CMU SCS

15-826: Multimedia Databases and Data Mining

Independent Component Analysis (ICA)

C. Faloutsos

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

CMU SCS

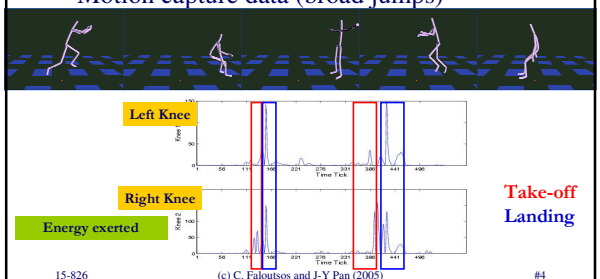
Outline

- Motivation
- Related work: PCA, ICA
- Proposed methods
  - Batch-AutoSplit
  - AutoSplit
  - Clustering-AutoSplit
- Experimental results and discussion
- Conclusions

CMU SCS

Motivation: (Q1) Find patterns in data

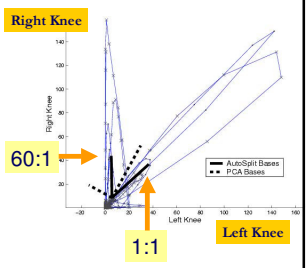
- Motion capture data (broad jumps)



CMU SCS

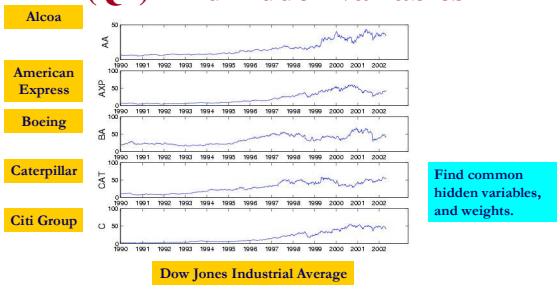
Motivation: (Q1) Find patterns in data

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



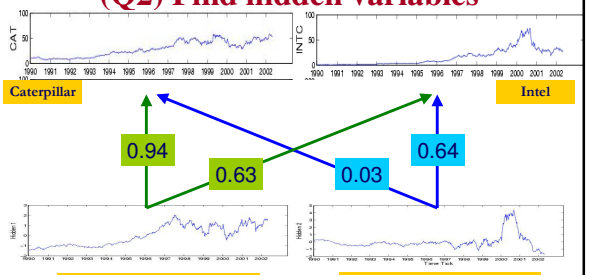
CMU SCS

Motivation: (Q2) Find hidden variables



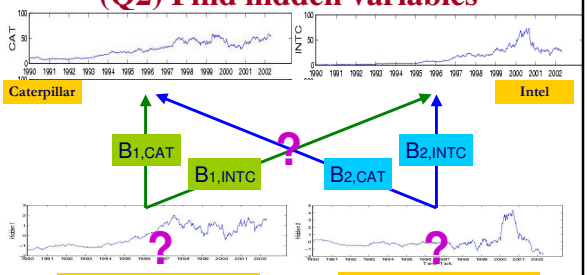
CMU SCS

Motivation: (Q2) Find hidden variables



CMU SCS

Motivation: (Q2) Find hidden variables



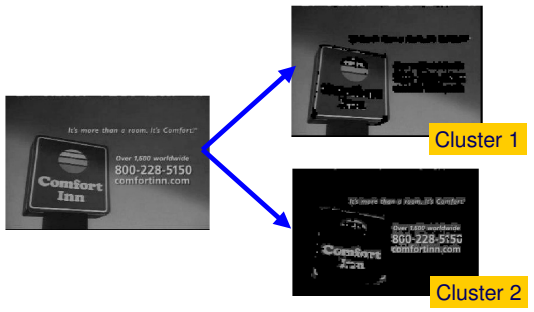
CMU SCS

Motivation: Find hidden variables

- ICA: also known as 'Blind Source Separation'
- 'cocktail party problem'
  - in a party, we can hear two concurrent conversations,
  - but separate them (and tune-in on one of them only)
- http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html
- (in stocks: one 'discussion' is the general economy trend; the other 'discussion' is the tech-stock boom

