

15-826: Multimedia (Databases) and Data Mining

Lecture #20:

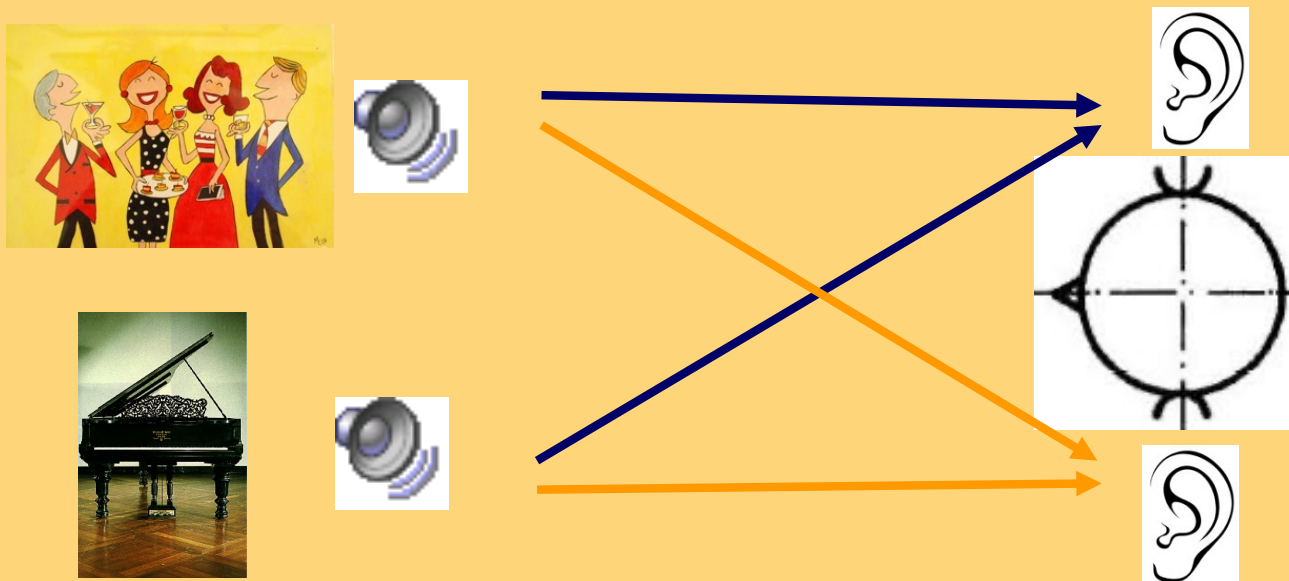
Independent Component Analysis (ICA)

Christos Faloutsos



Problem: BSS

- two sound sources in a cocktail party – separate them

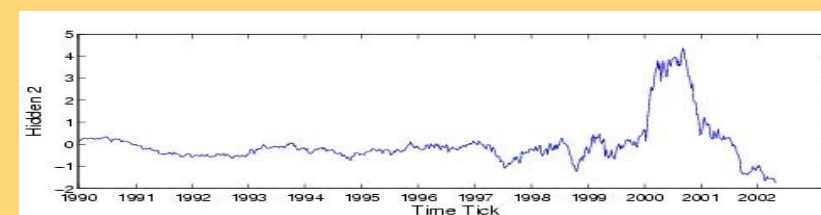
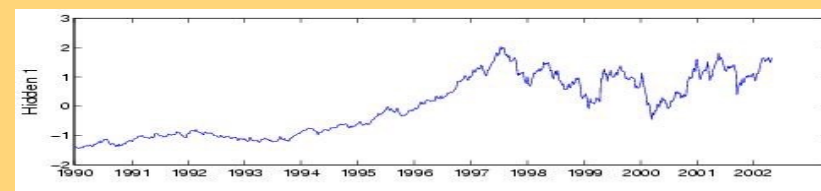
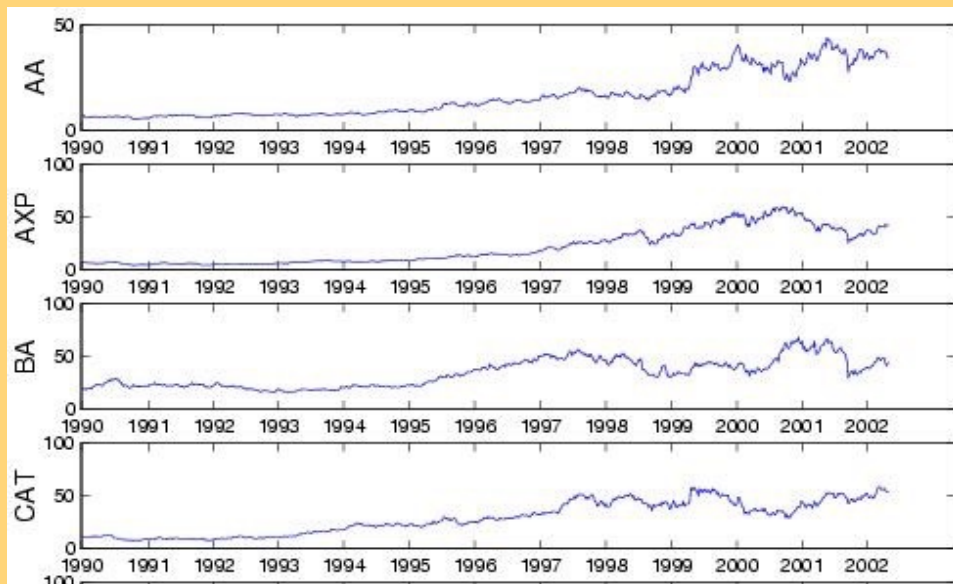


= “blind source separation” (BSS)
(= unknown sources, unknown mixing)



Problem

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

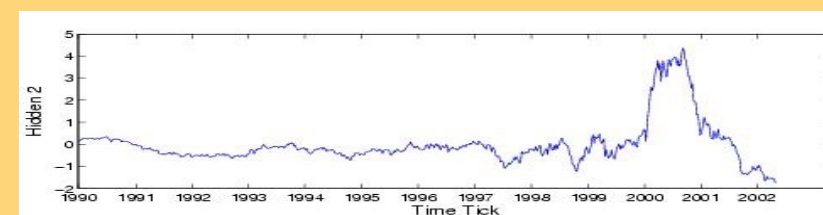
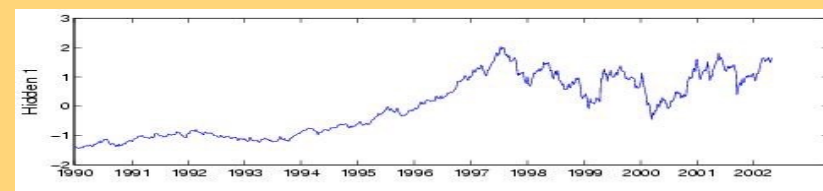
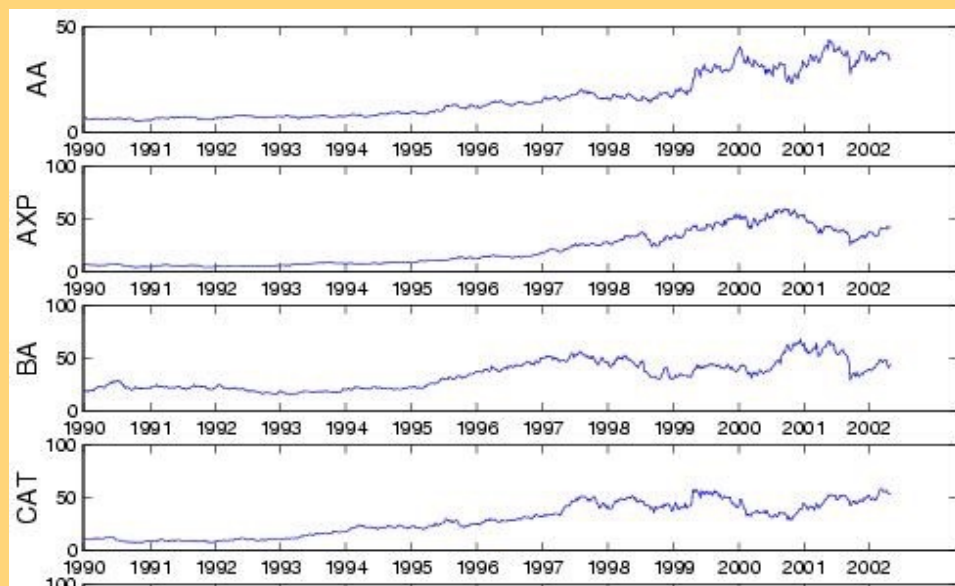
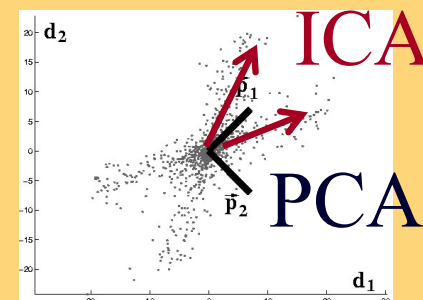




Answer

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

A: ~~SVD~~ ICA



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 - ICA

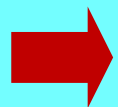
Goal: ‘Find **similar / interesting** things’

- Intro to DB
- ➔ • Indexing - similarity search
- Data Mining



Indexing - Detailed outline

- ...
- fractals
- text
- Singular Value Decomposition (SVD)
 - ...
 - Tensors
 - ICA
- multimedia
- ...





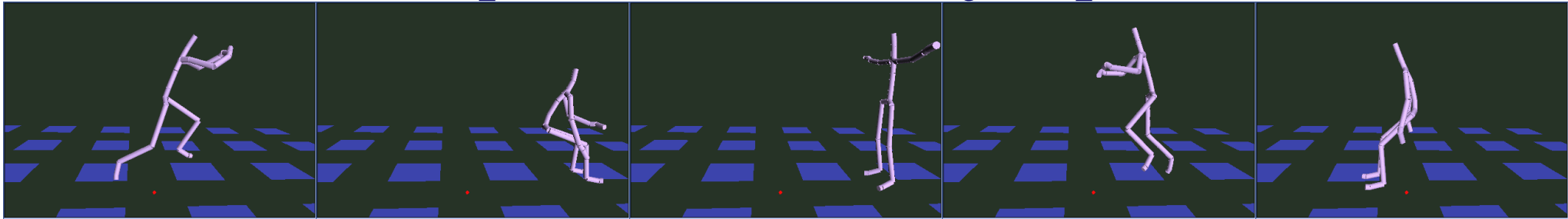
Outline - ICA

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

Motivation:

