Using Nonnegative Matrix and Tensor Factorizations for Topic and Scenario Detection and Tracking
Math and Stat Colloquium, Utah State University

Michael W. Berry
Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville
February 19, 2009

Collaborators

- Brett Bader (Sandia National Labs)
- Murray Browne (Turner Broadcasting System, Inc.)
- Paul Pauca, Bob Plemmons, Brian Lamb (Wake Forest)
- Amy Langville (College of Charleston)
- David Skillicorn, Parambir Keila (Queens U.)
- Stratis Gallopoulos, Ioannis Antonellis (U. Patras)

NNMF Origins

1. Nonnegative Matrix Factorization (NNMF)
2. Document Parsing and Term Weighting - ASRS
3. NNMF Classification of ASRS Documents
4. NNTF Classification of Enron Email
5. Discussion Tracking via PARAFAC/Tensor Factorization
6. Summary and References

- NNMF (Nonnegative Matrix Factorization) can be used to approximate high-dimensional data having nonnegative components.
- Lee and Seung (1999) demonstrated its use as a sum-by-parts representation of image data in order to both identify and classify image features.
- Xu et al. (2003) demonstrated how NNMF-based indexing could outperform SVD-based Latent Semantic Indexing (LSI) for some information retrieval tasks.
**NNMF for Image Processing**

Sparse NNMF versus Dense SVD Bases; Lee and Seung (1999)

**NNMF for Text Mining (Medlars)**

Interpretable NNMF feature vectors; Langville et al. (2006)

---

**Derivation**

- Given an $m \times n$ term-by-document (sparse) matrix $X$.
- Compute two reduced-dim. matrices $W, H$ so that $X \simeq WH$; $W$ is $m \times r$ and $H$ is $r \times n$, with $r \ll n$.
- Optimization problem:

$$
\min_{W,H} \|X - WH\|_F^2,
$$

subject to $W_{ij} \geq 0$ and $H_{ij} \geq 0$, $\forall i,j$.

- General approach: construct initial estimates for $W$ and $H$ and then improve them via alternating iterations.

---

**Minimization Challenges and Formulations**

[Berry et al., 2007]

- **Local Minima**: Non-convexity of functional $f(W, H) = \frac{1}{2} \|X - WH\|_F^2$ in both $W$ and $H$.
- **Non-unique Solutions**: $WDD^{-1}H$ is nonnegative for any nonnegative (and invertible) $D$.

**NNMF Formulations**:

- Information theoretic – Lee and Seung (2001)
- Compensate for feature redundancy – Guillamet et al. (2001)
- Extract spatially localized features – Wang et al. (2004)
Multiplicative Method (MM)

- Multiplicative update rules for $W$ and $H$ (Lee and Seung, 1999):
  1. Initialize $W$ and $H$ with nonnegative values, and scale the columns of $W$ to unit norm.
  2. Iterate for each $c$, $j$, and $i$ until convergence or after $k$ iterations:
     1. $H_{cj} \leftarrow H_{cj} \frac{(W^TX)_{cj}}{(W^TWH)_{cj} + \epsilon}$
     2. $W_{ic} \leftarrow W_{ic} \frac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$
     3. Scale the columns of $W$ to unit norm.
- Setting $\epsilon = 10^{-9}$ will suffice to avoid division by zero.

Lee and Seung MM Convergence

- **Convergence**: when the MM algorithm converges to a limit point in the interior of the feasible region, the point is a stationary point. The stationary point may or may not be a local minimum. If the limit point lies on the boundary of the feasible region, one cannot determine its stationarity [Berry et al., 2007].
- **Modifications**: Gonzalez and Zhang (2005) accelerated convergence somewhat but stationarity issue remains; Lin (2005) modified the algorithm to guarantee convergence to a stationary point; Dhillon and Sra (2005) derived update rules that incorporate weights for the importance of certain features of the approximation.

Alternating Least Squares Formulation

**Basic ALS Approach:**
ALS algorithms exploit the convexity of $W$ or $H$ (not both) in the underlying optimization problem. The basic iteration involves

- (LS) Solve for $H$ in $W^TWH = W^TX$.
- (NN) Set negative elements of $H$ to 0.
- (LS) Solve for $W$ in $HH^TW^T = HX^T$.
- (NN) Set negative elements of $W$ to 0.

