Problem: BSS

- two sound sources in a cocktail party – separate them

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

Q: how to extract \textit{sparse} hidden/latent variables?

Answer

Q: how to extract \textit{sparse} hidden/latent variables?
A: \textit{SVD ICA PCA}
Must-read Material


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

Hidden variables (= ‘topics’ = concepts)

Stock prices

Alcoa
American Express
Boeing
Citi Group

“General trend”

Trend#2

(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

• 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):

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

Good pattern ≈ sparse coding

b \_i alone, can describe x \_i.

(a) ICA representation of \( \tilde{x}_i \)

\[
\tilde{x}_i = h_{i,1} \tilde{b}_1 + h_{i,2} \tilde{b}_2
\]

PCA: first step of ICA

PCA finds the hyperplane. ICA finds the correct patterns.

Dimensionality reduction
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

Outline

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

Motivation:
Find hidden variables

Dow Jones Industrial Average

Alcoa

American Express

Boeing

Caterpillar

Citi Group

Find common hidden variables, and weights.
ICA: Like SVD, but sparse U

 Participation weight of row \( i \) to behavior \( j \)

Motivation:
Find hidden variables
Motivation:
Find hidden variables

"Hidden variable 1"
"Hidden variable 2"

"General trend"
???

Caterpillar
Intel

Motivation:
Find hidden variables

"Hidden variable 1"
"Hidden variable 2"

"General trend"
???

Caterpillar
Intel

Copyright (c) 2019 C. Faloutsos
Motivation: Find hidden variables

ICA: Like SVD, but sparse
ICA: Like SVD, but sparse

1st behavior
2nd behavior

Stock#1
Stock#2

General trend
Internet bubble
ICA: Like SVD, but sparse

Stock#1
General trend
Stock#2
Internet bubble

What else can ICA tell us?
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>

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

15-826 Copyright (c) 2019 C. Faloutsos
Companies related to hidden variable 2

<table>
<thead>
<tr>
<th>Company</th>
<th>Highest</th>
<th>Lowest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>0.641102</td>
<td>Philip Morris</td>
</tr>
<tr>
<td>Hewlett-Packard</td>
<td>0.621159</td>
<td>International Paper</td>
</tr>
<tr>
<td>GE</td>
<td>0.509164</td>
<td>Caterpillar</td>
</tr>
<tr>
<td>American Express</td>
<td>0.504871</td>
<td>Procter and Gamble</td>
</tr>
<tr>
<td>Disney</td>
<td>0.490529</td>
<td>Du Pont</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
  – 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?

- A: Sliding windows
**Step 1: Windowing**

- New stories (n=1659) (30 words) $X_{[nxm]}$
- Windowing
- $x_i = [1, 5, \ldots, 0]$
- $m=3887$ (dictionary size)

**Step 2: Interpret the patterns**

- $b'_i = [0, 0.7, \ldots, 0.6]$
- Top words: “animal”, “zoo”, …
- A hidden topic!

**Topics found**

<table>
<thead>
<tr>
<th>ID</th>
<th>Sorted word list</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Mckinne Sergeant sexual Major Arma</td>
</tr>
<tr>
<td>B</td>
<td>bomb Rudolph Clinic Atlanta Birmingham</td>
</tr>
<tr>
<td>C</td>
<td>Winfrey Beef Texa Oprah Cartl</td>
</tr>
<tr>
<td>D</td>
<td>Viagra Drug Impot Pill Doctor</td>
</tr>
<tr>
<td>E</td>
<td>Zamora Graham Kill Former John</td>
</tr>
<tr>
<td>F</td>
<td>Medal Olymp Gold Women Game</td>
</tr>
<tr>
<td>G</td>
<td>Pope Cube Castro Cuban Visit</td>
</tr>
<tr>
<td>H</td>
<td>Asia Econom Japan Econom Asian</td>
</tr>
<tr>
<td>I</td>
<td>Super Bowl Game Team Re</td>
</tr>
<tr>
<td>J</td>
<td>Peopl Tornado Florida Re bomb</td>
</tr>
</tbody>
</table>
### Step 3: Evaluate the patterns

#### True Topic

<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>

#### Sorted word list

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

AutoSplit finds correct topics.

---

### Step 3: Evaluate the patterns

#### AutoSplit

<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>

#### PCA

<table>
<thead>
<tr>
<th>ID</th>
<th>AutoSplit</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.
Step 3: Evaluate the patterns

<table>
<thead>
<tr>
<th></th>
<th>AutoSplit</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td></td>
</tr>
<tr>
<td>B'</td>
<td></td>
</tr>
<tr>
<td>C'</td>
<td></td>
</tr>
<tr>
<td>D'</td>
<td></td>
</tr>
<tr>
<td>E'</td>
<td></td>
</tr>
</tbody>
</table>

AutoSplit’s topics are better than PCA.

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, PCA

**Citation**


PAKDD 2004, Sydney, Australia
References


References

Software

• Open source software: ‘fastICA’
  [link](http://research.ics.aalto.fi/ica/fastica/)

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