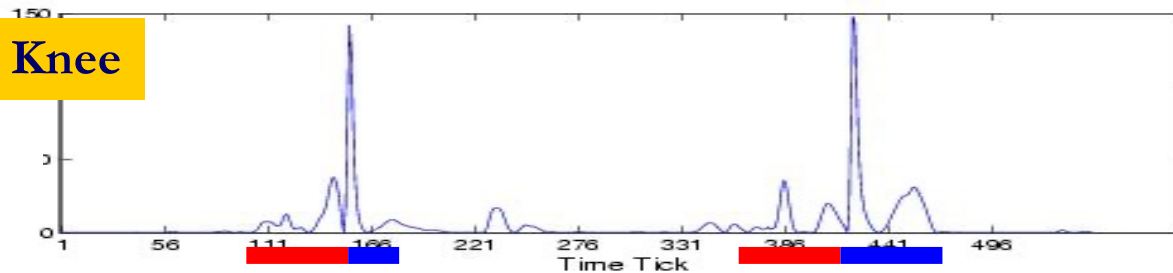
(Q1) Find patterns in data

- Motion capture data: broad jumps



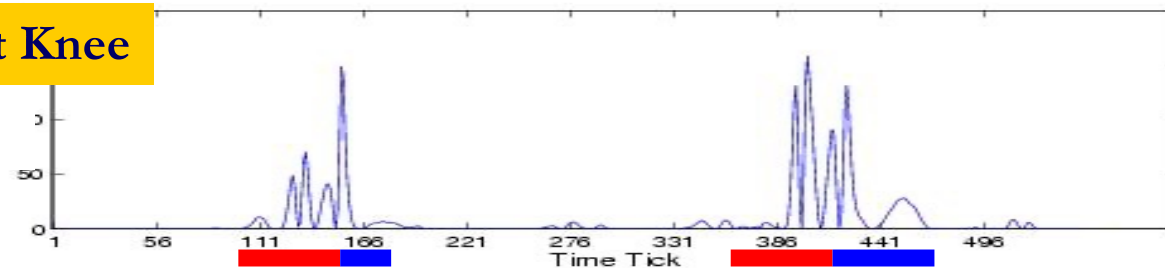
Left Knee

Energy exerted



Right Knee

Energy exerted

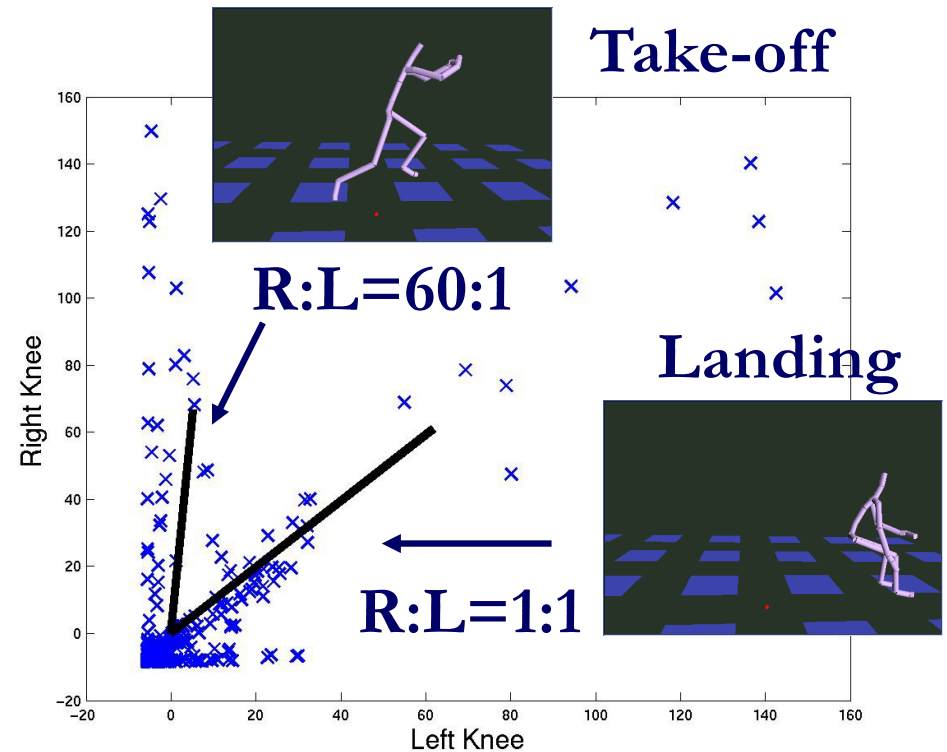


Take-off
Landing

Motivation:

(Q1) Find patterns in data

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

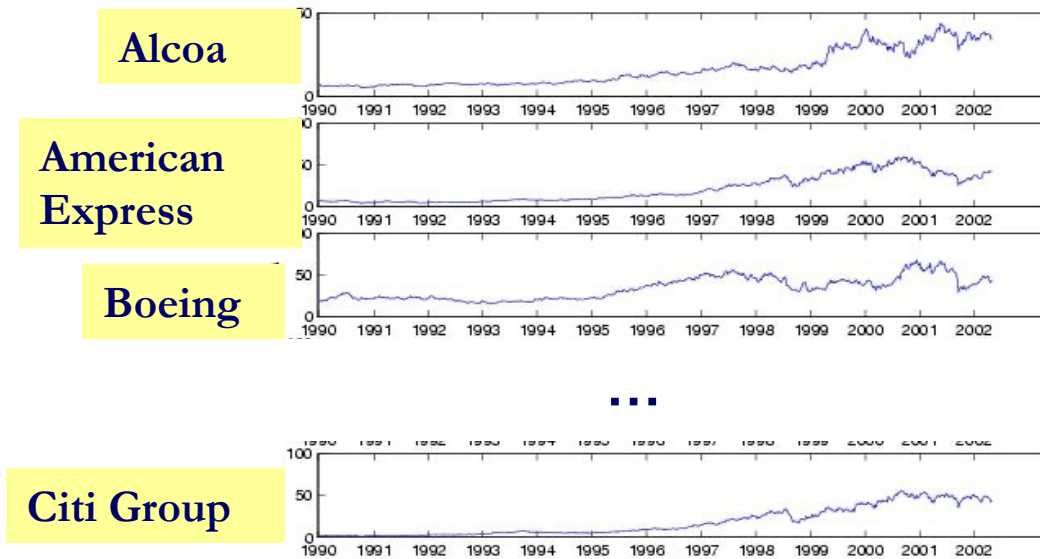


Each point is the measurement
at a time tick (total 550 points).

Motivation:

(Q2) Find hidden variables

Stock prices

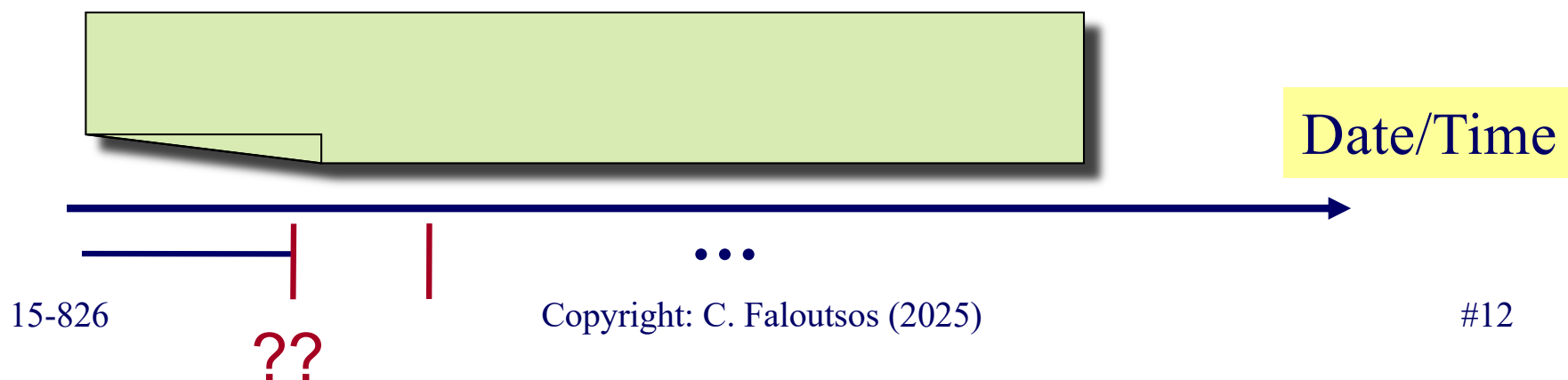


Hidden variables (= 'topics' = concepts)



(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

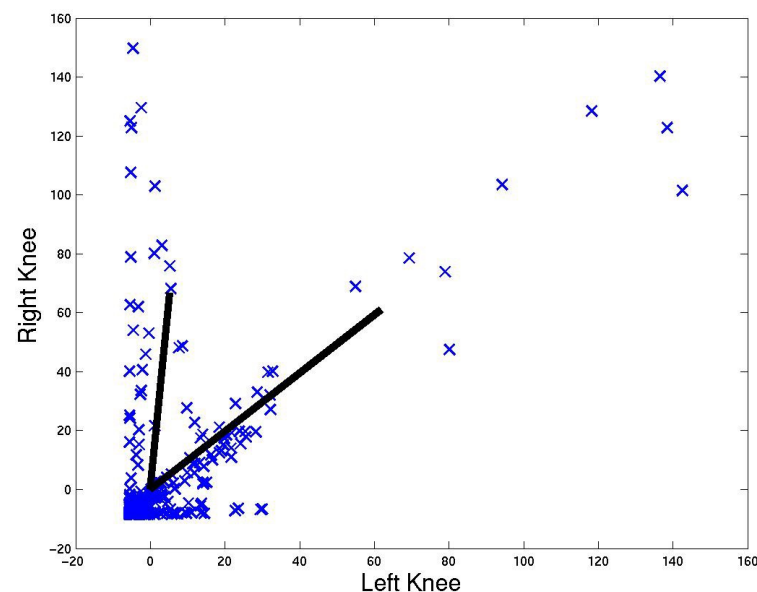
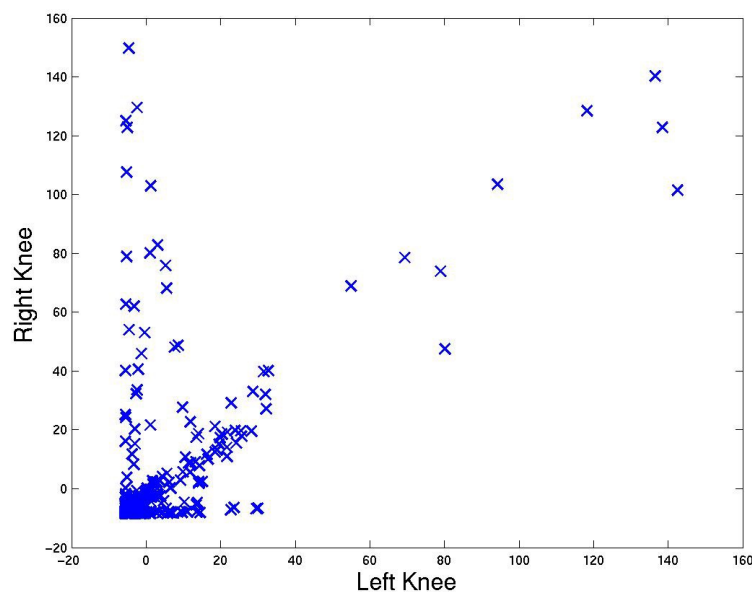




