Unusual Tensor Decompositions for Informatics Applications

Brett W. Bader
Sandia National Laboratories

NSF Tensor Workshop
February 20, 2009
Acknowledgements

- Richard Harshman (Univ. Western Ontario)
- Peter Chew (Sandia)
- Tammy Kolda (Sandia)
- Ahmed Abdelali (NMSU)
Tensor Decompositions

...and many more!

Each provides a different interpretation of the data
Temporal Analysis of Enron email using 3-way DEDICOM
Three-way DEDICOM

\[ X_x = AD_k RD_k A^T \quad k = 1, \ldots, K \]

- Introduced by Harshman (1978)
- DEcomposition into Dlrectional COMponents
- Columns of \( A \) are not necessarily orthogonal
- Central matrix \( R \) contains asymmetric information from \( X \)
- *Unique* solution with enough slices of \( X \) with sufficient variation
  - i.e., no rotation of \( A \) possible
  - greater confidence in interpretation of results
- Alternating algorithms; least-squares and approximations
- Early applications:
  - World trade (import/export matrices)
  - Car switching
- Variations: constrained DEDICOM
Application: Enron Email Analysis

- Links consist of email communications
- What can we learn about this network strictly from their communication patterns? (Social network analysis)
Case Study: Enron

- Enron created energy markets
- EnronOnline: e-trading business
  - natural gas
  - electric power
- Investigations
  - FERC
    - energy market manipulation
    - involved energy traders
  - SEC
    - accounting fraud
    - insider trading

Email communications at Enron (1998-2002)
Temporal Social Network Analysis

Email communications at Enron (1998-2002)
(data released by U.S. Federal Energy Regulatory Commission)

Emails among 184 employees over 44 months

Time series of communication graphs among employees

Adjacency array

DEDICOM

Joint work with R. Harshman (UWO) and T. Kolda
Roles of Employees

Bi-plots of two roles

L. Kitchen - President, Enron Online

J. Dasovich – Employee, Government Relationship Executive
J. Steffes – VP, Government Affairs
R. Shapiro – VP, Regulatory Affairs
S. Kean – VP, Chief of Staff
R. Sanders – VP, Enron Wholesale Services
S. Shackleton – ENA Legal
T. Jones – Financial Trading Group
M. Taylor – Manager, Financial Trading Group
ENA Legal

K. Watson
Transwestern Pipeline Company (ETS)

M. Lokay
Admin. Asst.
Transwestern Pipeline Company (ETS)

L. Donoho – Employee, Transwestern Pipeline Company (ETS)

M. McConnell – Employee, Transwestern Pipeline Company (ETS)
L. Blair – Employee, Northern Natural Gas Pipeline (ETS)

M. McConnell – Employee, Transwestern Pipeline Company (ETS)
L. Blair – Employee, Northern Natural Gas Pipeline (ETS)

L. Kitchen – President, Enron Online

J. Lavorato
CEO, Enron America

Identify shared characteristics to label group

Soft clustering

Bi-plots of two roles
Communication Patterns

- Mostly communication within roles
- Some asymmetric exchanges
Temporal Patterns

Communication patterns over time

- Legal
- Government & regulatory affairs
- Trade executives
- Pipeline employee

Normalized scale

Month

Enron crisis breaks; investigation begins
Filed for bankruptcy
Multilingual Text Analysis using PARAFAC2
PARAFAC2

\[ X_k \approx A_k C_k B^T \]

- Introduced by Harshman (1972)
- Less constrained than PARAFAC
- Related to 3-way DEDICOM
- Slices of \( A \) are constrained but not necessarily orthogonal
- *Unique* solution with enough slices of \( X \) with sufficient variation
  - i.e., no rotation of \( A \) possible
  - greater confidence in interpretation of results
- Alternating algorithms: least-squares and approximations
- Early applications:
  - Sets of cross-product matrices
  - Chromatographic data with retention time shifts
Cross-language Information
Retrieval (CLIR)

Web documents could be in any language

Languages on the web

Goal: Cluster documents by topic regardless of language
### Bible as Parallel Corpus

Linguistic differences among translations

<table>
<thead>
<tr>
<th>Translation</th>
<th>Terms</th>
<th>Total Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (King James)</td>
<td>12,335</td>
<td>789,744</td>
</tr>
<tr>
<td>Spanish (Reina Valera 1909)</td>
<td>28,456</td>
<td>704,004</td>
</tr>
<tr>
<td>Russian (Synodal 1876)</td>
<td>47,226</td>
<td>560,524</td>
</tr>
<tr>
<td>Arabic (Smith Van Dyke)</td>
<td>55,300</td>
<td>440,435</td>
</tr>
<tr>
<td>French (Darby)</td>
<td>20,428</td>
<td>812,947</td>
</tr>
</tbody>
</table>