**ALS Recovery and Constraints:**
- Unlike the MM algorithm, an element of $W$ (or $H$) that becomes 0 does not have to remain 0; method can escape/recover from a poor path.
- Paatero (1999) and Langville et al. (2006) have improved the computational complexity of the ALS approach; sparsity and nonnegativity contraints are enforced.
Alternating Least Squares Algorithms, contd.

**ALS Convergence:**
- Polak (1971) showed that every limit point of a sequence of alternating variable iterates is a stationary point.
- Lawson and Hanson (1995) produced the Non-Negative Least Squares (NNLS) that was shown to converge to a local minimum.
- The price for convergence of ALS algorithms is the usual high cost per iteration – Bro and de Jong (1997).

Enforcing Statistical Sparsity (Hoyer)

- From neural network applications, Hoyer (2002) enforced statistical sparsity for the weight matrix $H$ in order to enhance the parts-based data representations in the matrix $W$.
- Mu et al. (2003) suggested a regularization approach to achieve statistical sparsity in the matrix $H$: **point count regularization**; penalize the number of nonzeros in $H$ rather than $\sum_{ij} H_{ij}$.
- Goal of increased sparsity – better representation of parts or features spanned by the corpus ($X$) [Berry and Browne, 2005].

GD-CLS – Hybrid Approach

- First use MM to compute an approximation to $W$ for each iteration – a gradient descent (GD) optimization step.
- Then, compute the weight matrix $H$ using a constrained least squares (CLS) model to penalize non-smoothness (i.e., non-sparsity) in $H$ – common Tikhonov regularization technique used in image processing (Prasad et al., 2003).
- Convergence to a non-stationary point evidenced (proof still needed).

GD-CLS Algorithm

1. Initialize $W$ and $H$ with nonnegative values, and scale the columns of $W$ to unit norm.
2. Iterate until convergence or after $k$ iterations:
   1. $W_{ic} \leftarrow W_{ic} \frac{(XH^T)_{ic}}{(WHH^T)_{ic} + \epsilon}$, for $c$ and $i$
   2. Rescale the columns of $W$ to unit norm.
   3. Solve the constrained least squares problem:
      $$\min_{H_j} \{ \|X_j - WH_j\|_2^2 + \lambda \|H_j\|_2^2 \},$$
      where the subscript $j$ denotes the $j^{th}$ column, for $j = 1, \ldots, m$.
3. Any negative values in $H_j$ are set to zero. The parameter $\lambda$ is a regularization value that is used to balance the reduction of the metric $\|X_j - WH_j\|_2^2$ with enforcement of smoothness and sparsity in $H$. 
Two Penalty Term Formulation

- Introduce smoothing on $W_k$ (feature vectors) in addition to $H^k$:
\[
\min_{W, H} \{ \|X - WH\|_F^2 + \alpha \|W\|_F^2 + \beta \|H\|_F^2 \},
\]
where $\| \cdot \|_F$ is the Frobenius norm.
- Constrained NNMF (CNMF) iteration:
\[
H_{cj} \leftarrow H_{cj} \frac{(W^T X)_{cj} - \beta H_{cj}}{(W^T W)_{cj} + \epsilon}
\]
\[
W_{ic} \leftarrow W_{ic} \frac{(XH^T)_{ic} - \alpha W_{ic}}{(WHH^T)_{ic} + \epsilon}
\]

Improving Feature Interpretability

Gauging Parameters for Constrained Optimization

How sparse (or smooth) should factors $(W, H)$ be to produce as many interpretable features as possible?

To what extent do different norms ($L_1, L_2, L_\infty$) improve/degrade feature quality or span? At what cost?

Can a common nonnegative feature space be built from objects in both images and text? Are there opportunities for multimodal document similarity?