Outline - ICA

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

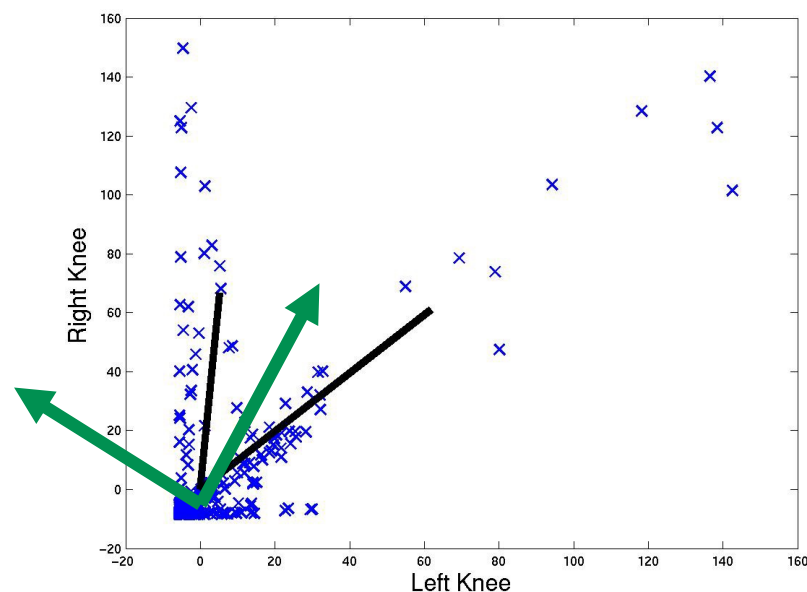
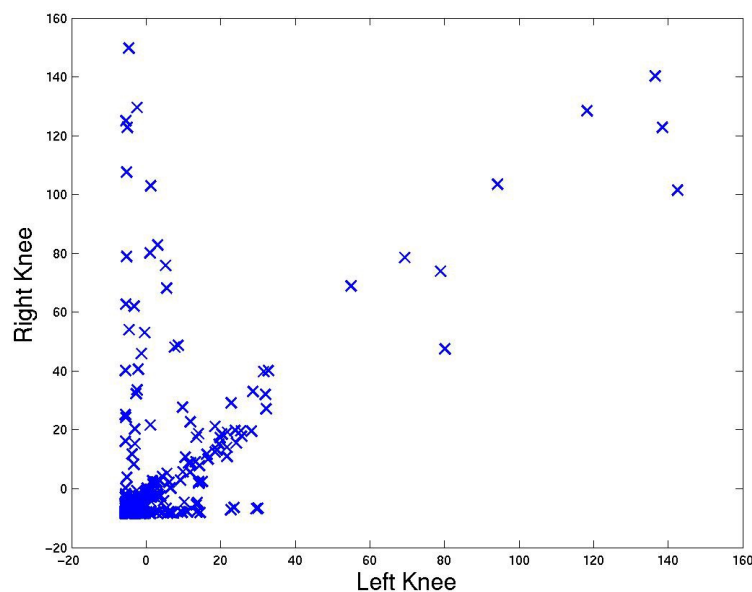
Formulation: Finding patterns



Given n data points,
each with m attributes.

Find patterns that describe
data properties the best.

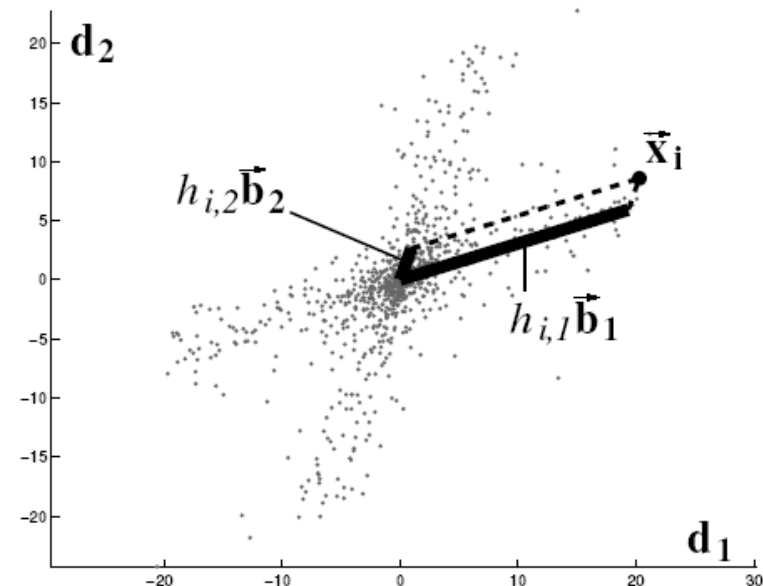
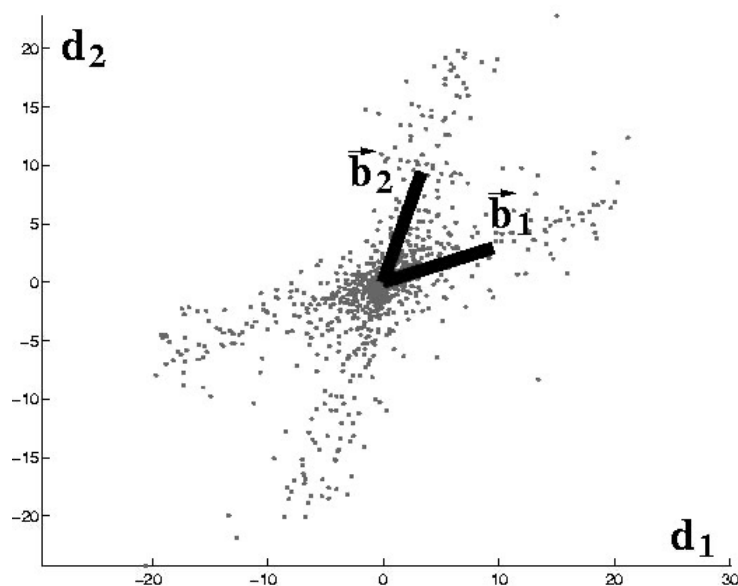
Formulation: Finding patterns



Given n data points,
each with m attributes.

**SVD/PCA: ORTHOGONAL
vectors**

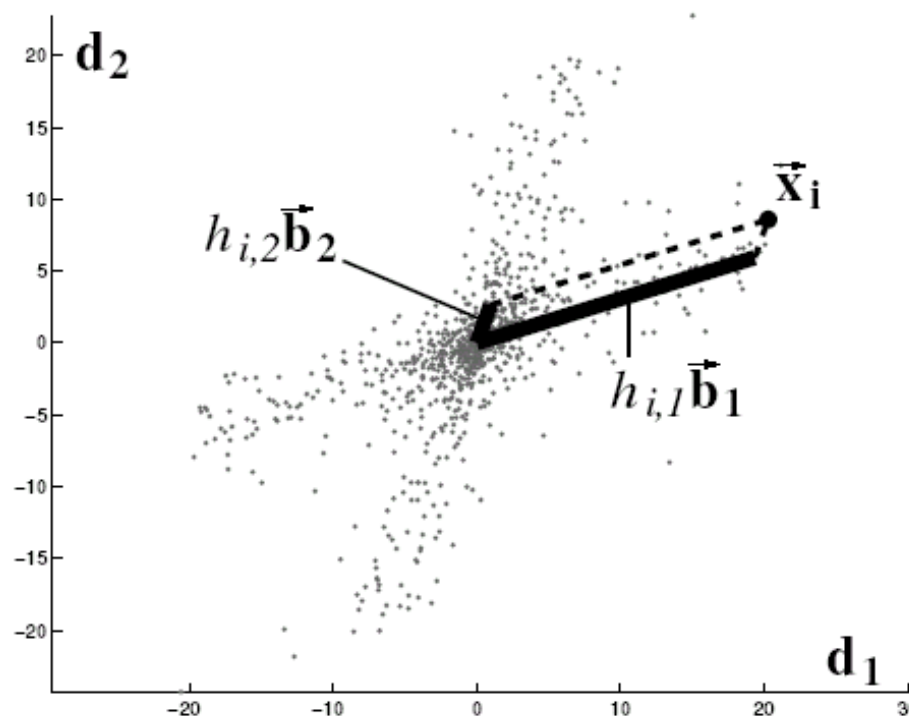
Linear representation



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

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

Patterns as data “vocabulary”



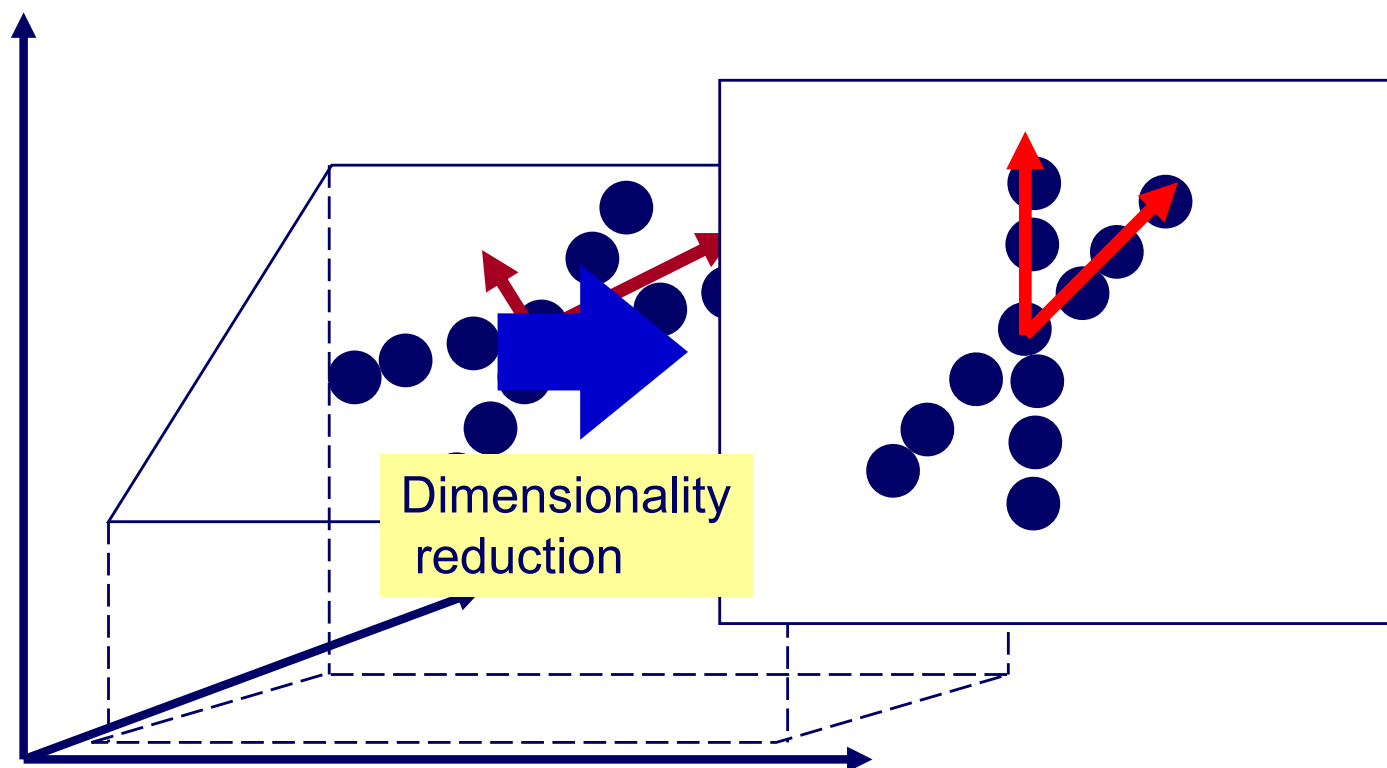
(a) ICA representation of $\vec{x}_i$

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

Good pattern
 $\approx$ sparse coding

b_1 alone, can
 describe x_i .

PCA: first step of ICA



PCA finds the hyperplane.

ICA finds the correct patterns.

