15-826: Multimedia Databases and Data Mining

Lecture #21:

Independent Component Analysis (ICA)

Jia-Yu Pan and Christos Faloutsos

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

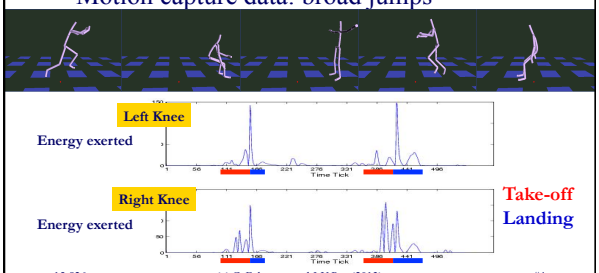
Outline

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

CMU SCS

Motivation: (Q1) Find patterns in data

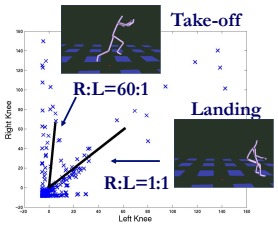
- Motion capture data: broad jumps



CMU SCS

Motivation: (Q1) Find patterns in data

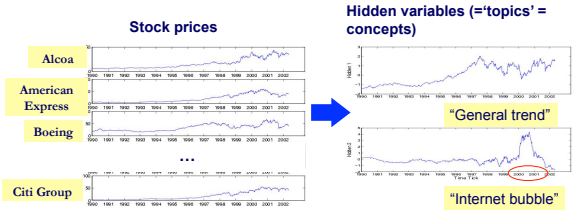
- Human would say
  - Pattern 1: along diagonal
  - Pattern 2: along vertical axis
- How to find these automatically?



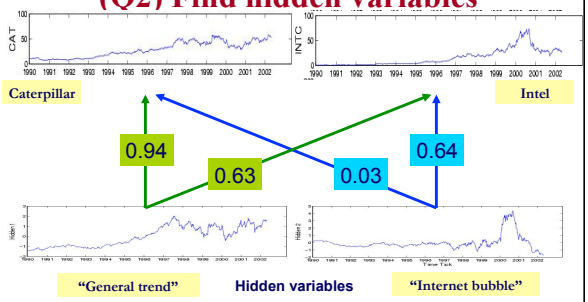
Each point is the measurement at a time tick (total 550 points).

CMU SCS

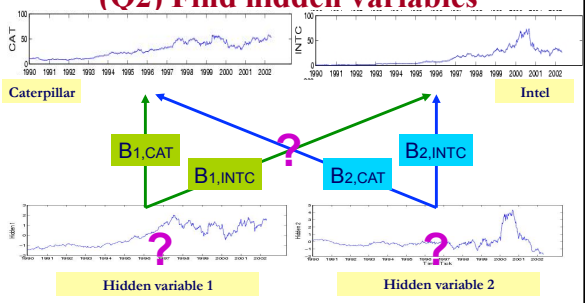
Motivation: (Q2) Find hidden variables



Motivation: (Q2) Find hidden variables

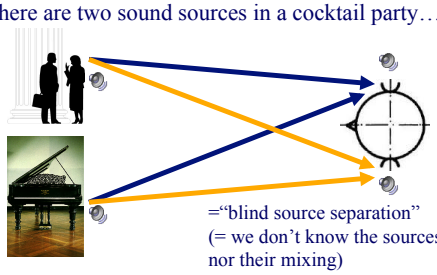


Motivation: (Q2) Find hidden variables



Motivation: Find hidden variables

- There are two sound sources in a cocktail party…

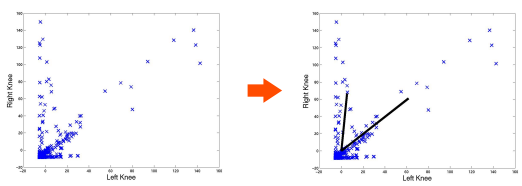


="blind source separation"
(= we don't know the sources, nor their mixing)

Outline

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

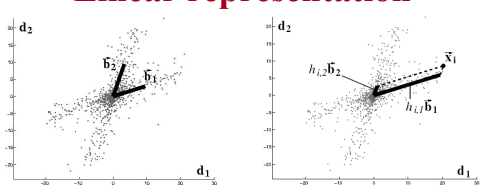
Formulation: Finding patterns



Given n data points, each with m attributes.