Anomaly Classification (ASRS)

- Classify events described by documents from the Airline Safety Reporting System (ASRS) into 22 anomaly categories; contest from SDM07 Text Mining Workshop.
- General Text Parsing (GTP) Software Environment in C++ [Giles et al., 2003] used to parse both ASRS training set and a combined ASRS training and test set:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Terms</th>
<th>ASRS Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>15,722</td>
<td>21,519</td>
</tr>
<tr>
<td>Training+Test</td>
<td>17,994</td>
<td>28,596 (7,077)</td>
</tr>
</tbody>
</table>

Web links:
- **GTP**: [http://www.cs.utk.edu/~lsi](http://www.cs.utk.edu/~lsi)
- **ASRS**: [http://www.cs.utk.edu/tmw07](http://www.cs.utk.edu/tmw07)

Term Weighting Schemes

- **Assessment of Term Importance**: for $m \times n$ term-by-message matrix $X = [x_{ij}]$, define
\[
x_{ij} = l_{ij}g_i d_j,
\]
where $l_{ij}$ is the local weight for term $i$ occurring in message $j$, $g_i$ is the global weight for term $i$ in the subcollection, and $d_j$ is a document normalization factor (set $d_j = 1$).

- **Common Term Weighting Choices**

<table>
<thead>
<tr>
<th>Name</th>
<th>Local</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>txx</td>
<td>Term Frequency</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>$l_{ij} = f_{ij}$</td>
<td>$g_i = 1$</td>
</tr>
<tr>
<td>lex</td>
<td>Logarithmic</td>
<td>Entropy (Define: $p_{ij} = f_{ij}/\sum_j f_{ij}$)</td>
</tr>
<tr>
<td></td>
<td>$l_{ij} = \log(1 + f_{ij})$</td>
<td>$g_i = 1 + (\sum_j p_{ij}\log(p_{ij})/\log n)$</td>
</tr>
</tbody>
</table>
**Parameterization of NNMF Classifier**

- **Important Control Parameters:**
  - \(\alpha\), the threshold on the relevance score or (target value) \(t_{ij}\) for document \(i\) and anomaly/label \(j\); we use \(R\) submatrix of \(H\) to cluster documents by the \(k\) features — assume documents describing similar anomalies share similar features.
  - \(\delta\), the threshold on the column elements of \(H\), which will filter out the association of features with both the training (\(R\)) and test (\(T\)) documents;
  - \(\sigma\), the percentage of documents used to define the training set (or number of columns of \(R\)).

**Initialization Schematic**

**Anomaly to Feature Mapping and Scoring Schematic**

**Training/Testing Performance (ROC Curves)**

- Best/Worst ROC curves (False Positive Rate versus True Positive Rate)

<table>
<thead>
<tr>
<th>Anomaly</th>
<th>Type (Description)</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>22</td>
<td>Security Concern/Threat</td>
<td>.9040</td>
</tr>
<tr>
<td>5</td>
<td>Incursion (collision hazard)</td>
<td>.8977</td>
</tr>
<tr>
<td>4</td>
<td>Excursion (loss of control)</td>
<td>.8296</td>
</tr>
<tr>
<td>21</td>
<td>Illness/Injury Event</td>
<td>.8201</td>
</tr>
<tr>
<td>12</td>
<td>Traffic Proximity Event</td>
<td>.7954</td>
</tr>
<tr>
<td>7</td>
<td>Altitude Deviation</td>
<td>.7931</td>
</tr>
<tr>
<td>18</td>
<td>Aircraft Damage/Encounter</td>
<td>.7250</td>
</tr>
<tr>
<td>11</td>
<td>Terrain Proximity Event</td>
<td>.7234</td>
</tr>
<tr>
<td>9</td>
<td>Speed Deviation</td>
<td>.7060</td>
</tr>
<tr>
<td>10</td>
<td>Uncommanded (loss of control)</td>
<td>.6784</td>
</tr>
<tr>
<td>13</td>
<td>Weather Issue</td>
<td>.6287</td>
</tr>
<tr>
<td>2</td>
<td>Noncompliance (policy/proc.)</td>
<td>.6009</td>
</tr>
</tbody>
</table>
Effects of Alternative Norms in the Objective Function

- CVX: Matlab Software for Disciplined Convex Programming (Grant, Boyd, and Ye – Stanford) used for ALP (Alternate Linear Programming)
- Factor first 100 ASRS docs only ($X$ is $773 \times 100$) with $k = 10, 20, 40$ features (rank)