Software

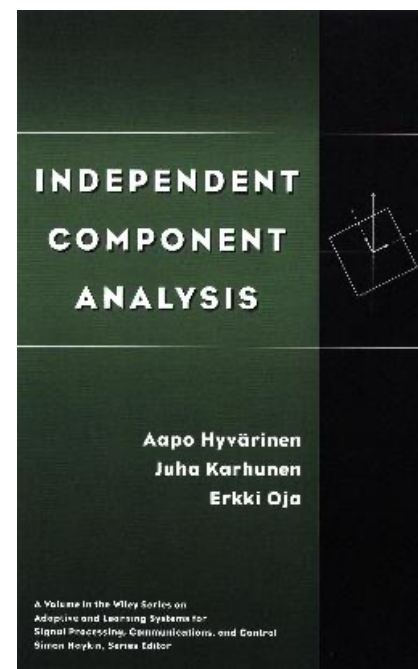
- Open source software: ‘fastICA’
<http://research.ics.aalto.fi/ica/fastica/>

- Or ‘autosplit’ :

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

References

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





Outline - ICA

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

Motivation: Find hidden variables

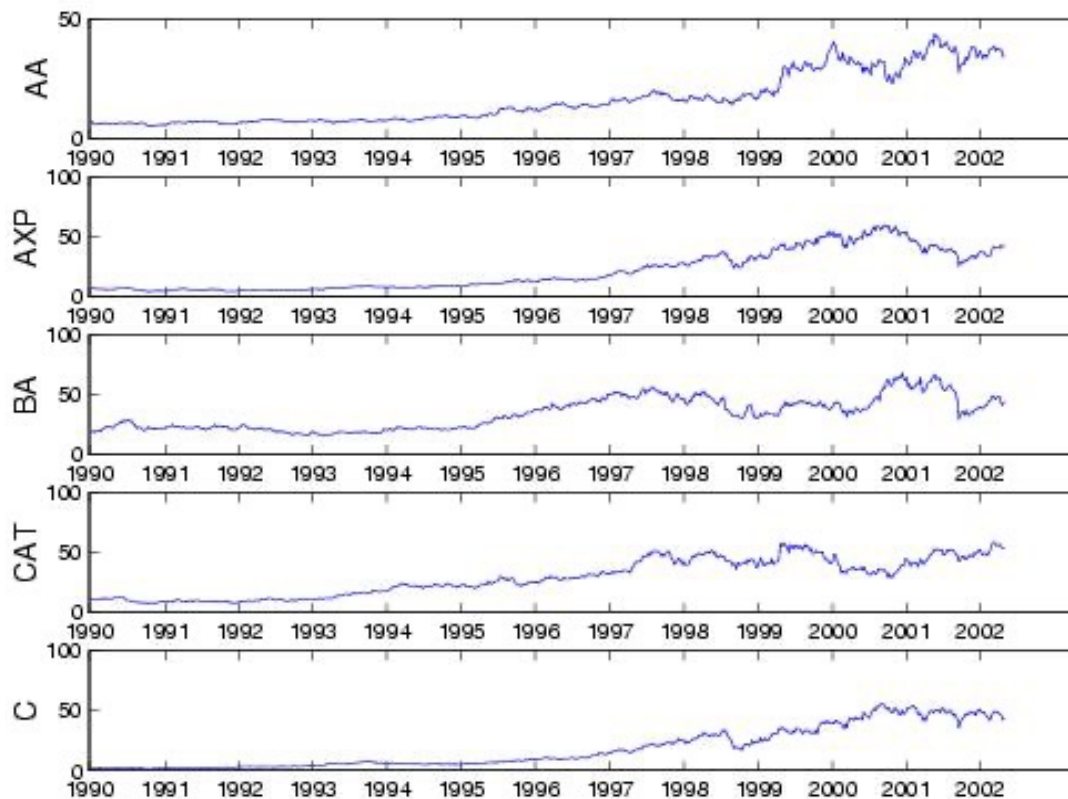
Alcoa

American
Express

Boeing

Caterpillar

Citi
Group



Dow Jones Industrial Average

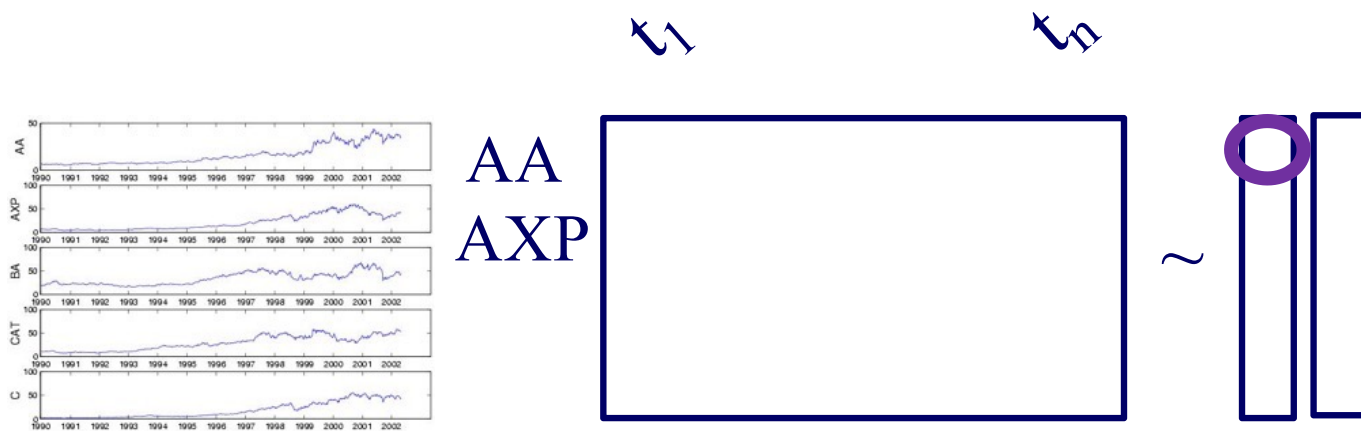
Find common
hidden variables,
and weights.

ICA: Like SVD, but sparse U

1st behavior

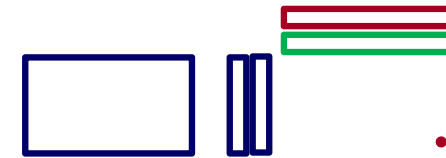


2nd behavior



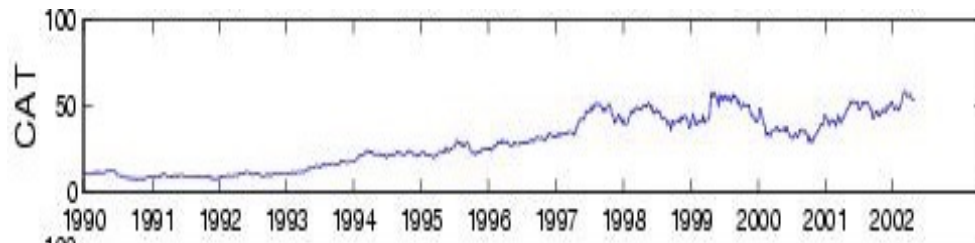
Participation weight of
row i to behavior j

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

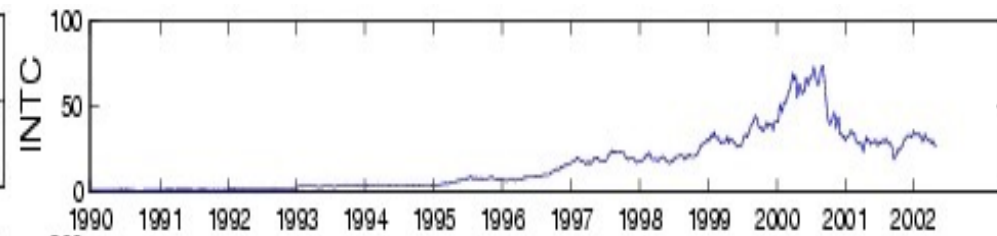


Motivation:

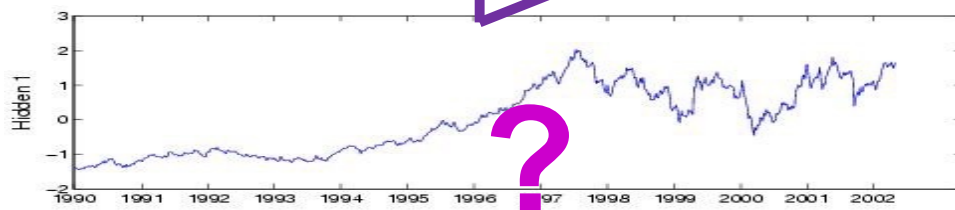
Find hidden variables $\Rightarrow$ behaviors