Find patterns that describe data properties the best.

Linear representation

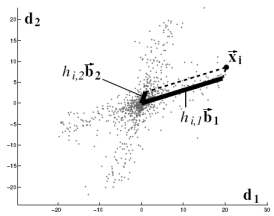


- Find vectors that describe the data set the best.
- Each point: linear combination of the vectors (patterns):

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

CMU SCS

Patterns as data "vocabulary"



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

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

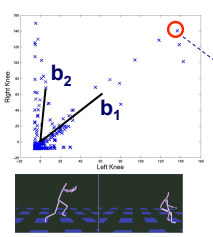
Good pattern ≈ sparse coding

Only b_1 is needed to describe x_i.

(Q) Given data x_i's, compute $h_{i,j}$'s and b_i's that are "sparse"?

CMU SCS

Patterns in motion capture data



Sparse ~ non-Gaussian ~ "Independent"

Left Right

$$\begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \vdots \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{bmatrix} = \begin{bmatrix} h_{1,1} & h_{1,2} \\ h_{2,1} & h_{2,2} \\ \vdots & \vdots \\ \vdots & \vdots \\ h_{n,1} & h_{n,2} \end{bmatrix} \begin{bmatrix} - \vec{\mathbf{b}}_1 - \\ - \vec{\mathbf{b}}_2 - \end{bmatrix}$$

?

$$\mathbf{X}_{\mathbf{nx2}} = \mathbf{H}_{\mathbf{nx2}}\mathbf{B}_{\mathbf{2x2}}$$

n=550 ticks

Data matrix — Hidden variables — Basis vectors

"Independent": e.g., minimize mutual information.

CMU SCS

Patterns in motion capture data

Sparse ~ non-Gaussian ~ "Independent"

$$\begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \vdots \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{bmatrix} = \begin{bmatrix} h_{1,1} & h_{1,2} \\ h_{2,1} & h_{2,2} \\ \vdots & \vdots \\ \vdots & \vdots \\ h_{n,1} & h_{n,2} \end{bmatrix} \begin{bmatrix} - \vec{\mathbf{b}}_1 - \\ - \vec{\mathbf{b}}_2 - \end{bmatrix}$$

?

$$\mathbf{X}_{\mathbf{nx2}} = \mathbf{H}_{\mathbf{nx2}}\mathbf{B}_{\mathbf{2x2}}$$

Data matrix — Hidden variables — Basis vectors

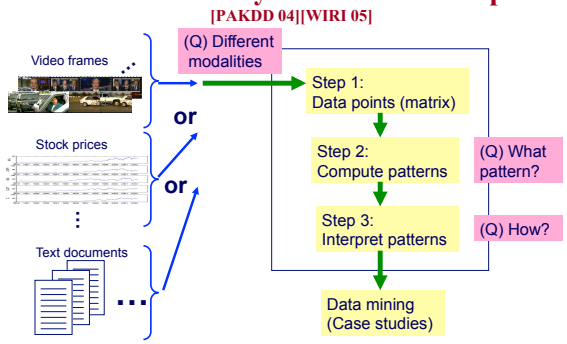
$\mathbf{X} = (\mathbf{U}\,\Lambda)\ \mathbf{V}^T$

Outline

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

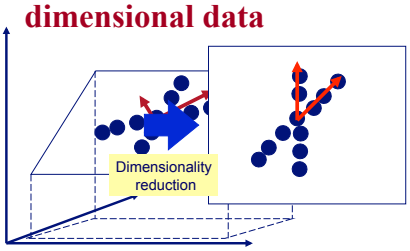
Pattern discovery with ICA: AutoSplit

[PAKDD 04][WIRI 05]



details

Finding patterns in high-dimensional data



PCA finds the hyperplane. ICA finds the correct patterns.

