

CMU SCS

15-826: Multimedia Databases and Data Mining

Addendum to Lecture #21: *Independent Component Analysis (ICA)*

Jia-Yu Pan and Christos Faloutsos


15-826
(c) C. Faloutsos and J-Y Pan (2011)
#1


CMU SCS

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

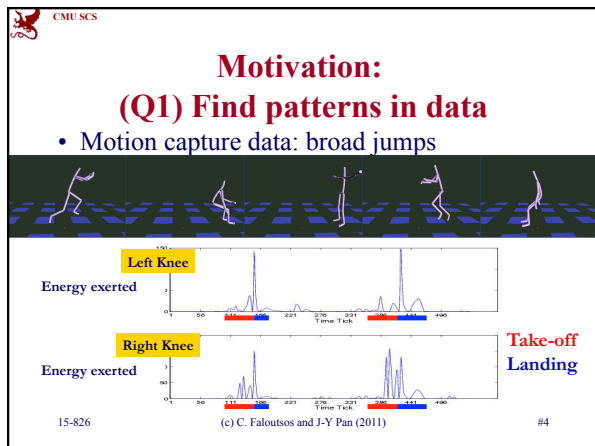
15-826
(c) C. Faloutsos and J-Y Pan (2011)
#2

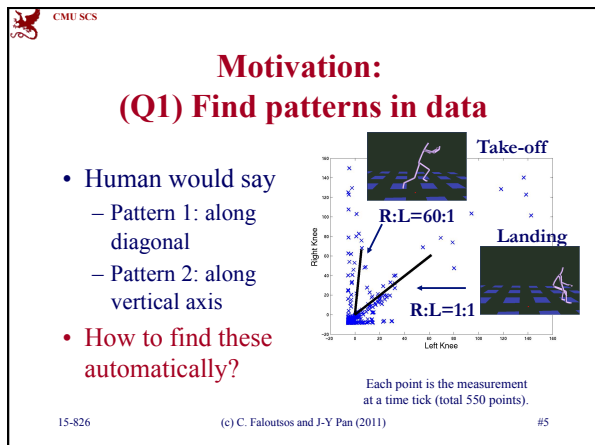

CMU SCS

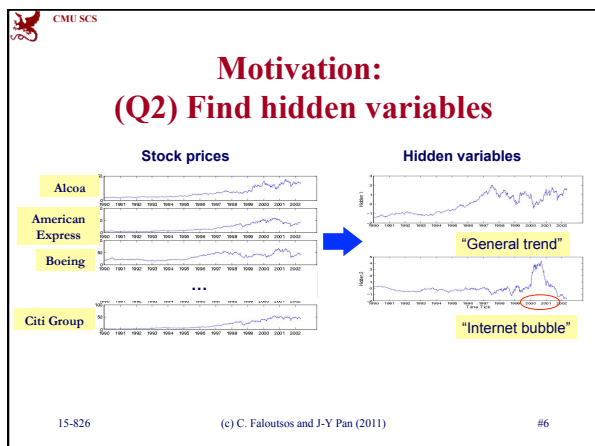
Outline

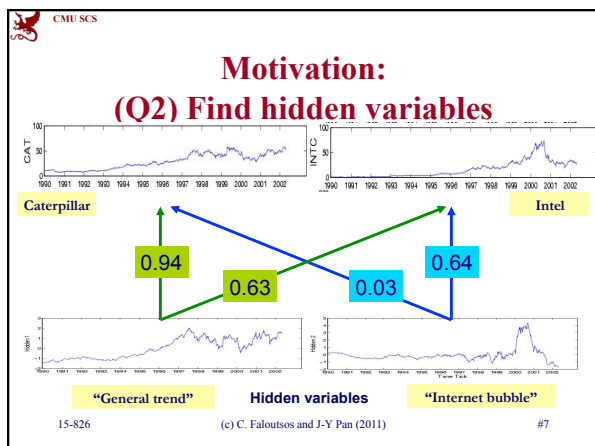
- Motivation
- Formulation
- PCA and ICA
- Example applications
- Conclusion

