15-826: Multimedia Databases and Data Mining

Independent Component Analysis (ICA)

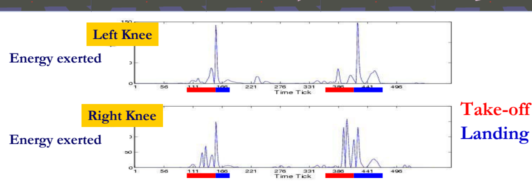
Jia-Yu Pan and Christos Faloutsos

Outline

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

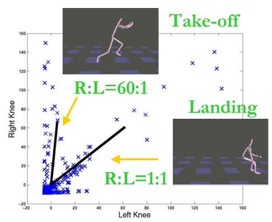
Motivation: (Q1) Find patterns in data

- Motion capture data: broad jumps



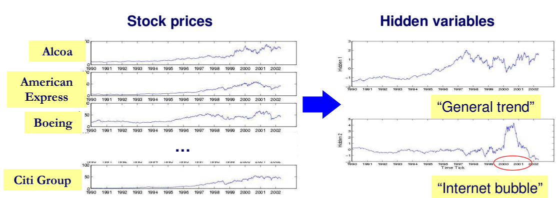
Motivation: (Q1) Find patterns in data

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

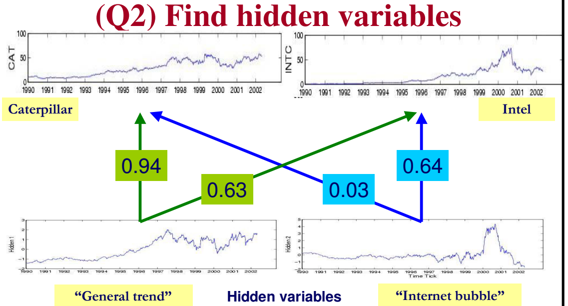


Each point is the measurement at a time tick (total 550 points).

Motivation: (Q2) Find hidden variables

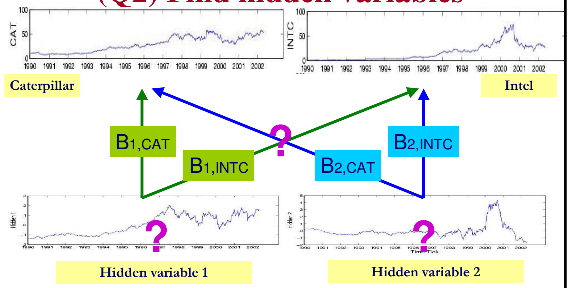


Motivation: (Q2) Find hidden variables



CMU SCS

Motivation: (Q2) Find hidden variables

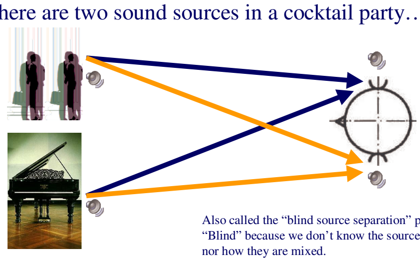


15-826 (c) C. Faloutsos and J-Y Pan (2007) #7

CMU SCS

Motivation: Find hidden variables

- There are two sound sources in a cocktail party…



Also called the "blind source separation" problem.
"Blind" because we don't know the sources, nor how they are mixed.

15-826 (c) C. Faloutsos and J-Y Pan (2007) #8

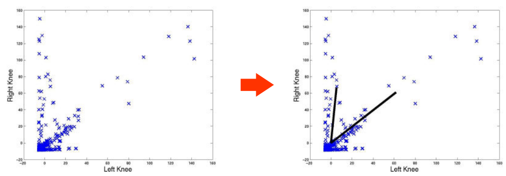
CMU SCS

Outline

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

15-826 (c) C. Faloutsos and J-Y Pan (2007) #9

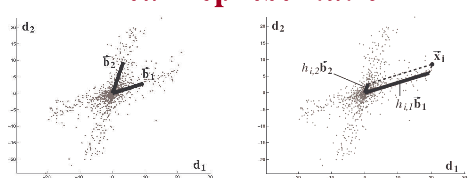
Formulation: Finding patterns



Given n data points, each with m attributes.