CMU SCS

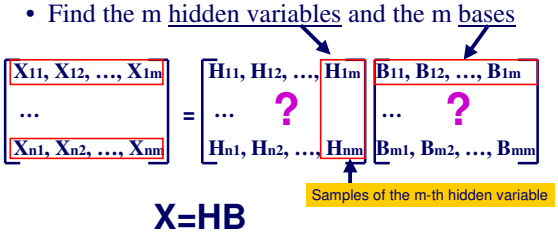
Motivation: (Q3) Segmentation



CMU SCS

Problem formulation

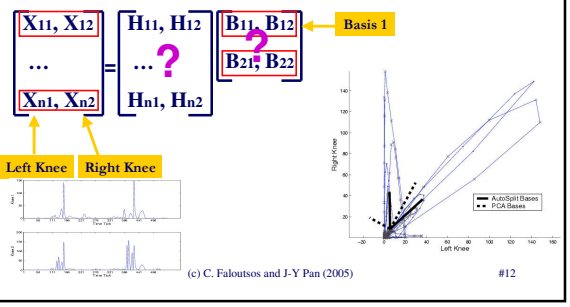
- Given n data items, each has m attributes
- Find the m hidden variables and the m bases



X=HB

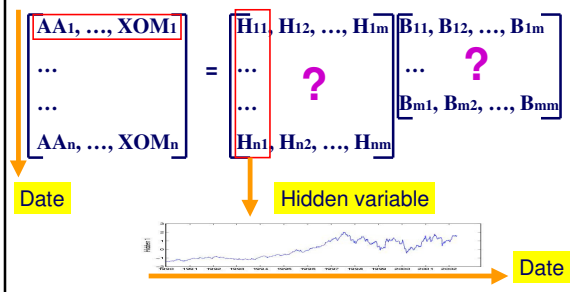
CMU SCS

Formulation: (Q1) Find patterns in data



CMU SCS

Formulation: (Q2) Find hidden variables



CMU SCS

Outline

- Motivation
- Related work: PCA, ICA
- Proposed methods
  - Batch-AutoSplit
  - AutoSplit
  - Clustering-AutoSplit
- Experimental results and discussion
- Conclusions

CMU SCS

Related Work: PCA

- Goal: Knowing **X**, find **H** and **B**, where
  X = H B
- Problem: Under-constrained

CMU SCS

Related Work: PCA

- PCA says
  - Choose bases/rows in B which are orthonormal, and
  - Find such bases that give smallest representation L2 error (for dimensional reduction)
- Matrices H and B can be solved by
  - SVD, neural networks, or many optimization methods

CMU SCS

Related Work: PCA

- Extremely popular
  - Latent Semantic Indexing [Deerwester+90]
  - KL transform [Duda,Hart,Stork00]
  - EigenFace [Turk,Pentlend91]
  - Multiple time series correlation [Guha,Gunopulos,Koudas03]
- But, there is room for improvement.

CMU SCS

Related work: ICA

- ICA says
  - Let h_i's (columns of H) be mutually independent.
- Many implementations
  - "Independence = ?":
    - No mutual information, Kurtosis, nonlinear decorrelation, 'sparse coding'
  - To solve for H,B:
    - Neural networks, optimization methods (gradient ascent, fixed-point, …)

CMU SCS

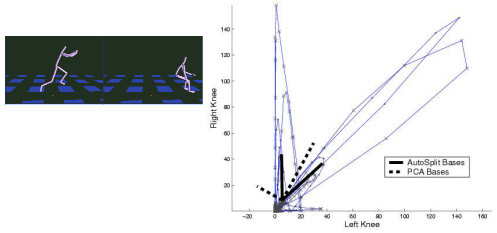
Outline

- Motivation
- Related work: PCA, ICA
- Proposed method: AutoSplit
  - Batch-AutoSplit
  - AutoSplit
  - Clustering-AutoSplit
- Experimental results and discussion
- Conclusions

CMU SCS

PCA sometimes misses essential features

- Best SVD axis: not always meaningful!