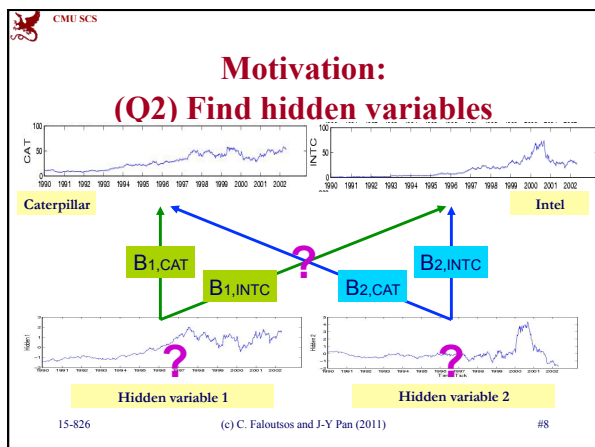
15-826
(c) C. Faloutsos and J-Y Pan (2011)
#3

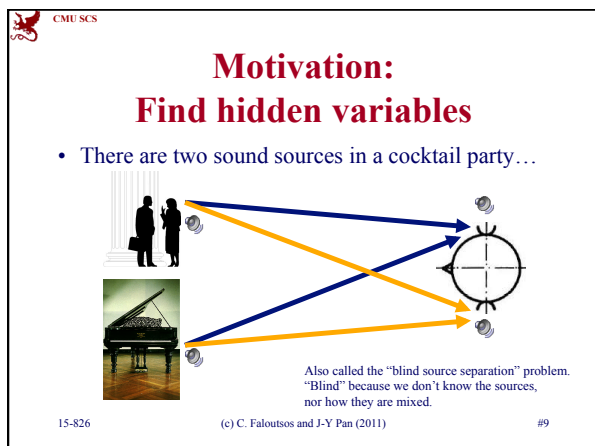















CMU SCS

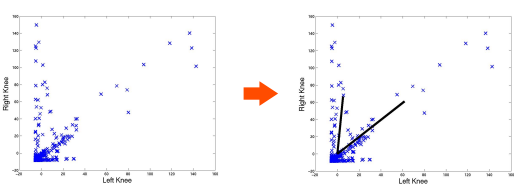
Outline

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

15-826
(c) C. Faloutsos and J-Y Pan (2011)
#10


CMU SCS


Formulation: Finding patterns



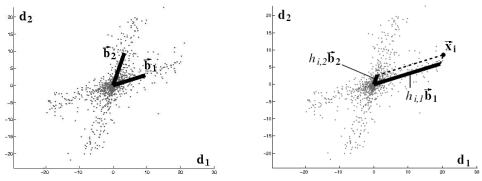
Given n data points, each with m attributes.

Find patterns that describe data properties the best.