Find patterns that describe data properties the best.

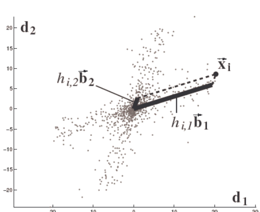
Linear representation



- Find patterns that are vectors that describe the data set the best.
- Each point is described as a linear combination of the vectors (patterns):

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

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

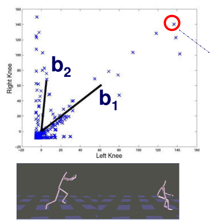
Only b_1 is needed to describe x_i.

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

CMU SCS

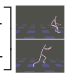
Patterns in motion capture data

Sparse ~ non-Gaussian
~ "Independent"



$$
\begin{array}{cc} \text{Left} & \text{Right} \end{array}
$$

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



$$\mathbf{X}_{\mathbf{n}\mathrm{x}2} = \mathbf{H}_{\mathbf{n}\mathrm{x}2}\mathbf{B}_{2\mathrm{x}2}$$

n=550 ticks

Data matrix — **Hidden variables** — **Basis vectors**

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

15-826 (c) C. Faloutsos and J-Y Pan (2007) #13

CMU SCS

Outline

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

15-826 (c) C. Faloutsos and J-Y Pan (2007) #14

CMU SCS

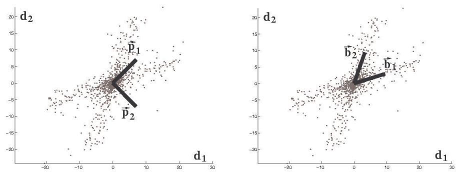
Basis vectors and hidden variables

- Goal: Knowing **X**, find **H** and **B**, where
  X = H B
- Problem: Under-constrained
  - Need additional assumptions/constraints

X: data set
H: hidden variables
B: basis vectors

15-826 (c) C. Faloutsos and J-Y Pan (2007) #15

PCA and ICA



PCA Vectors ICA Vectors

- PCA vectors: major variations
  - Together = good "low-rank approximation"/dimensional reduction
  - Individually ≠ meaningful patterns
- Luckily, ICA detects the major meaningful patterns.

PCA

- PCA (Principal Component Analysis)
  - Choose vectors which are orthonormal and
  - give smallest representation L2 error for dimensional reduction
- Matrices H and B can be solved by
  - SVD, neural networks, or many optimization methods

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.

ICA

- ICA (Independent Component Analysis)
  - Make hidden variables h_i's (columns of **H**) mutually independent.
- Many implementations
  - Many ways to define "independence"
  - Many ways to find the most independent **H**.
  - (**B** is found at the same time, since **X=HB**.)

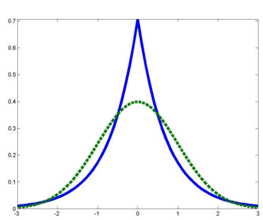
ICA

- Define "Independence": $p(h_i, h_j) = p(h_i)p(h_j)$
  - Zero mutual information
  - Non-Gaussianity, max. absolute Kurtosis
- To solve for H,B:
  - Neural networks, optimization methods (gradient ascent, fixed-point, …)

An non-gaussian distribution: Laplacian pdf

$$P(x) = \frac{\lambda}{2}\exp(-\lambda|x|)$$

Sharper at 0,
and more heavy tail
than Gaussian pdf



CMU SCS

Maximizing non-gaussianity

- Assume h_i ~ non-gaussian pdf (e.g. Laplacian pdf)
  - Fixed h_i values, what is the most likely "**B**" ?
    - (data point **x** is given and fixed)
  - Find **B**, s.t. likelihood $P(\mathbf{x}|\mathbf{B})$ is maximized.