CMU SCS

Batch-AutoSplit

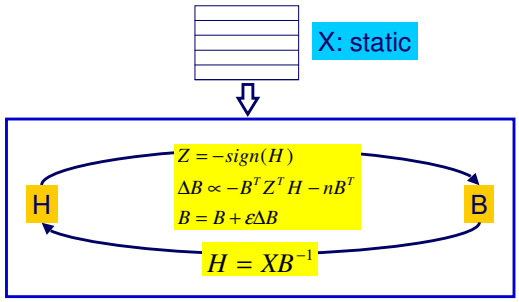
- Assume hidden variables to be "independent"
- Intuition: Independent ~ less redundant

Better data representation

- How to achieve this?

CMU SCS

Batch-AutoSplit

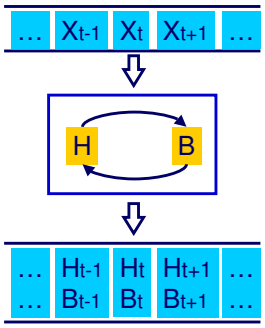


CMU SCS

skip

AutoSplit

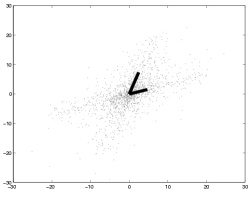
- X: streaming
- X_t: *t*-th window
- Our solution:
  - H, B are estimated iteratively.
  - Adapt to newly arrived data



CMU SCS

Batch-AutoSplit at work

'Sparse coding':
most points have ~zero for one or both coefficients

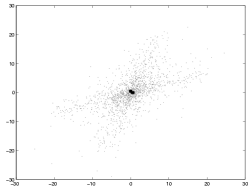


Converge

CMU SCS

Batch-AutoSplit at work

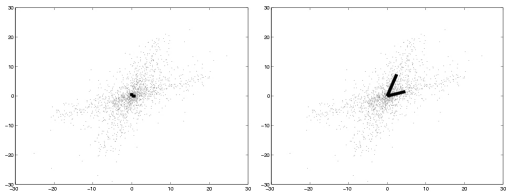
- The estimated bases every 50 iterations, up to iteration 2001.



"The X data set"

CMU SCS

Batch-AutoSplit at work

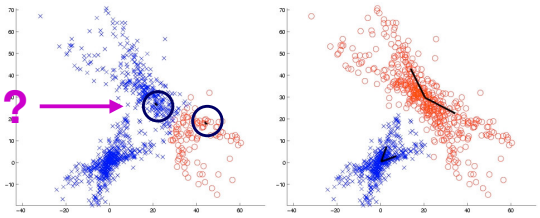


Initialize Converge

CMU SCS

skip

Clustering-AutoSplit at work



Initialize Converge

Where are the centers?
What are the bases?

CMU SCS

Outline

- Motivation
- Related work: PCA, ICA
- Proposed method: AutoSplit
  - Batch-AutoSplit
  - AutoSplit
  - Clustering-AutoSplit
- Experimental results and discussion
- Conclusions

CMU SCS

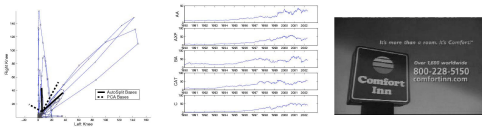
Experimental Results

- Q1: Find patterns (motion capture sequence)
- Q2: Find hidden variables (of stock prices)
- Q3: Separate text from background
- Q4: Text document segmentation
- Q5: Scalability

CMU SCS

Experimental Results

- Q1: Find patterns (motion capture sequence)
- Q2: Find hidden variables (of stock prices)
- Q3: Separate text from background



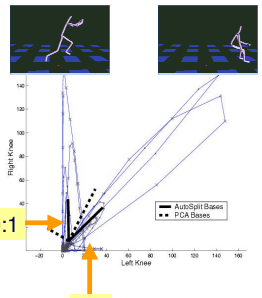
CMU SCS

Q1: Find patterns

- Patterns found
  - Landing: both knees
  - Take-off: right knee
  - Right-handed actor