<table>
<thead>
<tr>
<th>Anomaly</th>
<th>Area(ROC) for Rank-10</th>
<th>Anomaly Count in Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$L_1$</td>
<td>$L_2$</td>
<td>$L_\infty$</td>
</tr>
<tr>
<td>2</td>
<td>0.6124</td>
<td>0.3971</td>
<td><strong>0.6986</strong></td>
</tr>
<tr>
<td>5</td>
<td><strong>0.6914</strong></td>
<td>0.6173</td>
<td>0.2099</td>
</tr>
<tr>
<td>7</td>
<td>0.4560</td>
<td>0.5600</td>
<td>0.5360</td>
</tr>
<tr>
<td>8</td>
<td>0.7037</td>
<td>0.3580</td>
<td><strong>0.5062</strong></td>
</tr>
<tr>
<td>10</td>
<td>0.6071</td>
<td>0.6071</td>
<td>0.3929</td>
</tr>
<tr>
<td>12</td>
<td><strong>0.5900</strong></td>
<td>0.5650</td>
<td>0.3900</td>
</tr>
<tr>
<td>13</td>
<td>0.7500</td>
<td>0.6964</td>
<td>0.6607</td>
</tr>
<tr>
<td>18</td>
<td><strong>0.8750</strong></td>
<td>0.4286</td>
<td><strong>0.6429</strong></td>
</tr>
</tbody>
</table>
### Effects of Alternative Norms in the Objective Function

<table>
<thead>
<tr>
<th>Anomaly</th>
<th>Area(ROC) for Rank-20</th>
<th>Area(ROC) for Rank-40</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$L_1$</td>
<td>$L_2$</td>
</tr>
<tr>
<td>2</td>
<td>0.4067</td>
<td>0.5981</td>
</tr>
<tr>
<td>5</td>
<td>0.3951</td>
<td>0.6543</td>
</tr>
<tr>
<td>7</td>
<td>0.5200</td>
<td>0.5680</td>
</tr>
<tr>
<td>8</td>
<td>0.6296</td>
<td><strong>0.7160</strong></td>
</tr>
<tr>
<td>10</td>
<td><strong>0.6786</strong></td>
<td>0.5357</td>
</tr>
<tr>
<td>12</td>
<td>0.5200</td>
<td>0.6950</td>
</tr>
<tr>
<td>13</td>
<td><strong>0.4464</strong></td>
<td>0.4286</td>
</tr>
<tr>
<td>18</td>
<td><strong>0.6786</strong></td>
<td>0.2679</td>
</tr>
</tbody>
</table>

### Sparsity of NMF factors (without constraints) can vary with $L_p$ norm (and rank); $X^{773 \times 100}$

<table>
<thead>
<tr>
<th>Rank</th>
<th>Factor</th>
<th>Number of Nonzeros</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$L_1$</td>
</tr>
<tr>
<td>10</td>
<td>$W$</td>
<td>7730</td>
</tr>
<tr>
<td>10</td>
<td>$H$</td>
<td><strong>728</strong></td>
</tr>
<tr>
<td>20</td>
<td>$W$</td>
<td>15140</td>
</tr>
<tr>
<td>20</td>
<td>$H$</td>
<td><strong>1182</strong></td>
</tr>
<tr>
<td>40</td>
<td>$W$</td>
<td>27831</td>
</tr>
<tr>
<td>40</td>
<td>$H$</td>
<td><strong>70</strong></td>
</tr>
</tbody>
</table>

### Email Collection

- By-product of the FERC investigation of Enron (originally contained 15 million email messages).
- This study used the improved corpus known as the Enron Email set, which was edited by Dr. William Cohen at CMU.
- This set had over 500,000 email messages. The majority were sent in the 1999 to 2001 timeframe.
Ongoing, problematic, development of the Dabhol Power Company (DPC) in the Indian state of Maharashtra.

Deregulation of the Calif. energy industry, which led to rolling electricity blackouts in the summer of 2000 (and subsequent investigations).