15-826
(c) C. Faloutsos and J-Y Pan (2011)
#11


CMU SCS

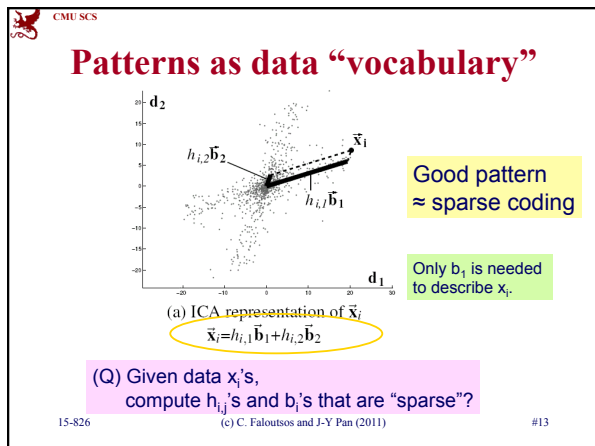
Linear representation

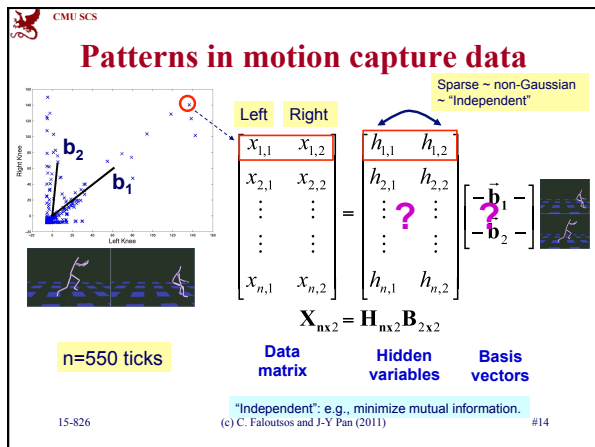


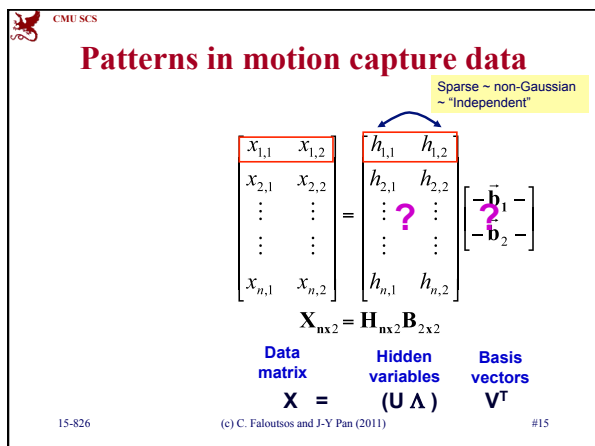
- Find vectors that describe the data set the best.
- Each point: linear combination of the vectors (patterns):


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

15-826
(c) C. Faloutsos and J-Y Pan (2011)
#12








CMU SCS


Outline

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

15-826

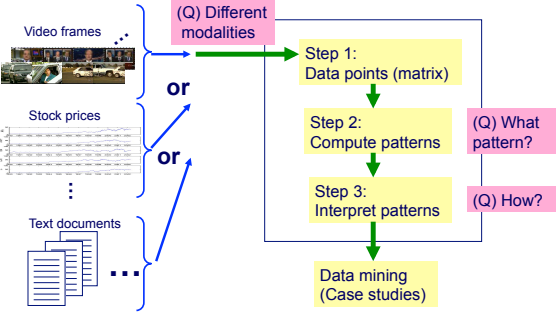
(c) C. Faloutsos and J-Y Pan (2011)

#16


CMU SCS

Pattern discovery with ICA: AutoSplit


[PAKDD 04][WIRI 05]



15-826

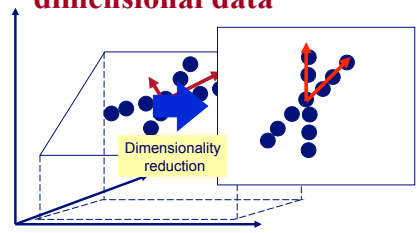
(c) C. Faloutsos and J-Y Pan (2011)

#17


CMU SCS

Finding patterns in high-dimensional data

details




PCA finds the hyperplane. ICA finds the correct patterns.

15-826

(c) C. Faloutsos and J-Y Pan (2011)


#18


CMU SCS

Outline

- Motivation
- Formulation
- PCA and ICA
- Example applications
 - ➡ – Find topics in documents
 - Hidden variables in stock prices
 - Visual vocabulary for retinal images
- Conclusion


15-826
(c) C. Faloutsos and J-Y Pan (2011)
#19


CMU SCS

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
(c) C. Faloutsos and J-Y Pan (2011)
#20


CMU SCS

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
(c) C. Faloutsos and J-Y Pan (2011)
#21

Topic discovery on text streams

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

15-826
(c) C. Faloutsos and J-Y Pan (2011)
#22

Topic discovery in documents

Step 1

New stories → Winding (n=1659, 30 words) → $X_{[n \times m]}$

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

m=3887 (dictionary size)

Step 2

$X_{[n \times m]} = H_{[n \times m]}$

$B_{[m' \times m']}$

(1) Find hyperplane (m'=10)

(2) Find patterns

Step 3

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

(Q) What does b'_i mean?

aaron animal zoo

15-826
(c) C. Faloutsos and J-Y Pan (2011)
#23

Step 3: Interpret the patterns

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

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
H	Asia	Economi	Japan	Econom	Asian
I	Super	Bowl	Game	Team	Re
J	Peopl	Tornado	Florida	Re	bomb

General idea: related to the data attributes

15-826


CMU SCS


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
#25


CMU SCS


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.

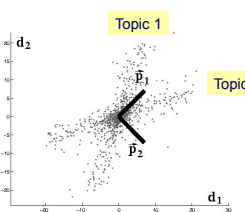
15-826
(c) C. Faloutsos and J-Y Pan (2011)
#26


CMU SCS

Step 3: Evaluate the patterns

	AutoSplit					
A						
B						
C						
D						
E						


	PCA					
A'						
B'						
C'						
D'						
E'						



PCA vectors mix the topics.

AutoSplit's topics are better than PCA.