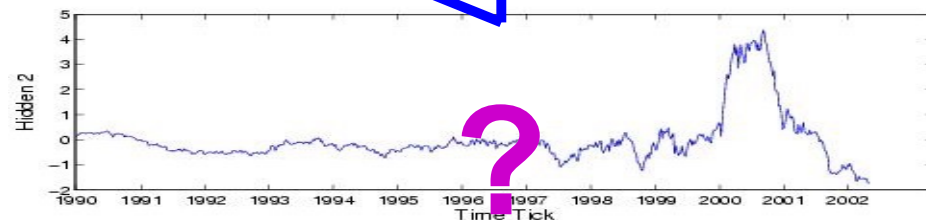
Caterpillar



Intel

B_{1,CAT}B_{1,INTC}B_{2,INTC}B_{2,CAT}

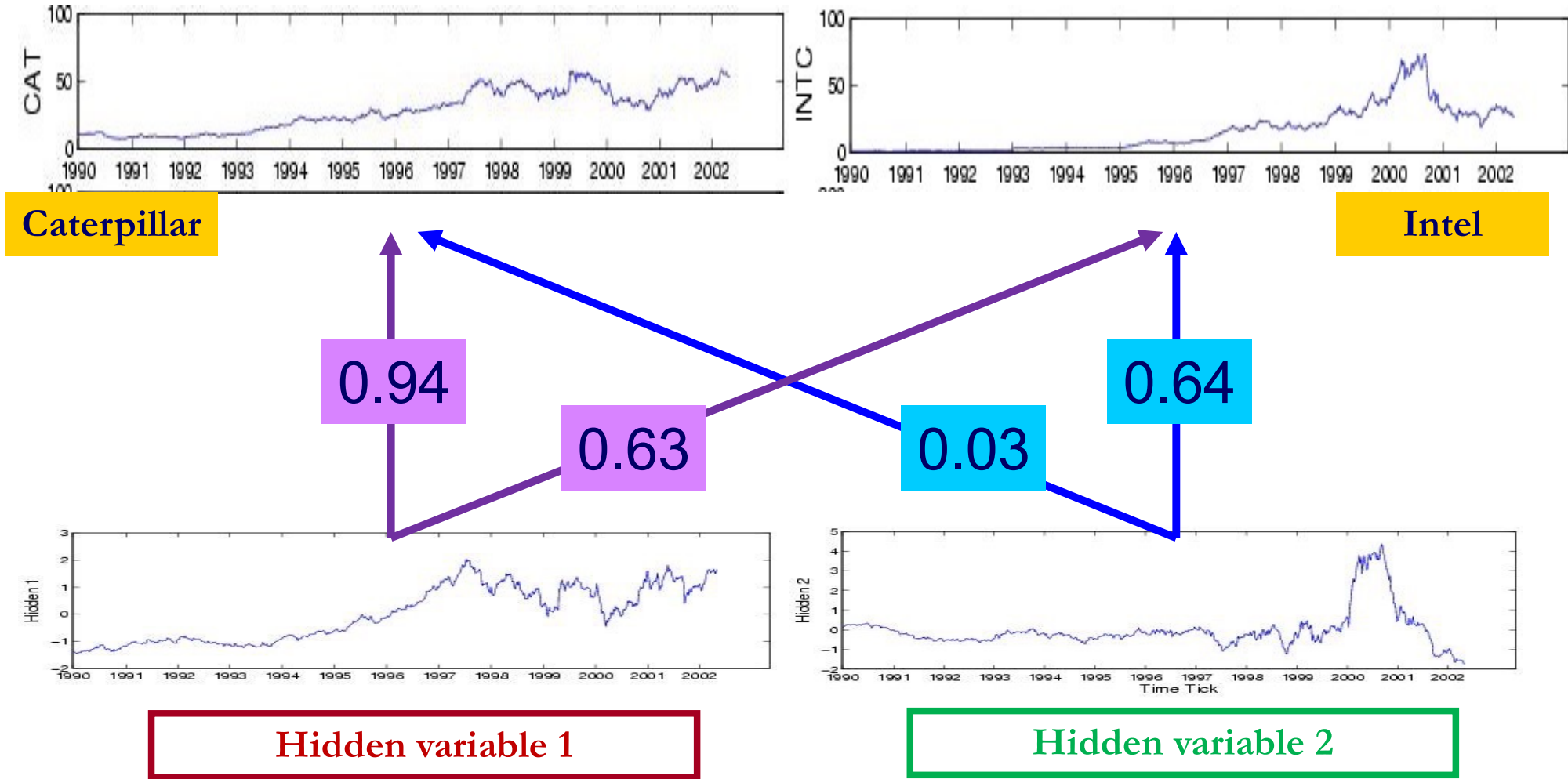
Hidden variable 1



Hidden variable 2

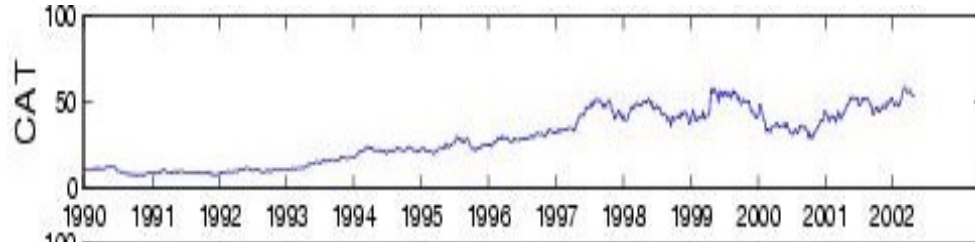
Motivation:

Find hidden variables

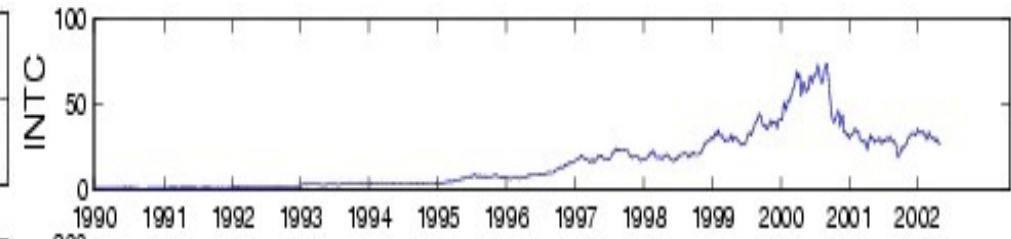


Motivation:

Find hidden variables



Caterpillar



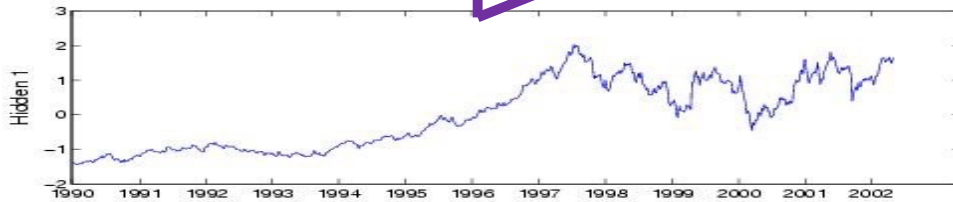
Intel

0.94

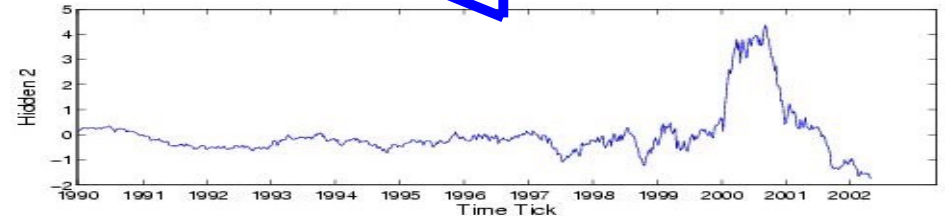
0.63

0.03

0.64



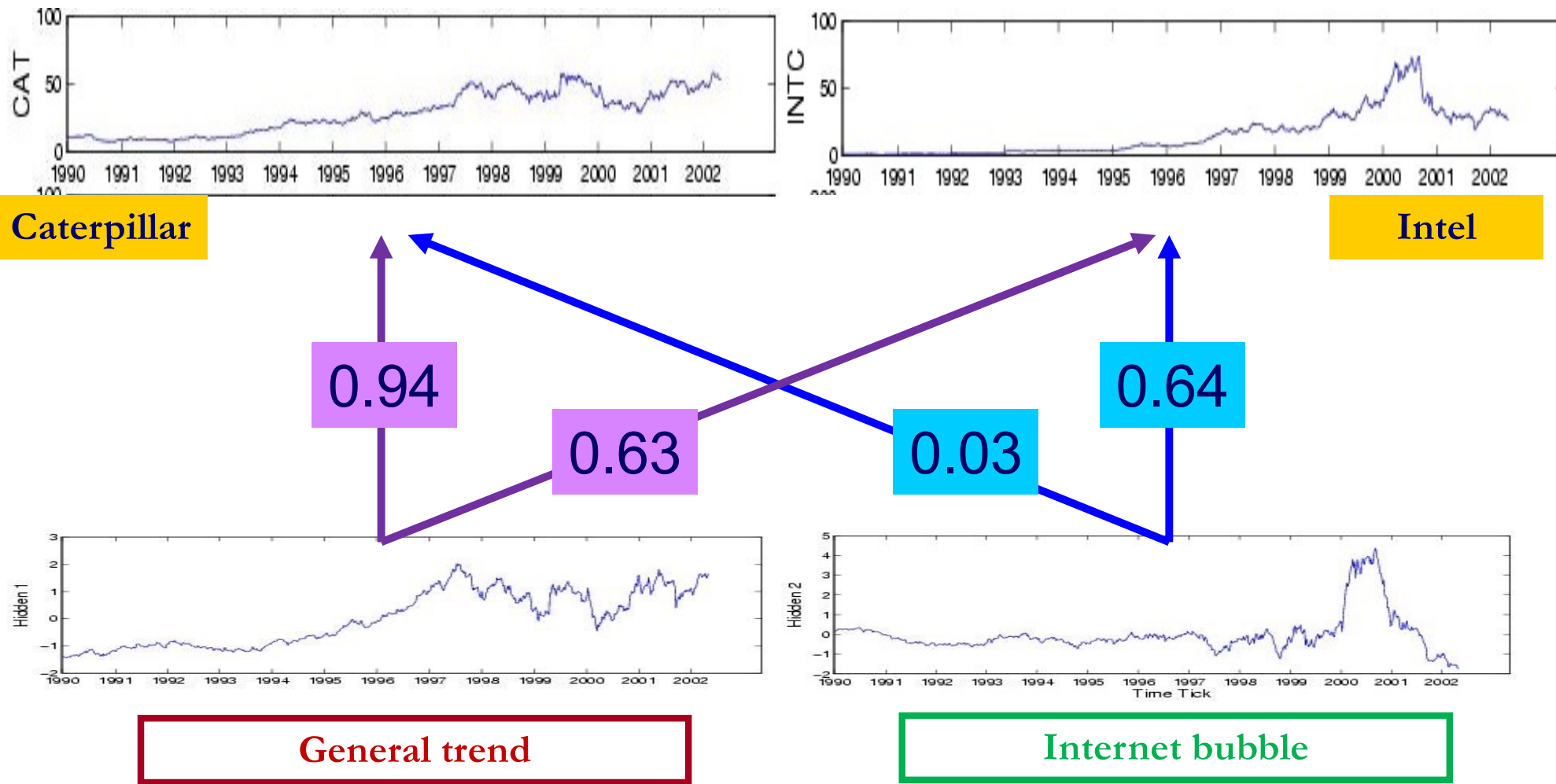
General trend



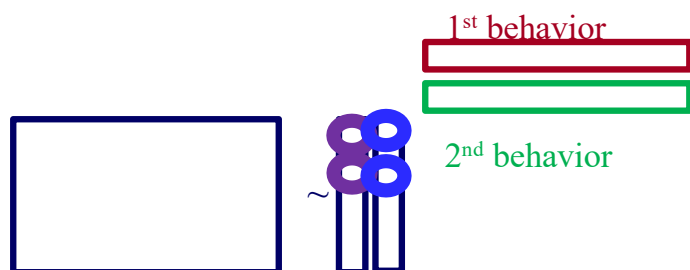
??

Motivation:

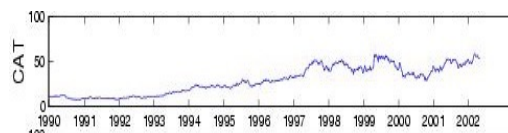
Find hidden variables



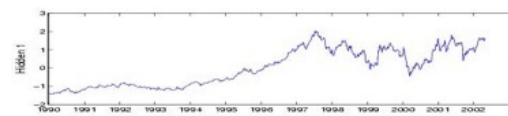
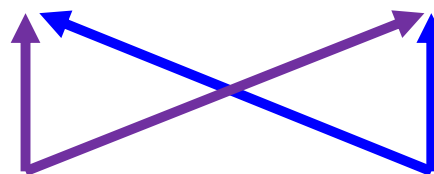
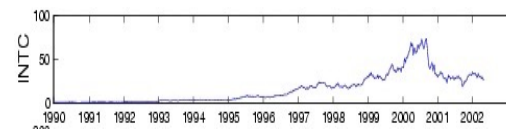
ICA: Like SVD, but sparse