Revelation of Enron’s deceptive business and accounting practices that led to an abrupt collapse of the energy colossus in October, 2001; Enron filed for bankruptcy in December, 2001.

PRIVATE with Log-Entropy Weighting

- Identify rows of $H$ from $X \sim WH$ or $H^k$ with $\lambda = 0.1; r = 50$ feature vectors ($W_k$) generated by GD-CLS:

<table>
<thead>
<tr>
<th>Feature Index ($k$)</th>
<th>Cluster Size</th>
<th>Topic Description</th>
<th>Dominant Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>497</td>
<td>California</td>
<td>ca, cpuc, gov, socalgas, sempra, org, sce, gmssr, aelaw, ci</td>
</tr>
<tr>
<td>23</td>
<td>43</td>
<td>Louise Kitchen</td>
<td>evp, fortune, britain, woman, ceo, avon, fiorina, cfo, hewlett, packard</td>
</tr>
<tr>
<td>26</td>
<td>231</td>
<td>Fantasy football</td>
<td>game, wr, qb, play, rb, season, injury, updated, fantasy, image</td>
</tr>
</tbody>
</table>

(Cluster size $\equiv$ no. of $H^k$ elements $> row_{max}/10$)

PRIVATE with Log-Entropy Weighting

- Additional topic clusters of significant size:

<table>
<thead>
<tr>
<th>Feature Index ($k$)</th>
<th>Cluster Size</th>
<th>Topic Description</th>
<th>Dominant Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>233</td>
<td>Texas</td>
<td>UT, orange, longhorn, longhorn[s], texas, true, truorange, recruiting, oklahoma, defensive</td>
</tr>
<tr>
<td>34</td>
<td>65</td>
<td>Enron collapse</td>
<td>partnership[s], fastow, shares, sec, stock, shareholder, investors, equity, lay</td>
</tr>
<tr>
<td>39</td>
<td>235</td>
<td>Emails about India</td>
<td>dabhol, dpc, india, mseb, maharashtra, indian, lenders, delhi, foreign, minister</td>
</tr>
</tbody>
</table>
2001 Topics Tracked by GD-CLS

Term Distribution in Feature Vectors

<table>
<thead>
<tr>
<th>Terms</th>
<th>Wt</th>
<th>Lambda 0.1</th>
<th>Lambda 0.01</th>
<th>Lambda 0.001</th>
<th>Alpha 0.1</th>
<th>Alpha 0.01</th>
<th>Alpha 0.001</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blackouts</td>
<td>0.508</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>Cal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stocks</td>
<td>0.511</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UT</td>
<td>0.517</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Texasfoot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronicle</td>
<td>0.523</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indian</td>
<td>0.527</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>India</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fastow</td>
<td>0.531</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gas</td>
<td>0.531</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFO</td>
<td>0.556</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Kitchen</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Californians</td>
<td>0.557</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>Cal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solar</td>
<td>0.570</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partnerships</td>
<td>0.576</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workers</td>
<td>0.577</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maharashtra</td>
<td>0.591</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>India</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mseb</td>
<td>0.605</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>India</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beach</td>
<td>0.611</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ljm</td>
<td>0.621</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tues</td>
<td>0.626</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPPS</td>
<td>0.644</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Cal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rebates</td>
<td>0.647</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ljm2</td>
<td>0.688</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Collapse</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\rho = 50$ features, lex term weighting, $\lambda = 0.1$


Annotation Project

- Subset of 2001 private collection:

<table>
<thead>
<tr>
<th>Month</th>
<th>Total</th>
<th>Classified</th>
<th>Usable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan,Sep</td>
<td>5591</td>
<td>1100</td>
<td>699</td>
</tr>
<tr>
<td>Feb</td>
<td>2804</td>
<td>900</td>
<td>460</td>
</tr>
<tr>
<td>Mar</td>
<td>3525</td>
<td>1200</td>
<td>533</td>
</tr>
<tr>
<td>Apr</td>
<td>4273</td>
<td>1500</td>
<td>705</td>
</tr>
<tr>
<td>May</td>
<td>4261</td>
<td>1800</td>
<td>894</td>
</tr>
<tr>
<td>June</td>
<td>4324</td>
<td>1025</td>
<td>538</td>
</tr>
<tr>
<td>Total</td>
<td>24778</td>
<td>7525</td>
<td>3829</td>
</tr>
</tbody>
</table>

- Approx. 40 topics identified after NNMF initial clustering with $k = 50$ features.