$$\vec{\mathbf{x}} = \vec{\mathbf{h}}\mathbf{B}$$

$$\Rightarrow P(\vec{\mathbf{x}}|\mathbf{B}) = \frac{P(\vec{\mathbf{h}})}{\det(\mathbf{B})}$$

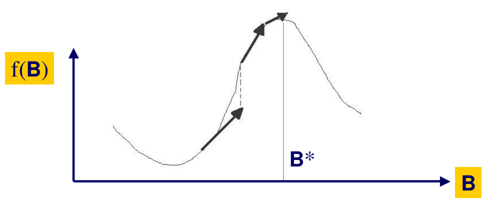
$$P(\vec{\mathbf{h}}) = P_{Laplacian}(\vec{\mathbf{h}}) = P_{Laplacian}(\vec{\mathbf{x}}\mathbf{B}^{-1})$$

$$\Rightarrow P(\vec{\mathbf{x}} \mid \mathbf{B}) = \frac{P_{Laplacian}(\vec{\mathbf{x}}\mathbf{B}^{-1})}{\det(\mathbf{B})}$$

CMU SCS

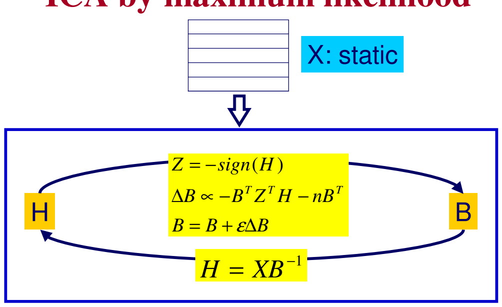
Maximize likelihood

- Likelihood $P(\mathbf{x}|\mathbf{B})$ is a function of **B**, f(**B**)
- Gradient ascent
  - To find **B** which maximizes $P(\mathbf{x}|\mathbf{B})$

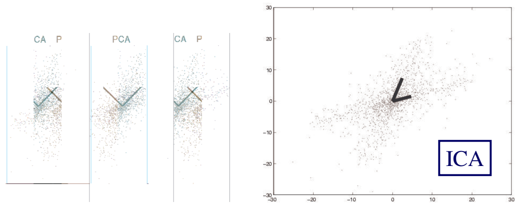


CMU SCS

ICA by maximum likelihood



Convergence of ICA



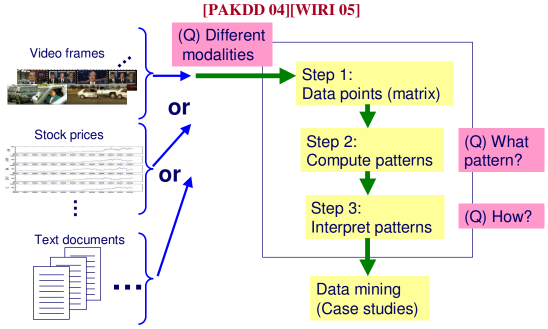
Start with the matrix B as an identity matrix.

Outline

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

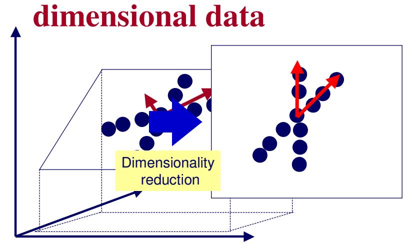
Pattern discovery with ICA: AutoSplit

[PAKDD 04][WIRI 05]



CMU SCS

Finding patterns in high-dimensional data



PCA finds the hyperplane. ICA finds the correct patterns.