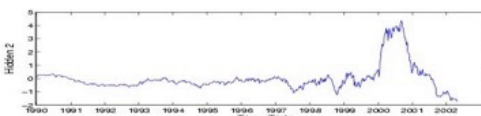
Stock#1



Stock#2

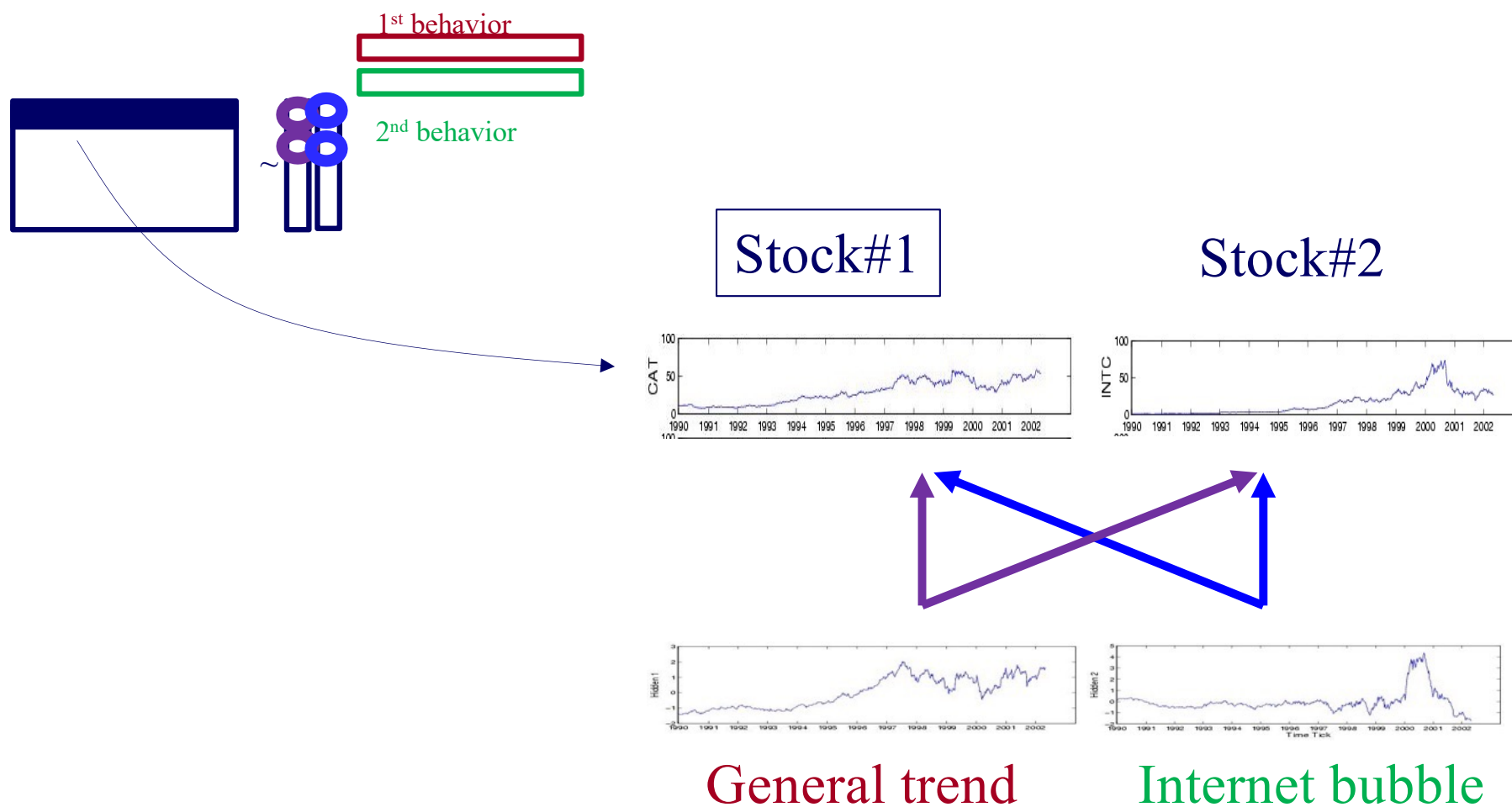


General trend

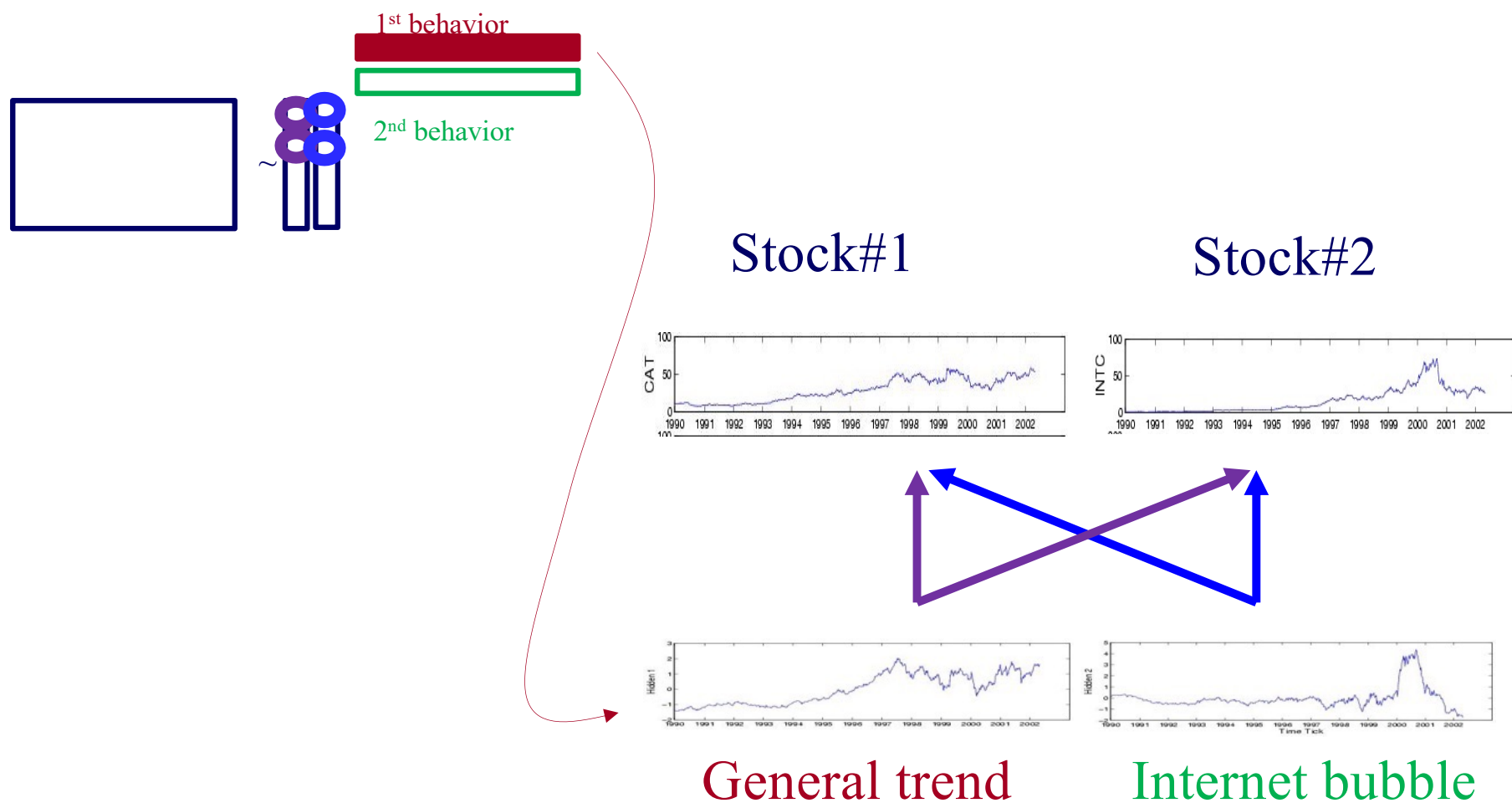


Internet bubble

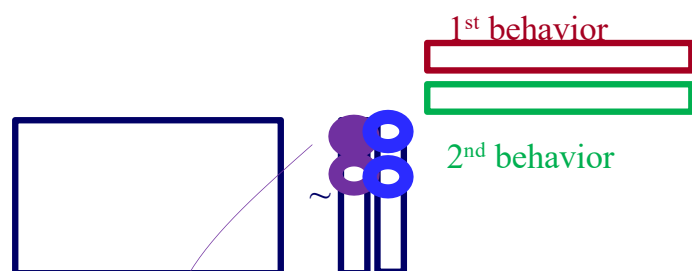
ICA: Like SVD, but sparse



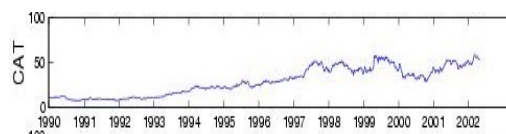
ICA: Like SVD, but sparse



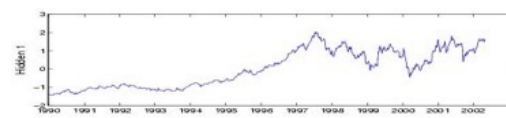
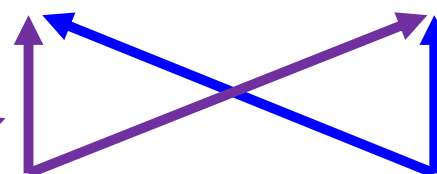
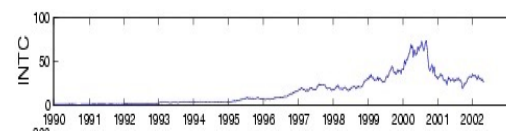
ICA: Like SVD, but sparse



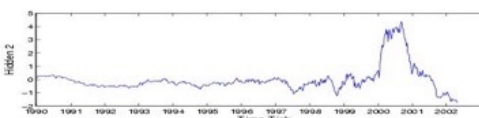
Stock#1



Stock#2



General trend



Internet bubble

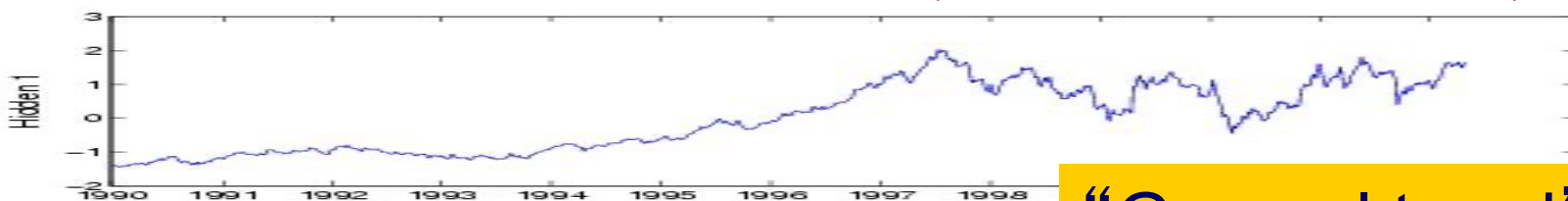
What else can ICA tell us?

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

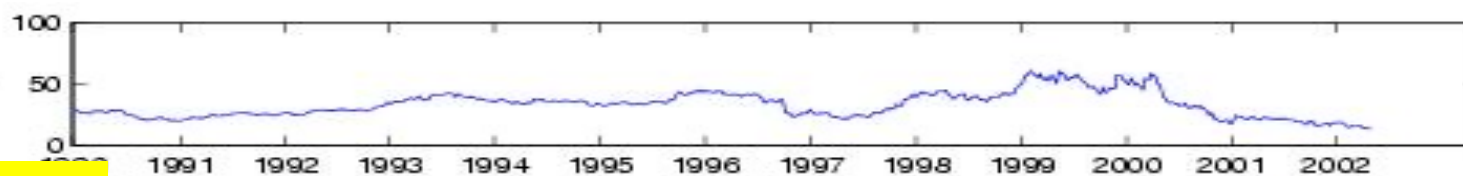
All companies are affected by the “general trend” variable (with weights 0.6~0.9), except AT&T.