Annotation Project, contd.

- Human classifiers: M. Browne (extensive background reading on Enron collapse) and B. Singer (junior Economics major).
- Classify email content versus type (see UC Berkeley Enron Email Analysis Group http://bailando.sims.berkeley.edu/enron_email.html
- As of June 2007, distributed by the by U. Penn LDC (Linguistic Data Consortium); see www.ldc.upenn.edu

Citation:

Dr. Michael W. Berry, Murray Browne and Ben Signer, 2007
2001 Topic Annotated Enron Email Data Set
Linguistic Data Consortium, Philadelphia
Multidimensional Data Analysis via PARAFAC

Mathematical Notation

- Kronecker product
  \[ A \otimes B = \begin{pmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & \ddots & \vdots \\ A_{m1}B & \cdots & A_{mn}B \end{pmatrix} \]

- Khatri-Rao product (columnwise Kronecker)
  \[ A \odot B = \left( A_1 \otimes B_1, \ldots, A_n \otimes B_n \right) \]

- Outer product
  \[ A_1 \odot B_1 = \begin{pmatrix} A_{11}B_{11} & \cdots & A_{11}B_{m1} \\ \vdots & \ddots & \vdots \\ A_{m1}B_{11} & \cdots & A_{m1}B_{m1} \end{pmatrix} \]

Temporal Assessment via PARAFAC

Email graph

PARAFAC Representations

- PARAllel FACtors (Harshman, 1970)
- Also known as CANDECOMP (Carroll & Chang, 1970)
- Typically solved by Alternating Least Squares (ALS)

Alternative PARAFAC formulations

\[ X_{ijk} \approx \sum_{i=1}^{r} A_{ir}B_{jr}C_{kr} \]

\[ X' = \sum_{i=1}^{r} A_i \odot B_i \odot C_i \], where \( X' \) is a 3-way array (tensor).

\[ X_k \approx A \ \text{diag}(C_k) \ \text{diag}(B) \]

\[ X^{I \times JK} \approx A(C \odot B)^T \], where \( X \) is matricized.
PARAFAC (Visual) Representations

**Scalar form**
\[ \mathbf{A} \mathbf{B}^T \]

**Outer product form**
\[ \| \mathbf{X}^{1 \times J \times K} - \mathbf{A} (\mathbf{C} \odot \mathbf{B})^T \|_F \]

**Tensor slice form**
\[ \mathbf{X}^{1 \times J \times K} = \mathbf{A} \mathbf{B}^T \]

**Matrix form**
\[ \| \mathbf{X}^{1 \times J \times K} - \mathbf{B} (\mathbf{C} \odot \mathbf{A})^T \|_F \]

**Nonnegative PARAFAC Algorithm**
- Adapted from (Mørup, 2005) and based on NNMF by (Lee and Seung, 2001)
  \[ \| \mathbf{X}^{1 \times J \times K} - \mathbf{A} (\mathbf{C} \odot \mathbf{B})^T \|_F = \| \mathbf{X}^{1 \times J \times K} - \mathbf{B} (\mathbf{C} \odot \mathbf{A})^T \|_F \]
- Minimize over \( \mathbf{A} \), \( \mathbf{B} \), \( \mathbf{C} \) using multiplicative update rule:
  \[ A_{i\rho} \leftarrow A_{i\rho} \frac{(X^{1 \times J \times K} Z)_{i\rho}}{(AZ^T Z)_{i\rho} + \epsilon}, \quad Z = (\mathbf{C} \odot \mathbf{B}) \]
  \[ B_{j\rho} \leftarrow B_{j\rho} \frac{(X^{1 \times J \times K} Z)_{j\rho}}{(BZ^T Z)_{j\rho} + \epsilon}, \quad Z = (\mathbf{C} \odot \mathbf{A}) \]
  \[ C_{k\rho} \leftarrow C_{k\rho} \frac{(X^{1 \times J \times K} Z)_{k\rho}}{(CZ^T Z)_{k\rho} + \epsilon}, \quad Z = (\mathbf{B} \odot \mathbf{A}) \]

