**15-826: Multimedia Databases and Data Mining**

Lecture #22:  
*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

---

**Must-read Material**

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

Hidden variables (="topics" = concepts)

"General trend"

"Internet bubble"
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):

\[
\bar{x}_i = h_{i,1} \bar{b}_1 + h_{i,2} \bar{b}_2
\]
Patterns as data “vocabulary”

Good pattern \( \approx \) sparse coding

(Q) Given data \( x_i \)'s, compute \( h_{ij} \)'s and \( b_i \)'s that are “sparse”?

Only \( b_i \) is needed to describe \( x_i \).

Patterns in motion capture data

\[
X_{m=2} = H_{m=2} B_{2 \times 2}
\]

Data matrix

Hidden variables

Basis vectors

n=550 ticks

"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 05]

Step 1:
- Data points (matrix)

Step 2:
- Compute patterns
- (Q) What pattern?

Step 3:
- Interpret patterns
- (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

**Step 1:** Windowing

**Step 2:**

- \( x_{[nxm]} \) = \[1, 5, \ldots, 0\]

**Step 3:**

- Hyperplane: \( b'_{i} = [0, 0.7, \ldots, 0.6] \)
- Patterns: "animal", "zoo", ...

**Step 3: Interpret the patterns**

- \( aaron \), \( zoo \), \( animal \)
- \( m=3887 \) (dictionary size)

**General idea:** related to the data attributes
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>winfrei beef texa oprah cattl</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.

AutoSplit’s topics are better than PCA.

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

\[
\begin{align*}
&\text{AA}_1, \ldots, \text{XOM}_1 \\
&\ldots \\
&\ldots \\
&\text{AA}_n, \ldots, \text{XOM}_n \\
&= B_{11}, B_{12}, \ldots, B_{1m} \\
&\ldots \\
&\ldots \\
&B_{m1}, B_{m2}, \ldots, B_{mm}
\end{align*}
\]

Characterize hidden variable by the companies it influences

<table>
<thead>
<tr>
<th>Company</th>
<th>B_{1,CAT}</th>
<th>B_{1,INTC}</th>
<th>B_{2,INTC}</th>
<th>B_{2,CAT}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caterpillar</td>
<td>0.94</td>
<td>0.03</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>Intel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Companies related to hidden variable 1

<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 0.021885</td>
</tr>
<tr>
<td>Boeing</td>
<td>0.911120</td>
<td>WalMart 0.624570</td>
</tr>
<tr>
<td>MMM</td>
<td>0.906542</td>
<td>Intel 0.638010</td>
</tr>
<tr>
<td>Coca Cola</td>
<td>0.903858</td>
<td>Home Depot 0.647774</td>
</tr>
<tr>
<td>Du Pont</td>
<td>0.900317</td>
<td>Hewlett-Packard 0.658768</td>
</tr>
</tbody>
</table>
Companies related to hidden variable 1

<table>
<thead>
<tr>
<th>B₁,j</th>
<th>Highest</th>
<th>Lowest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caterpillar</td>
<td>0.938512</td>
<td>AT&amp;T 0.021885</td>
</tr>
<tr>
<td>Boeing</td>
<td>0.911120</td>
<td>WalMart 0.624570</td>
</tr>
<tr>
<td>MMM</td>
<td>0.906542</td>
<td>Intel 0.638010</td>
</tr>
<tr>
<td>Coca Cola</td>
<td>0.903858</td>
<td>Home Depot 0.647774</td>
</tr>
<tr>
<td>Du Pont</td>
<td>0.900317</td>
<td>Hewlett-Packard 0.658768</td>
</tr>
</tbody>
</table>

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

2000-2001 “Internet bubble”

Companies related to hidden variable 2

<table>
<thead>
<tr>
<th>B₂,j</th>
<th>Highest</th>
<th>Lowest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>0.641102</td>
<td>Philip Morris -0.194843</td>
</tr>
<tr>
<td>Hewlett-Packard</td>
<td>0.621159</td>
<td>International Paper -0.089569</td>
</tr>
<tr>
<td>GE</td>
<td>0.509164</td>
<td>Caterpillar 0.031678</td>
</tr>
<tr>
<td>American Express</td>
<td>0.504871</td>
<td>Procter and Gamble 0.109576</td>
</tr>
<tr>
<td>Disney</td>
<td>0.490529</td>
<td>Du Pont 0.133337</td>
</tr>
</tbody>
</table>

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

• 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
• Rule of thumb: plot after PCA;
  – if ‘chicken-feet’, try ICA

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

References


Software

• Open source software: ‘fastICA’
  http://research.ics.tkk.fi/ica/fastica/

• Or ‘autosplit’:
  www.cs.cmu.edu/~jypan/software/autosplit_cmu.tar.gz