General trend (and outlier)

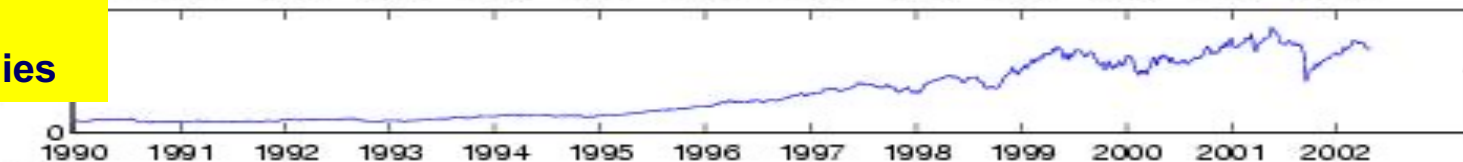


“General trend”

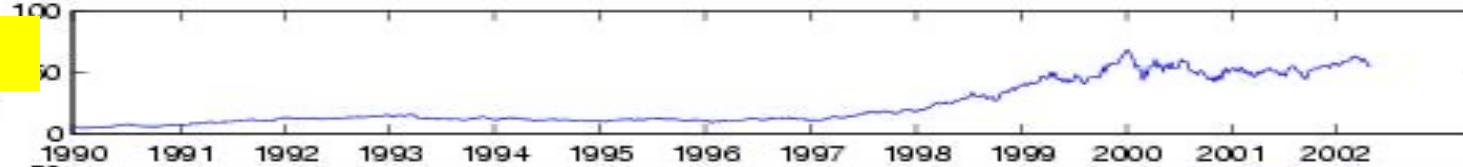
AT&T



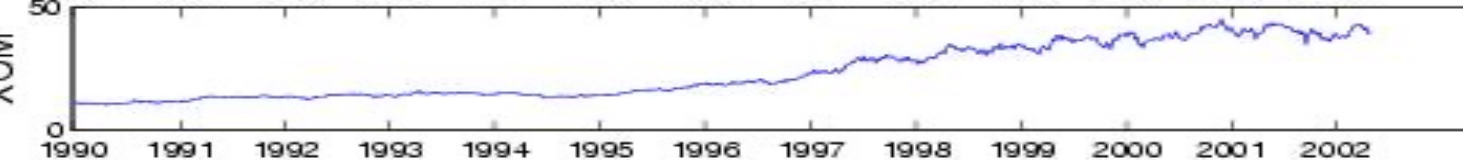
United Technologies



Walmart



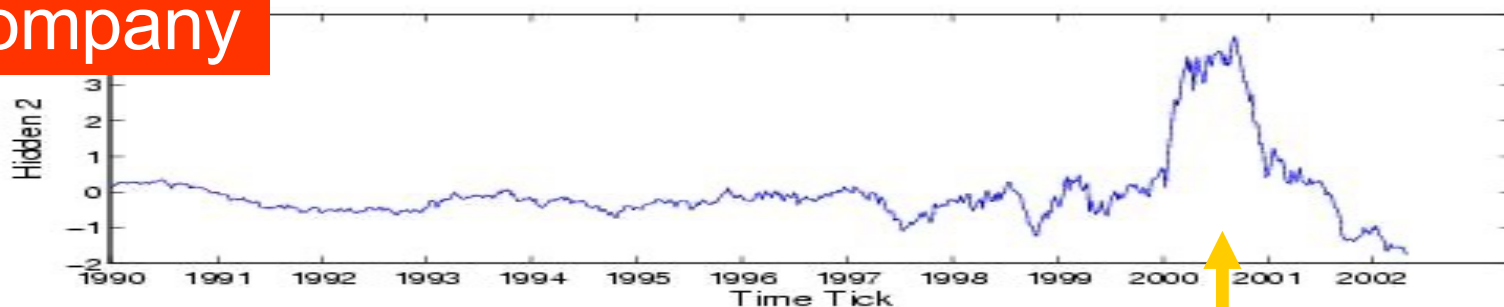
Exxon Mobil



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).

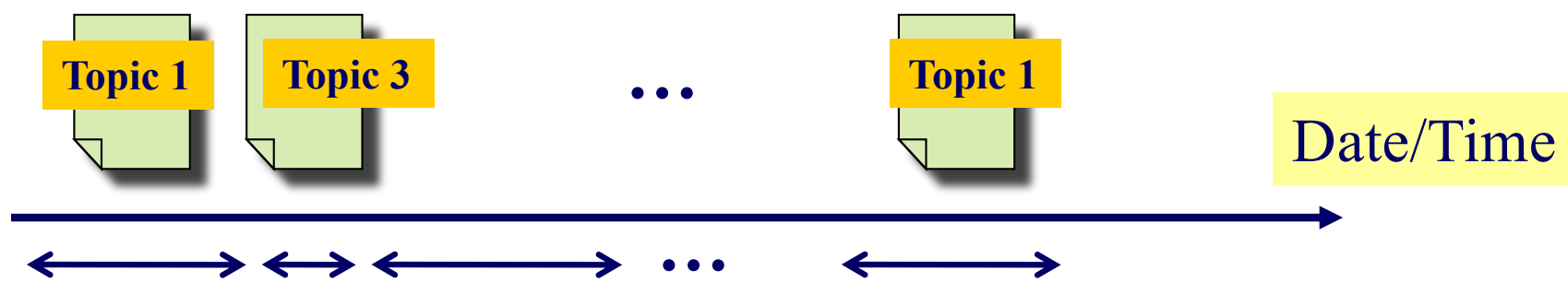


Outline - ICA

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

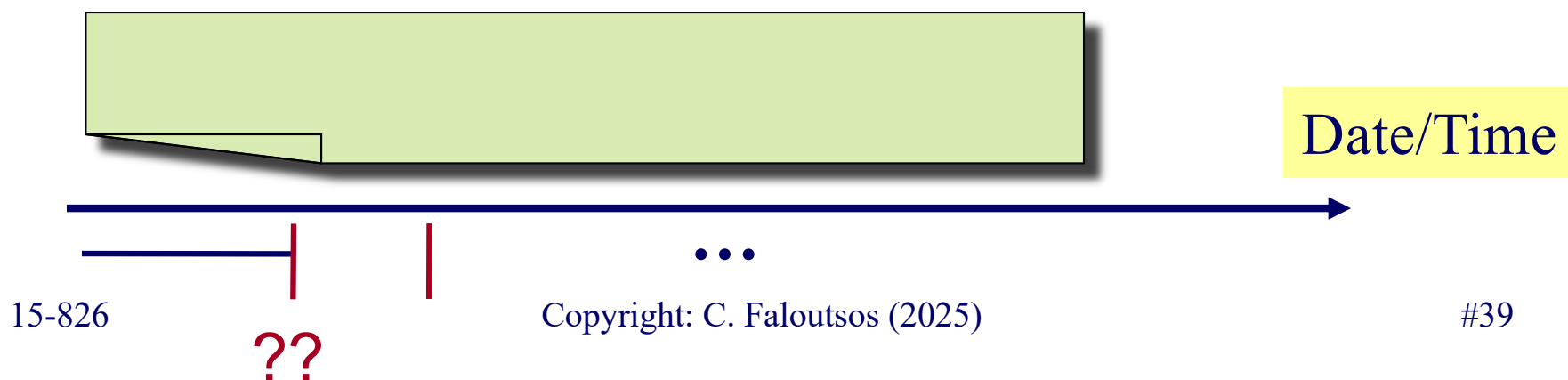
Topic discovery on text streams

- 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

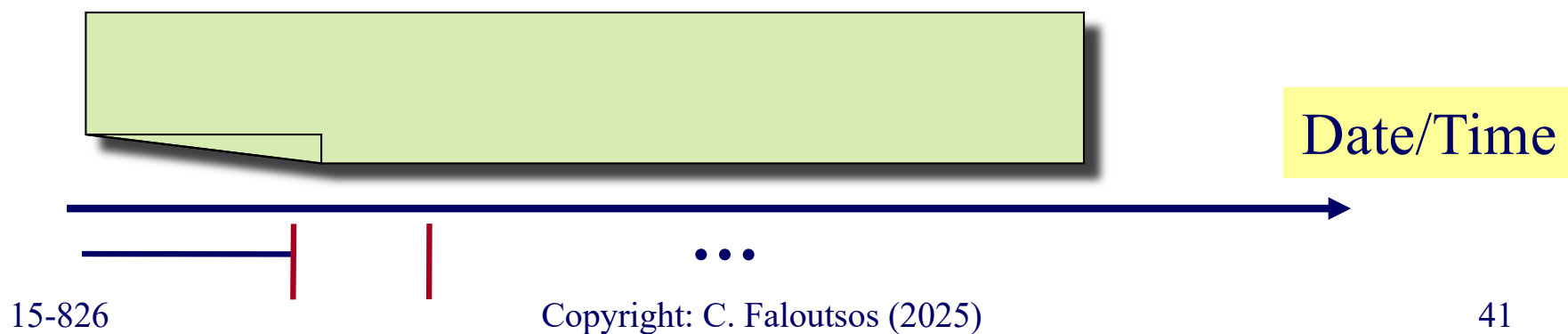


Topic discovery on text streams

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

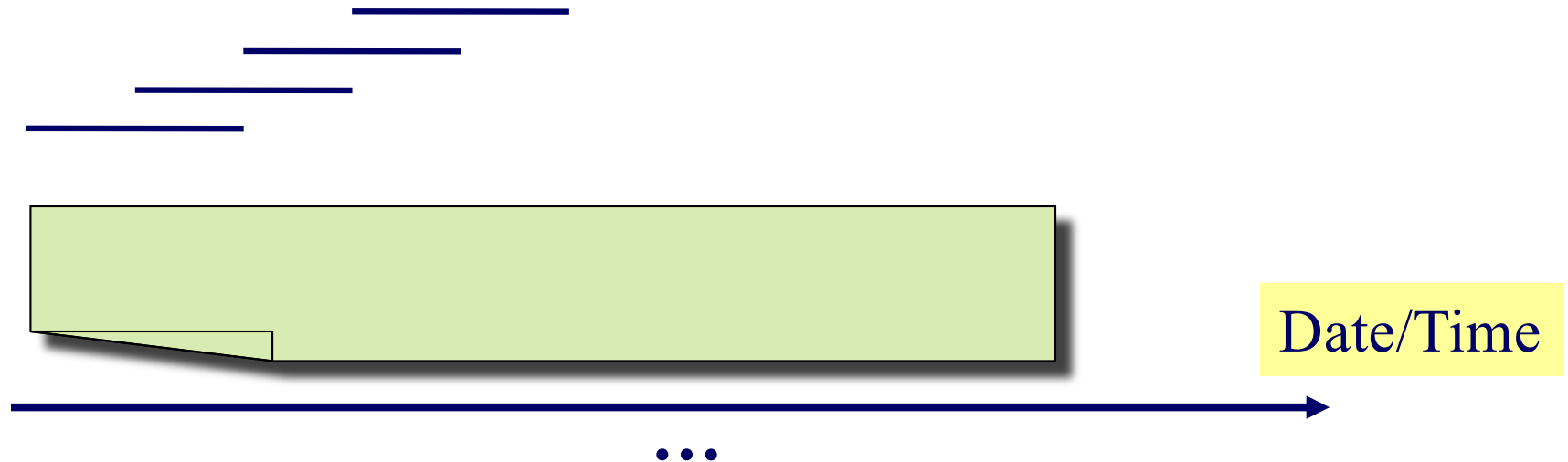


How to proceed?



How to proceed?