Discussion Tracking Using Year 2001 Subset
- 197 authors (From: user_id@enron.com) monitored over 12 months;
- Parsing 34,427 email subset with a base dictionary of 121,393 terms (derived from 517,431 emails) produced 69,157 unique terms; (term-author-month) array \( \mathbf{X} \) has \( \sim \) 1 million nonzeros.
- Rank-25 tensor: \( \mathbf{A} \) (69,157 \( \times \) 25), \( \mathbf{B} \) (197 \( \times \) 25), \( \mathbf{C} \) (12 \( \times \) 25)

<table>
<thead>
<tr>
<th>Month</th>
<th>Emails</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>7,050</td>
</tr>
<tr>
<td>Feb</td>
<td>6,387</td>
</tr>
<tr>
<td>Mar</td>
<td>6,871</td>
</tr>
<tr>
<td>Apr</td>
<td>7,382</td>
</tr>
<tr>
<td>May</td>
<td>5,989</td>
</tr>
<tr>
<td>Jun</td>
<td>2,510</td>
</tr>
<tr>
<td>Jul</td>
<td>2,166</td>
</tr>
<tr>
<td>Aug</td>
<td>2,074</td>
</tr>
<tr>
<td>Sep</td>
<td>2,192</td>
</tr>
<tr>
<td>Oct</td>
<td>5,719</td>
</tr>
<tr>
<td>Nov</td>
<td>4,011</td>
</tr>
<tr>
<td>Dec</td>
<td>1,382</td>
</tr>
</tbody>
</table>

Tensor-Generated Group Discussions
- NNTF Group Discussions in 2001
- 197 authors; 8 distinguishable discussions
- “Kaminski/Education” topic previously unseen
Gantt Charts from PARAFAC Models

Day-level Analysis for PARAFAC (Three Groups)
- Rank-25 tensor for 357 out of 365 days of 2001: $A$ (69 $\times$ 157 $\times$ 25), $B$ (197 $\times$ 25), $C$ (357 $\times$ 25)
- Groups 3,4,5 (out of 25 from $C$) [Bader et al., 2008]:

Day-level Analysis for NN-PARAFAC (Three Groups)
- Rank-25 tensor (best minimizer) for 357 out of 365 days of 2001: $A$ (69, 157 $\times$ 25), $B$ (197 $\times$ 25), $C$ (357 $\times$ 25)
- Groups 1,7,8 (out of 25 from $C$):

Day-level Analysis for NN-PARAFAC (Two Groups)
- Groups 20 (California Energy) and 9 (Football) (from $C$ factor of best minimizer) in day-level analysis of 2001:
Four-way Tensor Results (Sept. 2007)

- Apply NN-PARAFAC to term-author-recipient-day array (39,573 × 197 × 197 × 357); construct a rank-25 tensor (best minimizer among 10 runs).
- Goal: track more focused discussions between individuals/small groups; for example, betting pool (football).

![Graph showing conversation about FERC and Regional Transmission Organizations (RTOs).]

Four-way Tensor Results (October 2007)

- Four-way tensor exposed conversation confirming bank fraud related to the natural gas reserves in the Bammel Storage field (Texas)—“The Enron whistle-blower who wasn’t” by G. Farrell, USA Today, Oct. 11, 2007

NNTF Optimal Rank?

- No known algorithm for computing the rank of a $k$-way array for $k \geq 3$ [Kruskal, 1989].
- The maximum rank is not a closed set for a given random tensor.
- The maximum rank of a $m \times n \times k$ tensor is unknown; one weak inequality is given by
  \[ \max\{m, n, k\} \leq \text{rank} \leq \min\{m \times n, m \times k, n \times k\} \]
- For our rank-25 NNTF, the size of the relative residual norm suggests we are still far from the maximum rank of the 3-way and 4-way arrays.
Conclusions (NNMF for ASRS)

- Training phase was a good predictor of performance (for most anomaly categories).
- Obvious room for improvement in matching certain anomalies (e.g., 2. Noncompliance).
- Summarization of anomalies using NNMF features needs further work.
- Effects of sparsity contraints on NNMF versus element-wise filtering of H should be studied.
- Effects of using different $L_p$ norms in the objective function need more analysis.