15-826
(c) C. Faloutsos and J-Y Pan (2011)
#27


CMU SCS

Outline

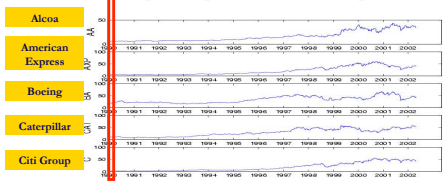
- Motivation
- Formulation
- PCA and ICA
- Example applications
 - Find topics in documents
 - Hidden variables in stock prices
- Conclusion

15-826
(c) C. Faloutsos and J-Y Pan (2011)
#28



CMU SCS

Find hidden variables (DJIA stocks)

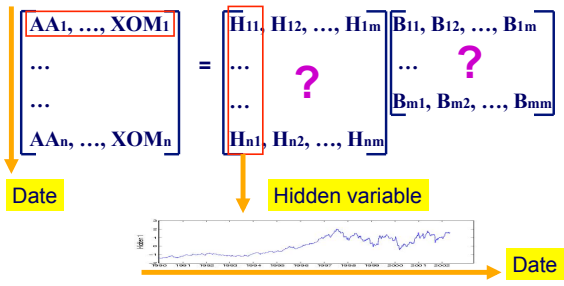
- Weekly DJIA closing prices
 - 01/02/1990-08/05/2002, n=660 data points
 - A data point: prices of 29 companies at the time



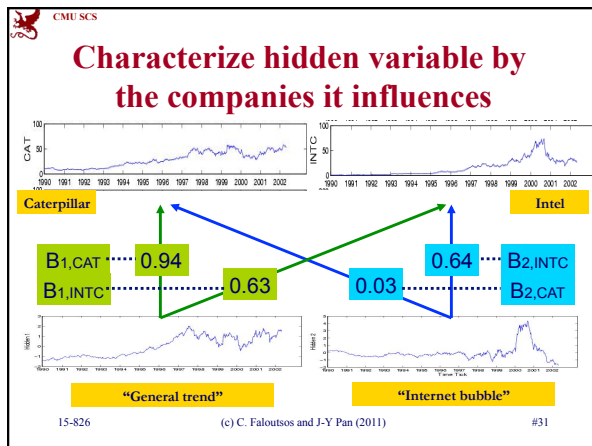
15-826
(c) C. Faloutsos and J-Y Pan (2011)
#29