CMU SCS

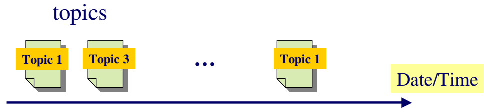
Outline

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

CMU SCS

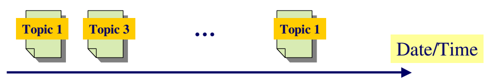
Topic discovery on text streams

- 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

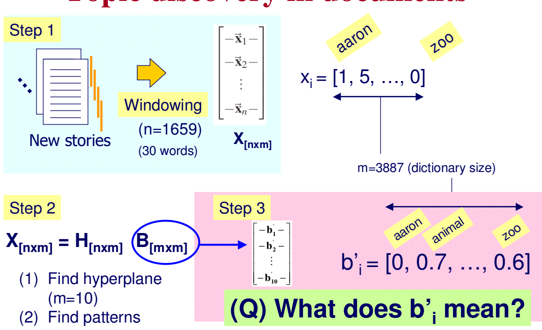


Topic discovery on text streams

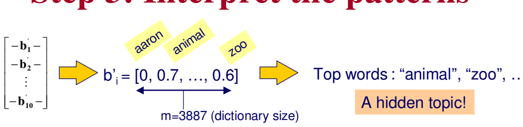
- Known: number of topics = 10
- Unknown: (1) topic of each document (2) topic description



Topic discovery in documents



Step 3: Interpret the patterns



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

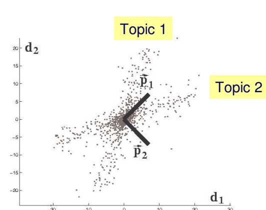
AutoSplit's topics are better than PCA.

CMU SCS

Step 3: Evaluate the patterns

	AutoSplit				
A					
B					
C					
D					
E					

	PCA				
A'					
B'					
C'					
D'					
E'					



PCA vectors mix the topics.

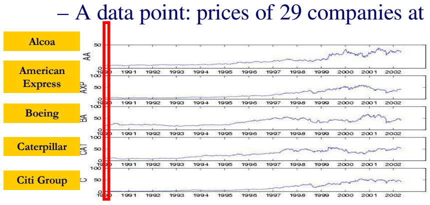
AutoSplit's topics are better than PCA.

Outline

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

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

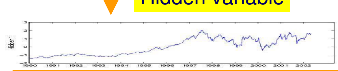


Formulation: Find hidden variables

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

Date

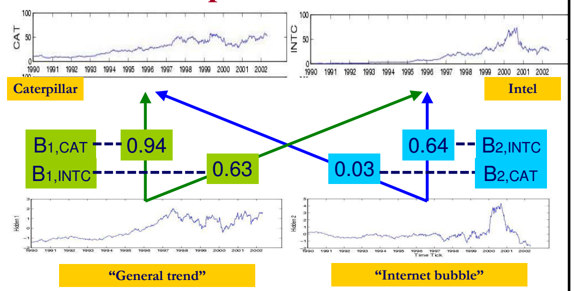
Hidden variable



Date

CMU SCS

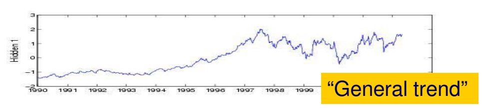
Characterize hidden variable by the companies it influences



CMU SCS

Companies related to hidden variable 1

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



CMU SCS

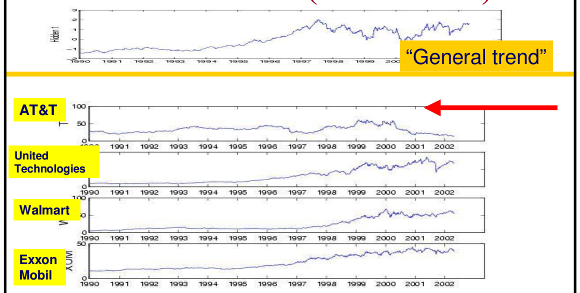
Companies related to hidden variable 1

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.

CMU SCS

General trend (and outlier)

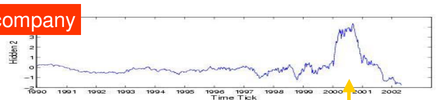


CMU SCS

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"

CMU SCS

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).