Take-off

Landing



60:1

1:1

m=2, n=550

CMU SCS

Q2: 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

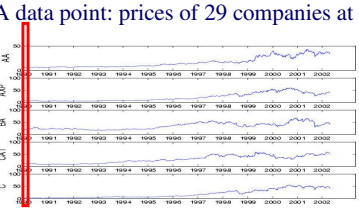
Alcoa

American Express

Boeing

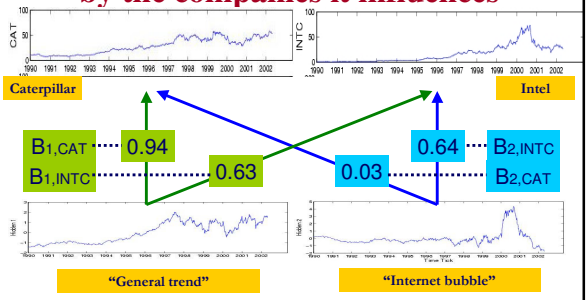
Caterpillar

Citi Group



CMU SCS

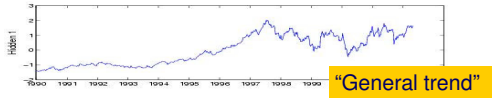
(Q2) Characterize hidden variable by the companies it influences



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

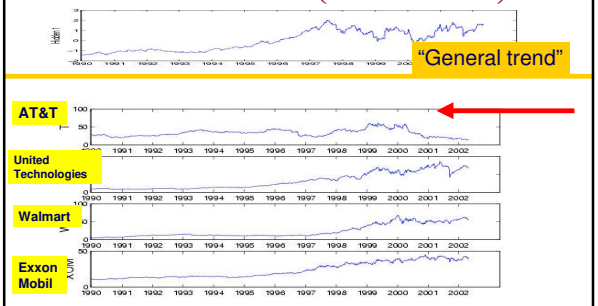
Companies related to hidden variable 1

$B_{1,j}$			
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.

CMU SCS

General trend (and outlier)

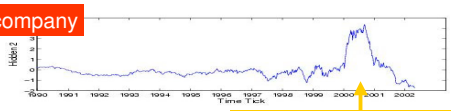


CMU SCS

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"

CMU SCS

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).

CMU SCS

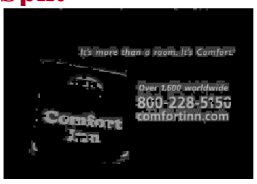
Q3: Separate text from background

- Task: Separate text from background in video frames
- Data points: 6x6 patches from the video frame (50% overlap)
- Use Cluster-AutoSplit to cluster the patches into 2 groups.

CMU SCS

Q3: Separate text from background - Clustering-AutoSplit

Cluster 1



Cluster 2

CMU SCS

Q3: Separate text from background

- Cluster-AutoSplit gives cleaner separation
  - Does not get confused by the background edges

CMU SCS

Q4:Application on text streams

- Task: Segmentation and topic discovery
- Data: CNN headline news (Jan.-Jun. 1998)
- Documents of 10 topics in one single text stream (sorted by date)

CMU SCS

Problem set-up

oprah ... beef ... ... drug ... ...

30 words 30 words ...

Each 30-word window:
a vector in *V*-dimensional space
V: vocabulary size
then,
•we do PCA to 10 eigenvectors/topics,
•and THEN we do ICA:

CMU SCS

The 10 topics

ID	Topic	#Articles
TP_1	Sgt. Gene McKinney is on trial for alleged sexual misconduct	60
TP_2	A bomb explodes in a Birmingham, AL abortion clinic	18
TP_3	The Cattle Industry in Texas sues Oprah Winfrey for defaming beef	45
TP_4	New impotency drug Viagra is approved for use	52
TP_5	Diane Zamora is convicted of helping to murder her lover's girlfriend	22
TP_6	1998 Winter Olympic games	20
TP_7	The Pope's historic visit to Cuba in Winter 1998	39
TP_8	The economic crisis in Asia	69
TP_9	Superbowl XXXII	23
TP_{10}	Tornado in Florida	38

CMU SCS

Process

- Windows of 30 words are taken
  - 50% overlap, 1659 windows
- Word vectors (3887-Dim)
- PCA to 10-Dim **(n=10, T=1659)**
  - Extract PCA bases for comparison
- ICA on the 10-Dim data vectors
  - Extract ICA bases
- Note: 10-Dim bases are mapped back to original space to identify member terms.