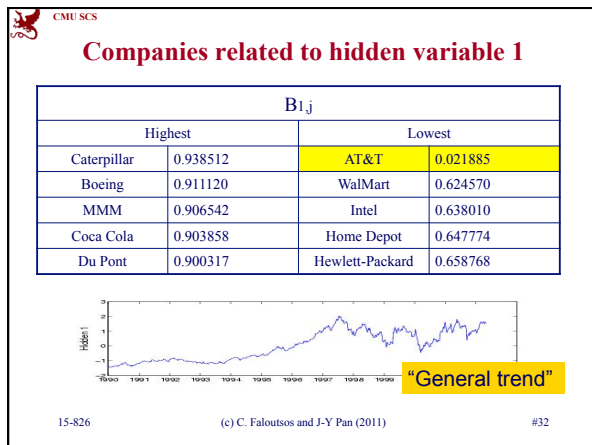

CMU SCS

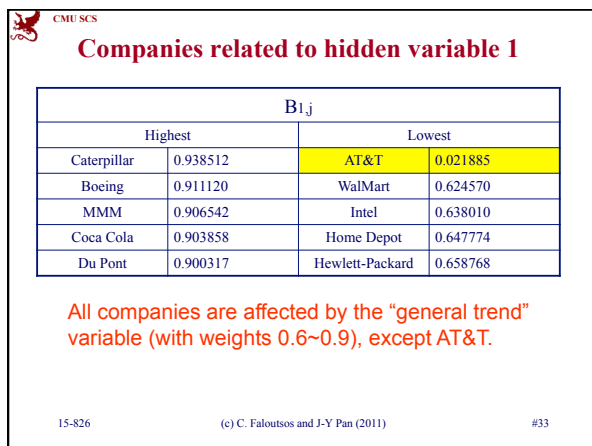
Formulation: Find hidden variables

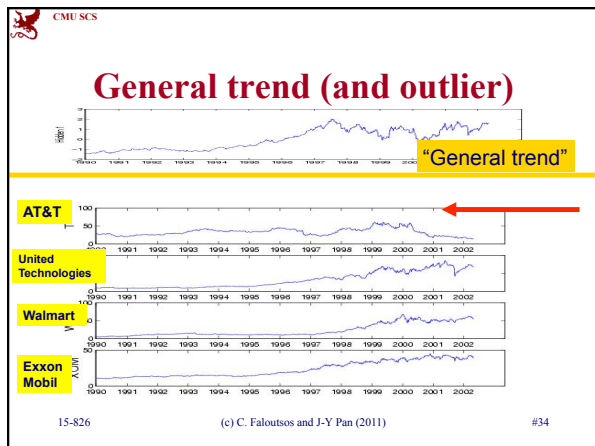


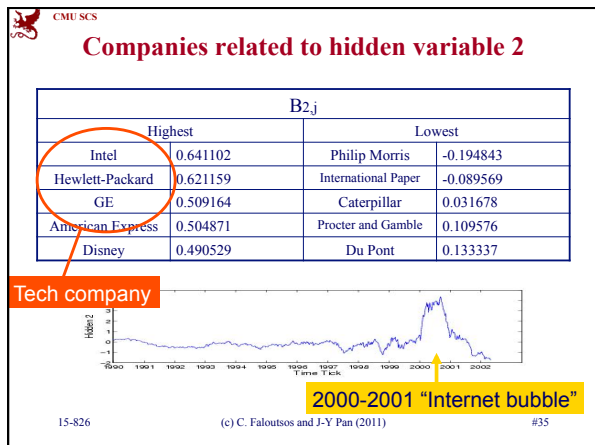
15-826
(c) C. Faloutsos and J-Y Pan (2011)
#30

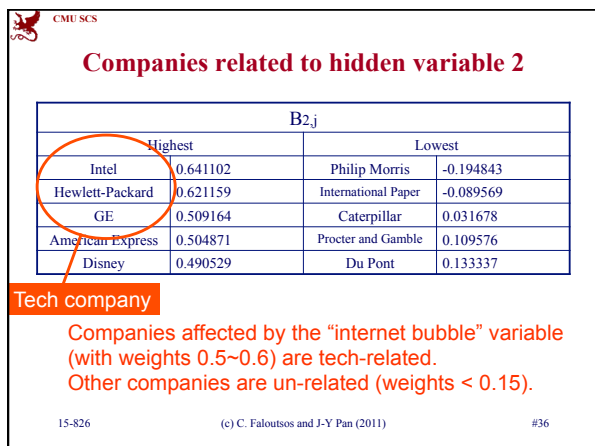















CMU SCS

Outline

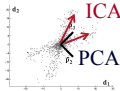
- Motivation
- Formulation
- PCA and ICA
- Example applications
 - Find topics in documents
 - Hidden variables in stock prices
 - Visual vocabulary for retinal images
- ➔ Conclusion

15-826
(c) C. Faloutsos and J-Y Pan (2011)
#37



CMU SCS

Conclusion

- ICA: more flexible than PCA in finding patterns.
- Many applications
 - Find topics and “vocabulary” for images
 - Find hidden variables in time series (e.g., stock prices)
 - Blind source separation




15-826
(c) C. Faloutsos and J-Y Pan (2011)
#38


CMU SCS


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
(c) C. Faloutsos and J-Y Pan (2011)
#39


CMU SCS

References

- Jia-Yu Pan, Andre Guilherme Ribeiro Balan, Eric P. Xing, Agma Juci Machado Traima, and Christos Faloutsos. Automatic Mining of Fruit Fly Embryo Images. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.
- Arnab Bhattacharya, Vebjørn Ljosa, Jia-Yu Pan, Mark R. Verardo, Hyungjeong Yang, Christos Faloutsos, and Ambuj K. Singh. ViVo: Visual Vocabulary Construction for Mining Biomedical Images. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM)*, 2005.
- Masafumi Hamamoto, Hiroyuki Kitagawa, Jia-Yu Pan, and Christos Faloutsos. A Comparative Study of Feature Vector-Based Topic Detection Schemes for Text Streams. In *Proceedings of International Workshop on Challenges in Web Information Retrieval and Integration (WIRI)*, 2005, pp.125-130.
- Jia-Yu Pan, Hiroyuki Kitagawa, Christos Faloutsos, and Masafumi Hamamoto. AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases. In *Proceedings of the The Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2004.


15-826
(c) C. Faloutsos and J-Y Pan (2011)
#40


CMU SCS

References

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

15-826
(c) C. Faloutsos and J-Y Pan (2011)
#41


CMU SCS

Software

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

15-826
(c) C. Faloutsos and J-Y Pan (2011)
#42
