 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

42

Carnegie Mellon

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.

Carnegie Mellon

Step 3: Evaluate the patterns

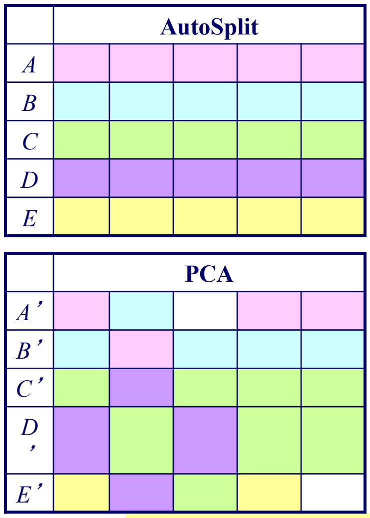
ID	AutoSplit				
A	mckinne	sergeant	sexual	major	armi
B	bomb	rudolph	clinic	atlanta	birmingham
C	winfrei	beef	texa	oprah	cattl
D	viagra	drug	Impot	pill	doctor
E	zamora	graham	kill	former	jone

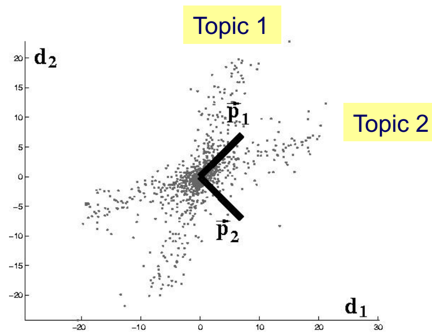
ID	PCA				
A’	mckinne	bomb	women	sexual	sergeant
B’	bomb	mckinne	rudolph	clinic	atlanta
C’	winfrei	viagra	texa	beef	oprah
D’	viagra	winfrei	drug	texa	beef
E’	zamora	viagra	winfrei	graham	olymp

AutoSplit’s topics are better than PCA.

Carnegie Mellon

Step 3: Evaluate the patterns



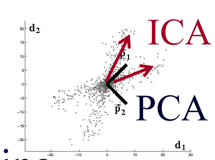


PCA vectors mix the topics.

AutoSplit's topics are better than PCA.

Carnegie Mellon

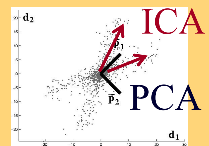
Conclusion

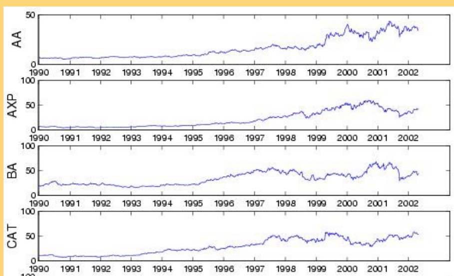


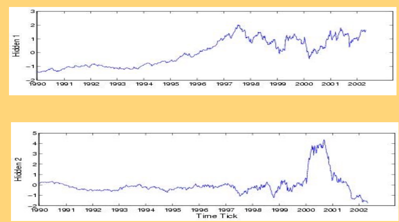
- ICA: more flexible than PCA in finding patterns.
- Many applications
  - Find hidden variables in time series (e.g., stock prices)
  - Blind source separation
- Rule of thumb: plot after PCA;
  - if 'chicken-feet', try ICA

Answer

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







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

47

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

Carnegie Mellon

References

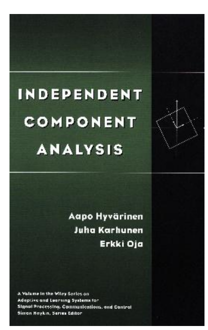
- Jia-Yu Pan, Andre Guilherme Ribeiro Balan, Eric P. Xing, Agma Juci Machado Traina, and Christos Faloutsos. *Automatic Mining of Fruit Fly Embryo Images*. *KDD,* 2006.
- Arnab Bhattacharya, Vebjorn Ljosa, Jia-Yu Pan, Mark R. Verardo, Hyungjeong Yang, Christos Faloutsos, and Ambuj K. Singh. *ViVo: Visual 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.

49

Carnegie Mellon

References

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



50

Carnegie Mellon

Software

- Open source software: ‘fastICA’
  http://research.ics.aalto.fi/ica/fastica/

- Or ‘autosplit’ :

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