Conclusions (NNMF/NNTF for Enron)

- GD-CLS/NNMF Algorithm can effectively produce a parts-based approximation $X \simeq WH$ of a sparse term-by-message matrix $X$.
- Smoothing on the features matrix (W) as opposed to the weight matrix H forces more reuse of higher weighted (log-entropy) terms; yields potential control vocabulary for topic tracking.
- Surveillance systems based on NNMF/NNTF algorithms show promise for monitoring discussions without the need to isolate or perhaps incriminate individuals.
- Potential applications include the monitoring/tracking of company morale, employee feedback to policy decisions, extracurricular activities, and blog discussions.

Research Tracks

- Further work needed in determining effects of alternative term weighting schemes (for $X$) and choices of control parameters (e.g., $\alpha, \beta$) for CNMF and NNTF/PARAFAC.
- How many dimensions (factors) for NNTF/PARAFAC are really needed for mining electronic mail and similar corpora? And, at what scale should each dimension be measured (e.g., time)?
- Compare convergence and accuracy of CNMF with NMU (Nonnegative Matrix Underapproximation); joint work with N. Gillis and F. Glineur (Univ. of Louvain, Belgium)

Improving Summarization and Steering

What versus why:

Extraction of textual concepts still requires human interpretation (in the absence of ontologies or domain-specific classifications).

How can previous knowledge or experience be captured for feature matching (or pruning)?

To what extent can feature vectors be annotated for future use or as the text collection is updated? What is the cost for updating NNMF/NNTF models?
NNTF for Visual Analytics (VA)

- VAST 2007 Contest: 1,455 news stories/emails/blog entries with underlying ecoterrorism activity to be uncovered.
- Who/What/When/Where questions using tagged entities (Person, Location, Organization, Money) and context (terms).
  (See [http://www.cs.umd.edu/hcil/VASTcontest07](http://www.cs.umd.edu/hcil/VASTcontest07))

Ecoterrorism Scenario of VAST 2007 Contest

<table>
<thead>
<tr>
<th>Month</th>
<th>Score</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>0.2</td>
<td>Group 6</td>
</tr>
<tr>
<td>Feb</td>
<td>0.4</td>
<td>Group 9</td>
</tr>
<tr>
<td>Mar</td>
<td>0.8</td>
<td>Group 15</td>
</tr>
<tr>
<td>Apr</td>
<td>1.0</td>
<td>Group 18</td>
</tr>
</tbody>
</table>

Score documents (news stories) against terms and entities from identifiable (classifiable) NTF factors:

Group 20

- Entities (15 total):
  - Scoring: [id, name]
    - 0.2252609: [scott roberts, iron and steel statistics bureau]
    - 0.1827936: [brazil, china]

- Terms (35 total):
  - Scoring: [id, term]
    - 0.2140977: [energy, economic]
    - 0.1915396: [china, nuclear]
    - 0.1502502: [power, oil]
    - 0.1321501: [beijing, roberts]

The sentences that contribute most to this story’s score:

- This is a very significant victory, said Zhang Luping, head of the Beijing Human and Animal Environmental Education Center.
- Beijing officials got a taste of the new attitude when they sought to discourage pet ownership through high license fees in 1995. A policy that was largely reversed under pressure from outspoken residents.

FutureLens: NNTF Output Visualization

- Java-based visualization environment (adapted from FeatureLens at HCIL, Univ. of Maryland)

For Further Reading

For Further Reading (contd.)

- P. Hoyer.
  Nonnegative Matrix Factorization with Sparseness Constraints.

- J.T. Giles and L. Wo and M.W. Berry.
  GTP (General Text Parser) Software for Text Mining.

- W. Xu, X. Liu, and Y. Gong.
  Document-Clustering based on Nonneg. Matrix Factorization.

- J.B. Kruskal.
  Rank, Decomp. , and Uniqueness for 3-way and n-way Arrays.