Outline

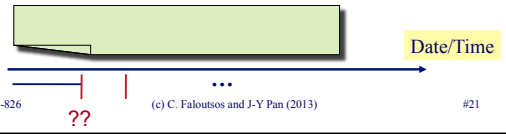
- Motivation
- Formulation
- PCA and ICA
- Example applications
  - Find topics in documents
  - Hidden variables in stock prices
  - Visual vocabulary for retinal images
- Conclusion

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

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



CMU SCS

Topic discovery on text streams

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

CMU SCS

Topic discovery in documents

Step 1

New stories

Windowing (n=1659) (30 words)

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

$X_{[nxm]}$

aaron zoo

x_i = [1, 5, …, 0]

m=3887 (dictionary size)

Step 2

$X_{[nxm']} = H_{[nxm']} B_{[m'xm']}$

(1) Find hyperplane (m'=10)
(2) Find patterns

Step 3

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

aaron animal zoo

b'_i = [0, 0.7, …, 0.6]

(Q) What does b'_i mean?

CMU SCS

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
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

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.

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

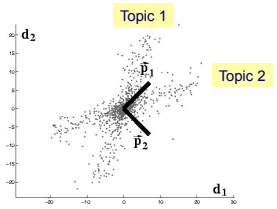
AutoSplit's topics are better than PCA.

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.

CMU SCS

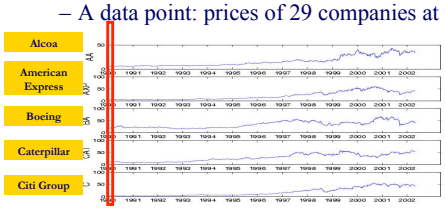
Outline

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

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



CMU SCS

Formulation: Find hidden variables

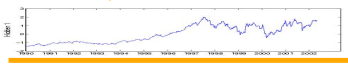
$$\begin{bmatrix} AA_1, \ldots, XOM_1 \\ \ldots \\ \ldots \\ AA_n, \ldots, XOM_n \end{bmatrix} = \begin{bmatrix} H_{11}, H_{12}, \ldots, H_{1m} \\ \ldots \\ \ldots \\ H_{n1}, H_{n2}, \ldots, H_{nm} \end{bmatrix} \begin{bmatrix} B_{11}, B_{12}, \ldots, B_{1m} \\ \ldots \\ B_{m1}, B_{m2}, \ldots, B_{mm} \end{bmatrix}$$

?

?

Date

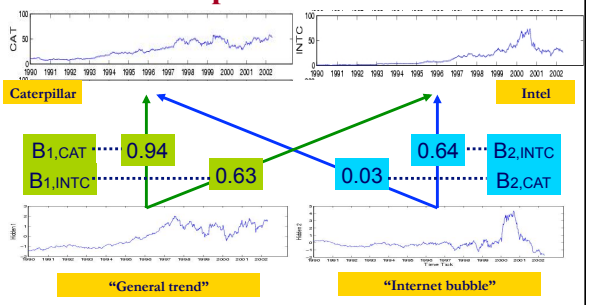
Hidden variable



Date

CMU SCS

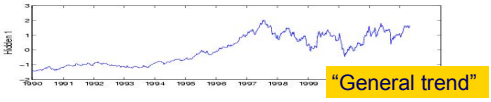
Characterize hidden variable by the companies it influences



CMU SCS

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



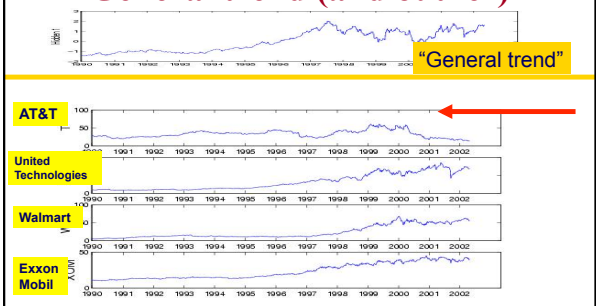
CMU SCS

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.

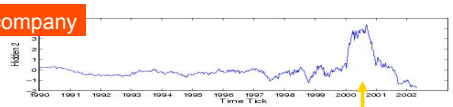
General trend (and outlier)



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"

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

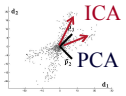
Companies affected by the "internet bubble" variable (with weights 0.5~0.6) are tech-related.
Other companies are un-related (weights < 0.15).

Outline

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

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



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

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. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (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. 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.

References

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

Software

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

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