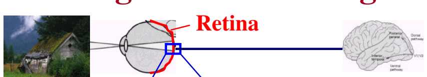
CMU SCS

Outline

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

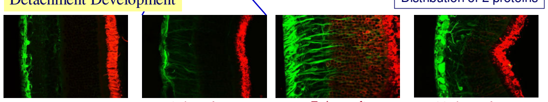
CMU SCS

Mining cat retinal images [ICDM 05]



Detachment Development

Distribution of 2 proteins



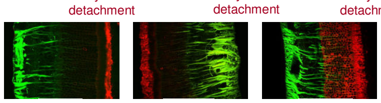
Normal

1 day after detachment

7 days after detachment

28 days after detachment

Treatment



1h3dr

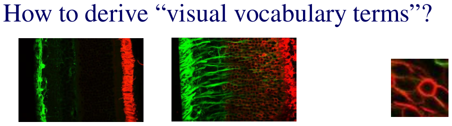
3d28dr

1d6dO_2

CMU SCS

"Vocabulary" for biomedical images?

- How to describe biomedical images?
- Analogy: the topics for text
  - Football reports: "touchdown", "punt", etc.
  - DB papers: "query", "optimization", etc.
- How to derive "visual vocabulary terms"?

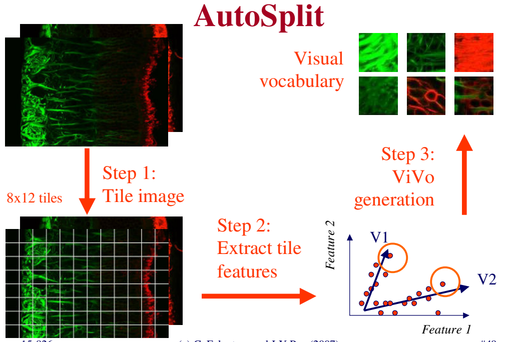


Normal

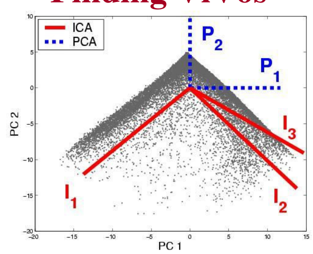
7 days after detachment

"spongy"

Visual Vocabulary (ViVo) by AutoSplit



Finding ViVos



Red lines indicate ViVos.

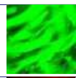
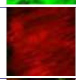
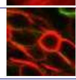
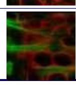
Each point is a tile.
Projected to the 1st and 2nd PCA vectors.
(Feature vector: 512 color structure features.)

Bio-mining with "ViVo"

- ✔ Visual Vocabulary for retinal images
  – using AutoSplit
- ➡ Evaluation
  – Qualitative: biological meanings of ViVos
  – Data mining: highlight "interesting" regions

CMU SCS

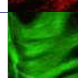
Biological interpretation of ViVos

ID	ViVo	Description	Condition
V1		GFAP in inner retina (Müller cells)	Healthy
V10		Healthy outer segments of rod photoreceptors	Healthy
V8		Redistribution of rod opsin into cell bodies of rod photoreceptors	Detached
V11		Co-occurring processes: Müller cell hypertrophy and rod opsin redistribution	Detached

CMU SCS

Biological interpretation of ViVos

ID	ViVo	Description	Condition
2		GFAP in hypertrophy Müller cells	Morphological changes in inner retina
3		GFAP in hypertrophy Müller cells	Morphological changes in inner retina
4		GFAP in hypertrophy Müller cells	Morphological changes in inner retina
5		Healthy outer segments of rod photoreceptors (rod opsin)	Healthy

ID	ViVo	Description	Condition
6		Rod photoreceptor cell body	Background labeling
7		GFAP in hypertrophy Müller cells	Morphological changes in inner retina
9		Outer segment degeneration (rod opsin)	Detached
12		GFAP in hypertrophy Müller cells	Morphological changes in inner retina

CMU SCS

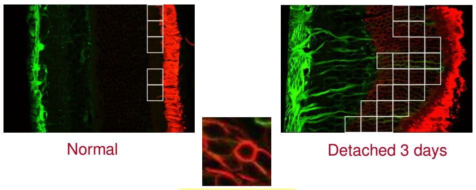
Bio-mining with "ViVo"