CMU SCS

Topics found (PCA)

pc_1	mckinnei	bomb	women	sexual	sergeant
pc_2	bomb	mckinnei	rudolph	clinic	atlanta
pc_3	winfrei	viagra	texa	beef	oprah
pc_4	viagra	winfrei	drug	texa	beef
pc_5	zamora	viagra	winfrei	graham	olymp
pc_6	zamora	graham	kill	viagra	jone
pc_7	pope	cuba	medal	olymp	castro
pc_8	asia	economi	japan	econom	asian
pc_9	bowl	super	re	peopl	medal
pc_{10}	peopl	tornado	super	bowl	florida

CMU SCS

Topics found (PCA)

pc_1	mckinnei	bomb	women	sexual	sergeant
pc_2	bomb	mckinnei	rudolph	clinic	atlanta
pc_3	winfrei	viagra	texa	beef	oprah
pc_4	viagra	winfrei	drug	texa	beef
pc_5	zamora	viagra	winfrei	graham	olymp
pc_6	zamora	graham	kill	viagra	jone
pc_7	pope	cuba	medal	olymp	castro
pc_8	asia	economi	japan	econom	asian
pc_9	bowl	super	re	peopl	medal
pc_{10}	peopl	tornado	super	bowl	florida

CMU SCS

Topics found (ICA)

h_1	mckinnei	sergeant	sexual	major	armi
h_2	bomb	rudolph	clinic	atlanta	birm-ingham
h_3	winfrei	beef	texa	oprah	cattl
h_4	viagra	drug	impot	pill	doctor
h_5	zamora	graham	kill	former	jone
h_6	medal	olymp	gold	women	game
h_7	pope	cuba	castro	cuban	visit
h_8	asia	economi	japan	econom	asian
h_9	super	bowl	game	team	re
h_{10}	peopl	tornado	florida	re	bomb

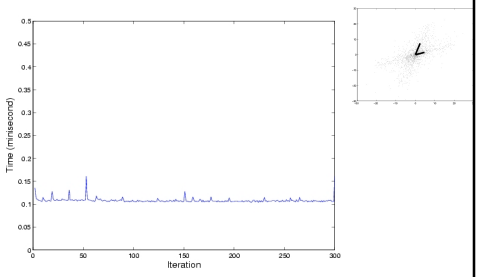
CMU SCS

Topics found (ICA)

h_1	mckinne	sergeant	sexual	major	armi
h_2	bomb	rudolph	clinic	atlanta	birm-ingham
h_3	winfrei	beef	texa	oprah	cattl
h_4	viagra	drug	impot	pill	doctor
h_5	zamora	graham	kill	former	jone
h_6	medal	olymp	gold	women	game
h_7	pope	cuba	castro	cuban	visit
h_8	asia	economi	japan	econom	asian
h_9	super	bowl	game	team	re
h_{10}	peopl	tornado	florida	re	bomb

CMU SCS

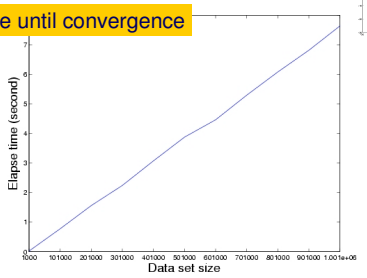
Q5: Scalability
Time to update H and B: O(1)



CMU SCS

Q5: Scale with data set size - linear

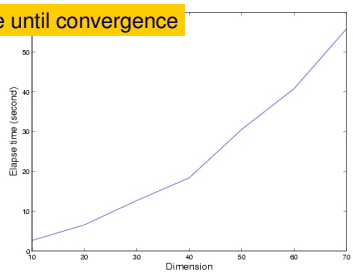
Run time until convergence



CMU SCS

Q5: Scale with dimensionality

Run time until convergence



CMU SCS

Conclusions

- AutoSplit (=**ICA = sparse coding = Blind Source Separation**)
  - Can find patterns (Q1)
  - Can find hidden variables (Q2)
  - **Better than PCA** (Q4)
  - Suitable for streams
  - Scalable (Q5)

CMU SCS

References

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