15-826: Multimedia Databases and Data Mining

Addendum to Lecture #21: Independent Component Analysis (ICA)
Jia-Yu Pan and Christos Faloutsos

Must-read Material

PAKDD 2004, Sydney, Australia

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?

Motivation:
(Q2) Find hidden variables

Stock prices

Hidden variables (="topics" = concepts)

“General trend”

“Internet bubble”
Motivation:
(Q2) Find hidden variables

Motivation:
(Q2) Find hidden variables

Motivation:
Find hidden variables

• There are two sound sources in a cocktail party…
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):
  \[ \tilde{x}_i = h_{i,1}\tilde{b}_1 + h_{i,2}\tilde{b}_2 \]
Patterns as data “vocabulary”

(Q) Given data $x_i$’s, compute $h_i$’s and $b_i$’s that are "sparse"?

Only $b_1$ is needed to describe $x_i$.

Patterns in motion capture data

$n=550$ ticks

$X = (U \Lambda V^T)$

Sparse ~ non-Gaussian ~ "Independent"

"Independent": e.g., minimize mutual information.
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 85]

Step 1: Data points (matrix)
Step 2: Compute patterns
Step 3: Interpret patterns

Data mining (Case studies)

(Q) What pattern?
(Q) Different modalities
(Q) How?

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

---

---
Topic discovery on text streams

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

Date/Time

Step 1
Windowing
(n=1659)
(30 words)

Step 2
\[ x_i = [1, 5, \ldots, 0] \]
m=3887 (dictionary size)

Step 3
(1) Find hyperplane (m'=10)
(2) Find patterns

b'_i = [0, 0.7, \ldots, 0.6]

(Q) What does \( b'_i \) mean?

Step 3: Interpret the patterns

Top words: “animal”, “zoo”, …

General idea: related to the data attributes

<table>
<thead>
<tr>
<th>BD</th>
<th>Mckinnon</th>
<th>Sergeant</th>
<th>sexual</th>
<th>Major</th>
<th>Army</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Smith</td>
<td>Rodeph</td>
<td>Chori</td>
<td>Atlanta</td>
<td>Benny</td>
</tr>
<tr>
<td>C</td>
<td>Wateii</td>
<td>Booth</td>
<td>Texas</td>
<td>Oprah</td>
<td>Card</td>
</tr>
<tr>
<td>D</td>
<td>Yoga</td>
<td>Deng</td>
<td>Inpart</td>
<td>PHI</td>
<td>Ducer</td>
</tr>
<tr>
<td>E</td>
<td>Zemore</td>
<td>Vicklows</td>
<td>KILL</td>
<td>Former</td>
<td>Idea</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F</th>
<th>Asia</th>
<th>Economic</th>
<th>Japan</th>
<th>Ecomon</th>
<th>Asian</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>Super</td>
<td>Hard</td>
<td>Game</td>
<td>Train</td>
<td>No</td>
</tr>
<tr>
<td>J</td>
<td>People</td>
<td>Turado</td>
<td>Florid</td>
<td>Re</td>
<td>North</td>
</tr>
</tbody>
</table>
Step 3: Evaluate the patterns

<table>
<thead>
<tr>
<th>ID</th>
<th>True Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sgt. Gene McKinney is on trial for alleged sexual misconduct</td>
</tr>
<tr>
<td>2</td>
<td>A bomb explodes in a Birmingham, AL, abortion clinic</td>
</tr>
<tr>
<td>3</td>
<td>The cattle industry in Texas sues Oprah Winfrey for defaming beef</td>
</tr>
<tr>
<td>4</td>
<td>New impotency drug Viagra is approved for use</td>
</tr>
<tr>
<td>5</td>
<td>Diane Zamora is convicted of helping to murder her lover’s girlfriend</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Sorted word list</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>mckinne sergeant sexual major armi</td>
</tr>
<tr>
<td>B</td>
<td>bomb rudolph clinic atlanta birmingham</td>
</tr>
<tr>
<td>C</td>
<td>winfrey beef texa oprah cath</td>
</tr>
<tr>
<td>D</td>
<td>viagra drug impot pill doctor</td>
</tr>
<tr>
<td>E</td>
<td>zamora graham kill former jone</td>
</tr>
</tbody>
</table>

AutoSplit finds correct topics.

Step 3: Evaluate the patterns

<table>
<thead>
<tr>
<th>ID</th>
<th>AutoSplit</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>mckinne sergeant sexual major armi</td>
</tr>
<tr>
<td>B</td>
<td>bomb rudolph clinic atlanta birmingham</td>
</tr>
<tr>
<td>C</td>
<td>winfrey beef texa oprah cath</td>
</tr>
<tr>
<td>D</td>
<td>viagra drug impot pill doctor</td>
</tr>
<tr>
<td>E</td>
<td>zamora graham kill former jone</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td>mckinne bomb women sexual sergeant</td>
</tr>
<tr>
<td>B'</td>
<td>bomb mckinne rudolph clinic atlanta</td>
</tr>
<tr>
<td>C'</td>
<td>winfrey viagra texa beef oprah</td>
</tr>
<tr>
<td>D'</td>
<td>viagra winfrey drug texa beef</td>
</tr>
<tr>
<td>E'</td>
<td>zamora viagra winfrey graham olymp</td>
</tr>
</tbody>
</table>

AutoSplit’s topics are better than PCA.

AutoSplit’s topics are better than PCA.

PCA vectors mix the topics.
Outline

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

\[
\begin{bmatrix}
AA_1, \ldots, XOM_1 \\
\vdots \\
AA_n, \ldots, XOM_n
\end{bmatrix}
= \begin{bmatrix}
H_{11}, H_{12}, \ldots, H_{1m} \\
\vdots \\
H_{n1}, H_{n2}, \ldots, H_{nm}
\end{bmatrix}
\begin{bmatrix}
B_{11}, B_{12}, \ldots, B_{1m} \\
\vdots \\
B_{m1}, B_{m2}, \ldots, B_{mm}
\end{bmatrix}
\]
Characterize hidden variable by the companies it influences

<table>
<thead>
<tr>
<th></th>
<th>Highest</th>
<th>Lowest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caterpillar</td>
<td>0.938512</td>
<td>AT&amp;T</td>
</tr>
<tr>
<td>Boeing</td>
<td>0.911120</td>
<td>WalMart</td>
</tr>
<tr>
<td>MMM</td>
<td>0.906542</td>
<td>Intel</td>
</tr>
<tr>
<td>Coca Cola</td>
<td>0.903858</td>
<td>Home Depot</td>
</tr>
<tr>
<td>Du Pont</td>
<td>0.900317</td>
<td>Hewlett-Packard</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>$B_{2,j}$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>0.641102</td>
<td>Philip Morris</td>
<td>-0.194843</td>
</tr>
<tr>
<td>HP</td>
<td>0.621159</td>
<td>International Paper</td>
<td>-0.089569</td>
</tr>
<tr>
<td>GE</td>
<td>0.509164</td>
<td>Caterpillar</td>
<td>0.031678</td>
</tr>
<tr>
<td>AMEX</td>
<td>0.504871</td>
<td>Procter and Gamble</td>
<td>0.195576</td>
</tr>
<tr>
<td>Disney</td>
<td>0.490529</td>
<td>Du Pont</td>
<td>0.133337</td>
</tr>
</tbody>
</table>

Tech company

2000-2001 “Internet bubble”

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

• 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


• 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