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

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

Motivation:
(Q1) Find patterns in data
• Motion capture data: broad jumps

Motion capture data: broad jumps
Energy exerted
Left Knee
Right Knee
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

Hidden variables (="topics" = concepts)

- "General trend"
- "Internet bubble"

Motivation:
(Q2) Find hidden variables

Hidden variable 1
Hidden variable 2

"General trend"  Hidden variables  "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):

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

Good pattern \( \approx \) sparse coding

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

Only \( b_1 \) is needed to describe \( x_c \).

Patterns in motion capture data

\[
X_{m2} = H_{m2}B_{2x2}
\]

\( n=550 \) ticks

Data matrix

Hidden variables

Basis vectors

“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
- Step 3: Interpret patterns

(Q) Different modalities

(Q) What pattern?

(Q) How?

Data mining (Case studies)

Finding patterns in high-dimensional data

PCA finds the hyperplane. ICA finds the correct patterns.

Dimensionality reduction

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

Date/Time
**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 in documents**

- **Step 1**
  - Windowing
  - New stories
  - (n=1659) (30 words)

- **Step 2**
  - $X_{[nxm]} = H_{[nxm]} \cdot B_{[m'x]}$
    - (1) Find hyperplane $m'=10$
    - (2) Find patterns

- **Step 3**
  - $x_i = [1, 5, ..., 0]$
  - $b' = [0, 0.7, ..., 0.6]$
  - (Q) What does $b'$ mean?

---

**Step 3: Interpret the patterns**

- Top words: “animal”, “zoo”, ...
- A hidden topic!

---

**Topics found**

- **ID**
  - A: Mckinne
  - B: bomb
  - C: Winfree
  - D: Viagra
  - E: Zamora

- **Sorted word list**
  - A: Sergeant, sexual, Major, Arm
  - B: Bomb, Rudolph, Clinic, Atlanta, Birmingham
  - C: Beef, Tesa, Oprah, Catt
  - D: Drug, Impot, Pill, Doctor
  - E: Graham, Kill, Yormur, Jon

---

**General idea: related to the data attributes**

- H: Asia, Econom, Japan, Econom, Asian
- J: Super, Bowl, Game, Team, Re
- J: Peopl, Tornado, Florida, Re, Bomb
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.

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>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’s topics are better than PCA.

Step 3: Evaluate the patterns

<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>winfrei viagra texa beef oprah</td>
</tr>
<tr>
<td>D</td>
<td>viagra winfrei drug texa beef</td>
</tr>
<tr>
<td>E</td>
<td>zamora viagra winfrei graham olymp</td>
</tr>
</tbody>
</table>

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

\[
\begin{align*}
&\text{AA}_1, \ldots, \text{XOM}_1 \\
&\ldots \\
&\ldots \\
&\text{AA}_n, \ldots, \text{XOM}_n \\
&= \begin{bmatrix}
B_{11}, B_{12}, \ldots, B_{1m} \\
\vdots \\
\vdots \\
B_{m1}, B_{m2}, \ldots, B_{mm}
\end{bmatrix} \\
&H_{11}, H_{12}, \ldots, H_{1m} \\
&\quad \ldots \\
&\quad \ldots \\
&H_{n1}, H_{n2}, \ldots, H_{nm}
\end{align*}
\]

Characterize hidden variable by the companies it influences

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>B_{1,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.

Companies related to hidden variable 2

<table>
<thead>
<tr>
<th>B_{2,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>0.590164</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>

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

Conclusion

References


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


Software

- Open source software: ‘fastICA’

- Or ‘autosplit’:
  [www.cs.cmu.edu/~jypan/software/autosplit_cmu.tar.gz](http://www.cs.cmu.edu/~jypan/software/autosplit_cmu.tar.gz)