- Languages convey information in different number of words
  - Isolating language: One morpheme per word
    - e.g., "He travelled by hovercraft on the sea." Largely isolating, but travelled and hovercraft each have two morphemes per word.
  - Synthetic language: High morpheme-per-word ratio
    - e.g., Aufsichtsratsmitgliederversammlung => "On-view-council-with-limbs-gathering" meaning "meeting of members of the supervisory board".
Term-Doc Matrix

Term-by-verse matrix for all languages

Look for co-occurrence of terms in the same verses and across languages to capture latent concepts.
Latent Semantic Indexing

Term-by-verse matrix for all languages

Bible verses
English
Spanish
Russian
Arabic
French

Truncated SVD

A_k = U_k \sum_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T

Projection

Document feature vector

dimension 1 0.1375
dimension 2 0.1052
dimension 3 0.0341
dimension 4 0.0441
dimension 5 -0.0087
dimension 6 0.0410
dimension 7 0.1011
dimension 8 0.0020
dimension 9 0.0518
dimension 10 0.0822
dimension 11 -0.0101
dimension 12 -0.1154
dimension 13 -0.0990
dimension 14 0.0228
dimension 15 -0.0520
dimension 16 0.1096
dimension 17 0.0294
dimension 18 0.0495
dimension 19 0.0553
dimension 20 0.1598

term x concept

Project new documents of interest into subspace of $U\Sigma^{-1}$ and compute cosine similarities
Quran as Test Set

- Quran is translated into many languages, just like the Bible
- 114 suras (or chapters)
- More variation across translations = harder clustering task
Performance Metrics

- **MP5**: Average multilingual precision at 5 (or n) documents
  - The average percentage of the top 5 documents that are translations of the query document
  - Calculated as an average for all languages
  - Essentially, MP5 measures success in multilingual clustering
## LSA Results

5 languages, 240 latent dimensions

<table>
<thead>
<tr>
<th>Method</th>
<th>Average MP5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD/LSA</td>
<td>65.5%</td>
</tr>
</tbody>
</table>

Documents tend to cluster more by language than by topic
New Approach: Multi-matrix Array

(Chew, Bader, Kolda, Abdelali, 2007)

Term-by-verse matrix for each language

Array size: 55,300 x 31230 x 5 with 2,765,719 nonzeros
$Tucker_1$

$X_1 \approx Tucker_1 \approx U_1$
# Tucker1 Results

5 languages, 240 latent dimensions

<table>
<thead>
<tr>
<th>Method</th>
<th>Average MP5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD/LSA</td>
<td>65.5%</td>
</tr>
<tr>
<td>Tucker1</td>
<td>71.3%</td>
</tr>
</tbody>
</table>

Only minor improvement because each $U_k$ is not orthogonal.
PARAFAC2

Where each $U_k$ is orthonormal and $S_k$ is diagonal

$$X_k \approx U_k H S_k V^T$$

(Harshman, 1972)
PARAFAC2 Results

5 languages, 240 latent dimensions

<table>
<thead>
<tr>
<th>Method</th>
<th>Average MP5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD/LSA</td>
<td>65.5%</td>
</tr>
<tr>
<td>Tucker1</td>
<td>71.3%</td>
</tr>
<tr>
<td>PARAFAC2</td>
<td>78.5%</td>
</tr>
</tbody>
</table>

Modest improvement over LSA

Why PARAFAC2?
Tensor Methods and Modeling: Why the Proliferation?

- N-way interactions in real world applications
- Next frontier after matrix linear algebra
  - Lots of low hanging fruit
- New mathematical and computational challenges
  - Differences from matrix problems (e.g., rank of 2x2x2)
  - Original algorithms developed in different research communities
Thoughts on Future Directions of Tensor-based Computation and Modeling

- Need scalable algorithms
  - Fast, efficient for large-scale problems
  - Handle constraints
    - non-negativity
    - sparsity
    - orthogonality
    - etc.
  - Parallel algorithms

- Match models to applications
  - Requires creativity by domain experts and tensor researchers
  - Sometimes not a straightforward extension from matrix approaches
  - Danger of reinventing what’s already in the literature
    - psychometrics
Questions?

bwbader@sandia.gov

http://www.sandia.gov/~bwbader/