- Visual Vocabulary for retinal images
  - using AutoSplit
- Evaluation
  - Qualitative: biological meanings of ViVos
  - Data mining: highlight "interesting" regions

CMU SCS

Finding "distinguishing ViVos"

- Given: Images of two classes
  - Find the class-distinguishing ViVo ("DiVo")
  - Highlight distinguishing regions



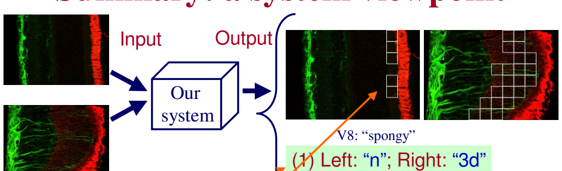
Normal

Detached 3 days

DiVo: "spongy"

CMU SCS

Summary: a system viewpoint



(1) Left: "n"; Right: "3d"

Accurate classification

(2) Regions shown (V8): "cells of rod photoreceptors"

DiVo analysis

(3) Description: "Detachment occurs!"

ViVo interpretation

"Rod opsin distributes from outer segment into cell bodies."

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

CMU SCS

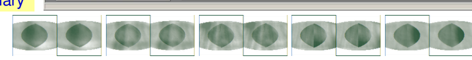
Conclusion

- 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

CMU SCS

Vocabulary for embryo gene expressions

Vocabulary



with André Balan, Christos Faloutsos, Eric P. Xing

CMU SCS

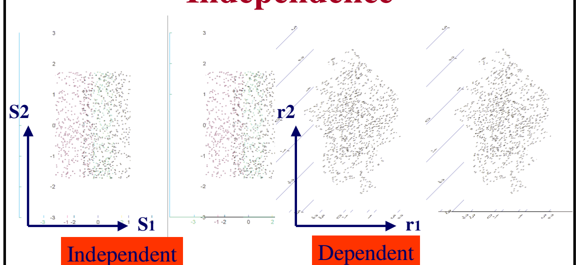
References

- 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.

Acknowledgement

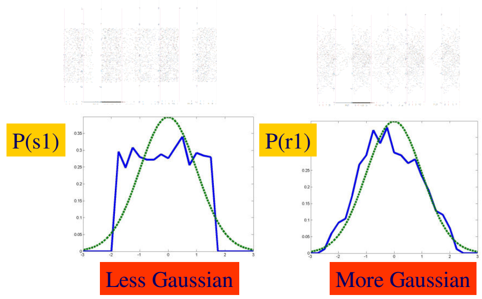
- Prof. Tai Sing Lee
- Prof. Hiroyuki Kitagawa
- Prof. HyungJeong Yang
- Masafumi Hamamoto
- Prof. Nancy Pollard
- Prof. Jessica Hodgins
- Prof. Michael Lewicki
- Prof. Eric Xing
- CMU Informedia project
- UCSB DB Lab
- CMU bio-imaging center
- CMU graphics lab

Independence



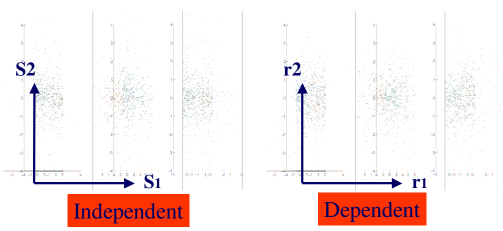
CMU SCS

Non-Gaussianity and Independence



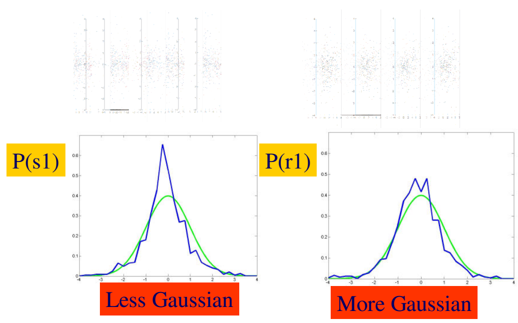
CMU SCS

Non-Gaussianity and Independence



CMU SCS

Non-Gaussianity and Independence



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

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