- A: Sliding windows



Topic discovery in documents

Step 1

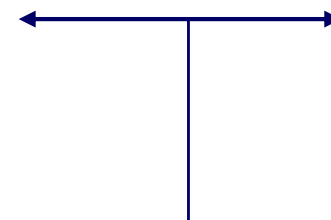


Windowing
(n=1659)
(30 words)

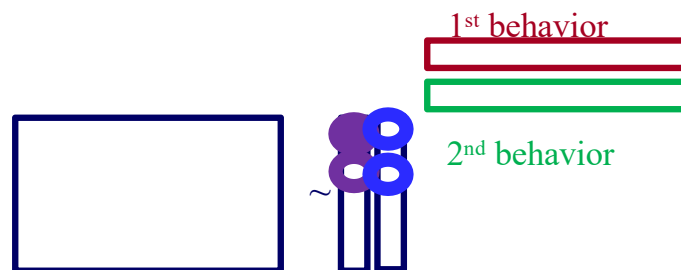
$$\begin{bmatrix} -\vec{x}_1- \\ -\vec{x}_2- \\ \vdots \\ -\vec{x}_n- \end{bmatrix}$$

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

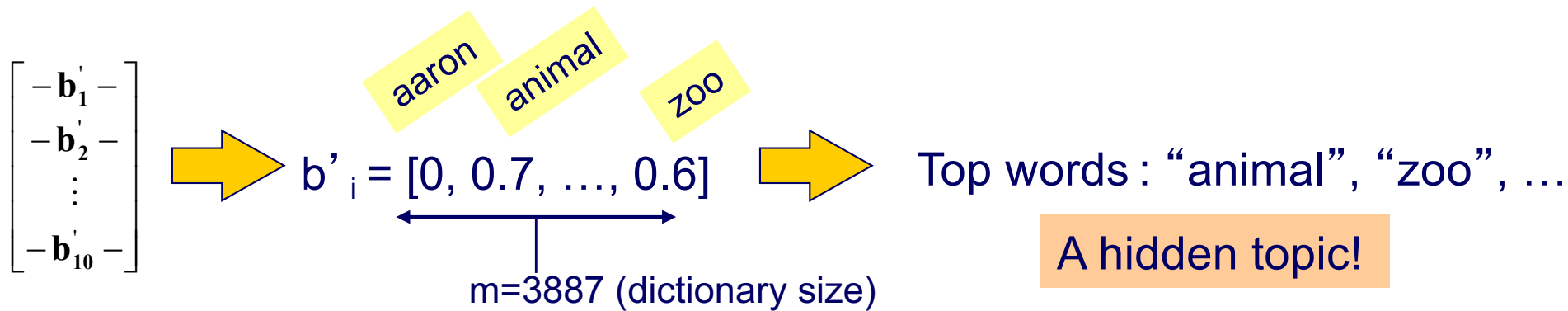
$$\mathbf{x}_i = [1, 5, \dots, 0]$$



m=3887 (dictionary size)



Step 3: Interpret the patterns



Topics found

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

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.

Step 3: Evaluate the patterns

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

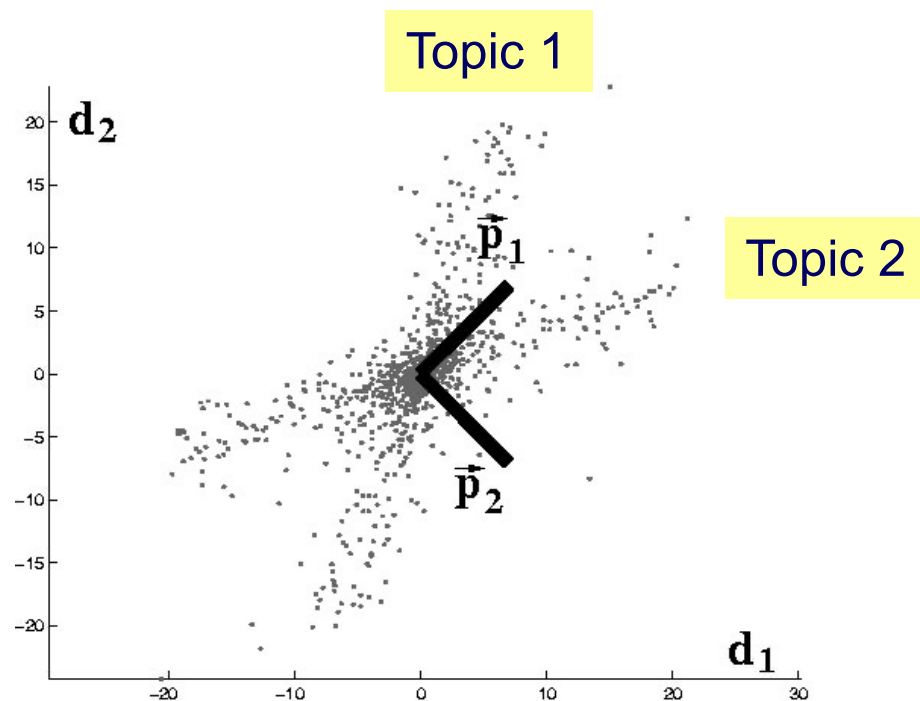
ID	PCA				
<i>A'</i>	mckinne	bomb	women	sexual	sergeant
<i>B'</i>	bomb	mckinne	rudolph	clinic	atlanta
<i>C'</i>	winfrei	viagra	texa	beef	oprah
<i>D'</i>	viagra	winfrei	drug	texa	beef
<i>E'</i>	zamora	viagra	winfrei	graham	olymp

AutoSplit's topics are better than PCA.

Step 3: Evaluate the patterns

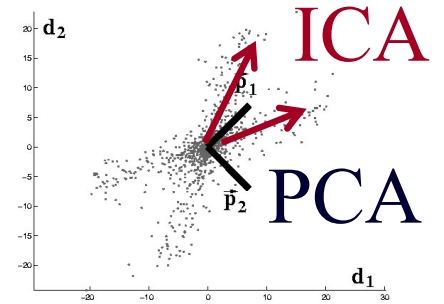
	AutoSplit				
<i>A</i>					
<i>B</i>					
<i>C</i>					
<i>D</i>					
<i>E</i>					

	PCA				
<i>A'</i>					
<i>B'</i>					
<i>C'</i>					
<i>D'</i>					
<i>E'</i>					



PCA vectors mix the topics.

Conclusion



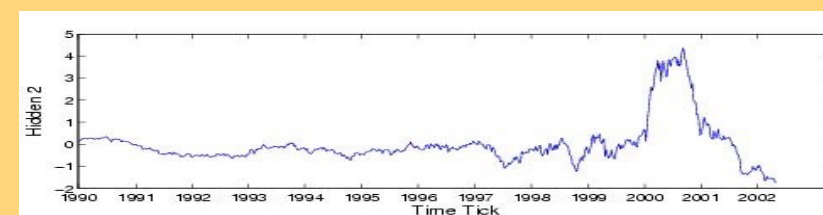
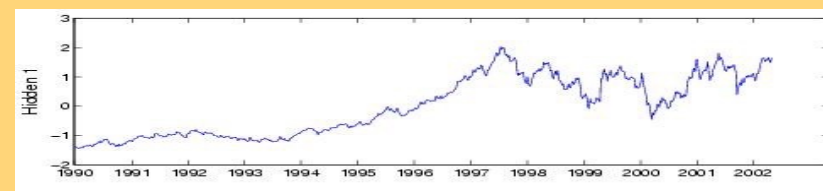
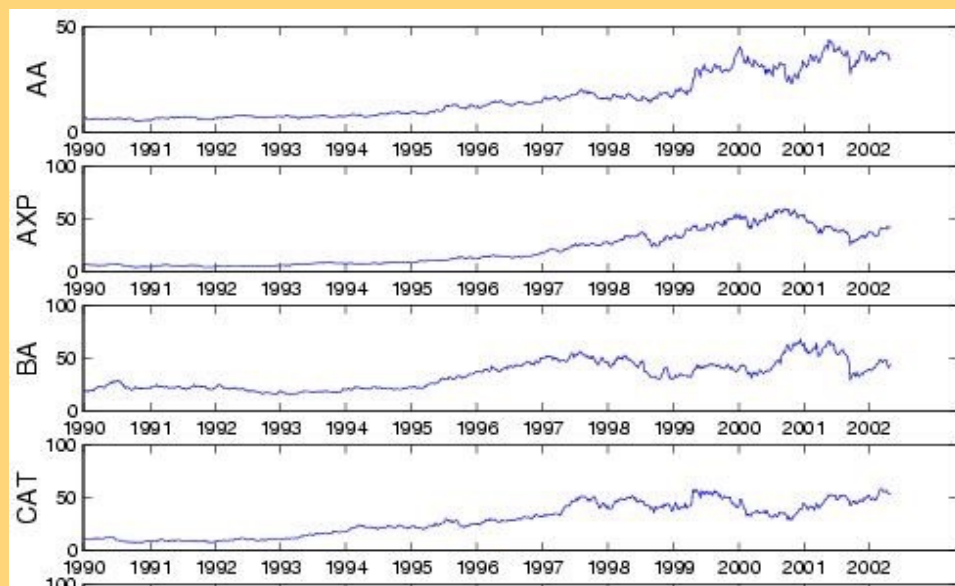
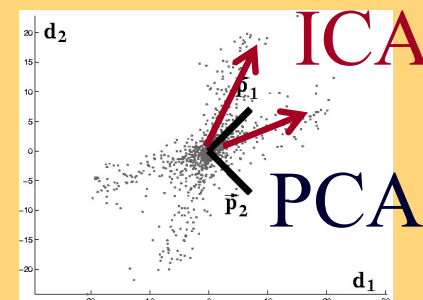
- ICA: more flexible than PCA in finding patterns.
- Many applications
 - Find hidden variables in time series (e.g., stock prices)
 - Blind source separation
- Rule of thumb: plot after PCA;
 - if ‘chicken-feet’, try ICA



Answer

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

A: ~~SVD~~ ICA



SVD, ICA: special cases of Artificial Neural Networks

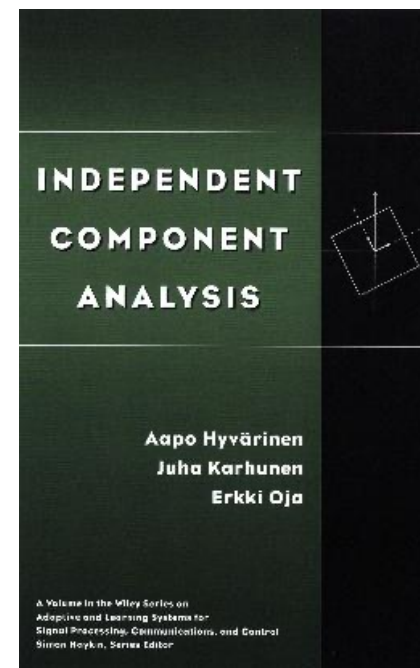
- Aapo Hyvärinen, Juha Karhunen, Erkki Oja: *Independent Component Analysis*

A.N.N. autoencoder with

- Linear transfer function
- 1 hidden layer
- L2 penalty

= SVD

(page 136-137, sec. 6.2.4)



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

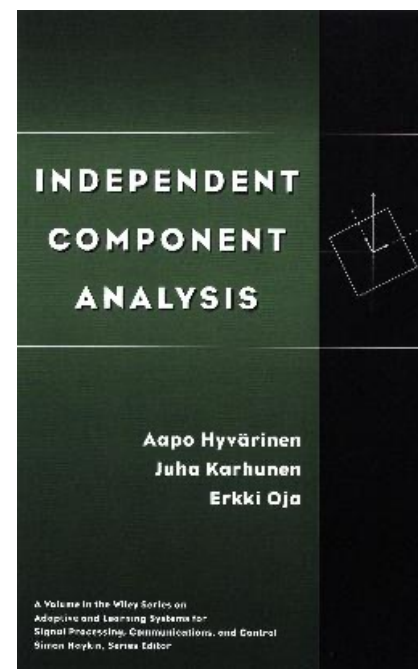


References

- Jia-Yu Pan, Andre Guilherme Ribeiro Balan, Eric P. Xing, Agma Juci Machado Traina, and Christos Faloutsos. *Automatic Mining of Fruit Fly Embryo Images. 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.

References

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



Software

- Open source software: ‘fastICA’
<http://research.ics.aalto.fi/ica/fastica/>
- Also on scikit-learn: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html>
- Or ‘autosplit’ :
www.cs.cmu.edu/~jypan/software/autosplit_cmu.tar.gz