Joint NMF for Hybrid Clustering based on Content and Connection Structure

Rundong Du\textsuperscript{1}, Barry Drake\textsuperscript{2,3} and Haesun Park\textsuperscript{2}

1. School of Mathematics
2. School of Computational Science and Engineering
3. Georgia Tech Research Institute

Georgia Institute of Technology, Atlanta, GA

SIAM ALA
HKBU, May, 2018

Supported in part by
Constrained Low Rank Approximations for Scalable Data Analytics

Objectives:
- Model text and graph clustering problems
- Design, verify, and deploy scalable numerical alg. and software
- Develop divide-and-conquer methods to handle problems of larger size for various computing environments

Goal: Orders of magnitude speed improvements over existing data analytics methods and solutions of higher quality

Why CLRA?
- Utilize advances in numerical linear algebra and optimization
- Exploit software such as BLAS and LAPACK
- Behavior of algorithms easier to analyze
- Facilitates design of MPI based algorithms for scalable solutions
- Can easily be modified for various problem demands, e.g. adaptive methods
Clustering: data clustering, topic modeling, graph clustering, community detection, hybrid clustering...
Nonnegative Matrix Factorization (NMF)

(Lee&Seung 99, Paatero&Tapper 94)

Given $X \in \mathbb{R}_+^{m \times n}$ and a desired rank $k \ll \min(m, n)$, find $W \in \mathbb{R}_+^{m \times k}$ and $H \in \mathbb{R}_+^{k \times n}$ s.t. $X \approx WH$.

Notation:
- $\mathbb{R}_+$: nonnegative real numbers
- $\min_{W \geq 0, H \geq 0} \|X - WH\|_F$
- Nonconvex

**NMF for Clustering?**

Objective functions for K-means and NMF may look the same:

$$
\min \sum_i \|x_i - w_{\sigma_i}\|^2 = \min \|X - WH\|^2_F
$$

(Ding et al. 05; Kim & Park, 08; Xu et al. S03; Cai et al. 08; Kim & Park Bio 07, etc.)

$\sigma_i = j$ when $i$-th point is assigned to $j$-th cluster ($j \in \{1, \cdots, k\}$).

But, the constraints are different:

- **K-means**: $H \in \{0, 1\}^{k \times n}$, $1_k^T H = 1_n^T$
- **NMF**: $W \geq 0$, $H \geq 0$
Block Coordinate Descent (BCD) for NMF

\[
\min f(z) = f(W, H) = \|X - WH\|_F, \text{ s.t. } z \in Z = Z_1 \times \cdots \times Z_p
\]

- **BCD** generates \(z^{(k+1)} = (z_1^{(k+1)}, \ldots, z_p^{(k+1)})\) by
  \[
  z_i^{(k+1)} = \arg \min_{\xi \in Z_i} f(z_1^{(k+1)}, \ldots, z_{i-1}^{(k+1)}, \xi, z_{i+1}^{(k)}, \ldots, z_p^{(k)})
  \]

- **Th. (Bertsekas, 99)**: Suppose \(f\) is continuously differentiable over the Cartesian product of closed, convex sets \(Z_1, Z_2, \ldots, Z_p\) and for each \(i\), the minimum is uniquely attained. Then every limit point of the sequence generated by the BCD method \(\{z^{(k)}\}\) is a stationary point.
**NMF for text clustering:** (J. Kim and HP, SISC 11; J.Kim, Y. He, and HP, JOGO 14)

**SymNMF for graph clustering:** (D. Kuang, S. Yun, and HP, JOGO 15)

<table>
<thead>
<tr>
<th>Input</th>
<th>Eigenbasis</th>
<th>Nonnegative basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature-Data matrix</td>
<td>SVD/PCA</td>
<td>NMF/Affine NMF</td>
</tr>
<tr>
<td>Data-Data matrix</td>
<td>Spectral clustering</td>
<td>SymNMF</td>
</tr>
</tbody>
</table>
NMF Performance for Clustering and Topic Modeling

20Newsgroups (36,568 × 18,221)

- HierNMF2
- DC-NMF
- CLUTO
- nmf-flat
- kmeans–hier
- kmeans–flat

NMI vs. k for different methods.

Reuter (12,411 × 7,984)

- HierNMF2
- DC-NMF
- Mallet–LDA
- AnchorRecovery
- XRAY–greedy
- Hottopixx

NMI vs. k for different methods.

20Newsgroups (36,568 × 18,221)

- HierNMF2
- DC-NMF
- Mallet–LDA
- AnchorRecovery
- XRAY–greedy
- Hottopixx

NMI vs. k for different methods.

Cora (154,134 × 29,169)

- HierNMF2
- DC-NMF
- Mallet–LDA
- AnchorRecovery
- XRAY–greedy
- Hottopixx

NMI vs. k for different methods.

Source: R. Du, D. Kuang, B. Drake, HP, to appear in JOGO
Methods Compared:

- CLUTO
  (Y. Zhao and G. Karypis, 01)
- HierNMF2
  (D. Kuang and HP, 13)
- DC-NMF
  (D. Kuang et al., 17)
- AnchorRecovery
  (S. Arora et al., 13)
- Mallet-LDA
  (A. K. McCallum, 02; D. Blei et al., 03)
- XRAY
  (A. Kumar et al., 13)
- Hottopixx
  (V. Bittorf et al., 12)

Data size (# of topics):

- RCV1: 149K x 765K (60)
- Wiki-4.5M: 2.3M x 4.1M (80)

HierNMF2 on Wiki4.5M found 80 topics in 43.1 min on MacbookPro, Intel Core i7 2.6 GHz, 4 cores, 16 GB memory. WEKA K-means did not finish. CLUTO ran out of memory.

SmallK [http://smallk.github.io](http://smallk.github.io)
JointNMF from NMF and SymNMF

\[
\min_{W \geq 0, H \geq 0} \| X - WH \|_F \quad \min_{H \geq 0} \| S - H^T H \|_F
\]
NMF: content/text clustering
SymNMF: graph clustering
\[
X \in \mathbb{R}^{m \times n}: \text{term} \times \text{doc} \quad S \in \mathbb{R}^{n \times n}: \text{doc} \times \text{doc}, S^T = S
\]
\[
W \in \mathbb{R}^{m \times k}, H \in \mathbb{R}^{k \times n}, k \ll \min\{m, n\}
\]

JointNMF for Hybrid Clustering:
\[
\min_{W \geq 0, H \geq 0} \alpha_1 \| X - WH \|_F^2 + \alpha_2 \| S - H^T H \|_F^2
\]
JointNMF and Block Coordinate Descent (BCD)

Formulation:

$$\min_{W \geq 0, H \geq 0} \| X - WH \|_F^2 + \alpha \| S - H^T H \|_F^2$$

Recast for the BCD framework:

$$\min_{W, H, \tilde{H} \geq 0} \| X - WH \|_F^2 + \alpha \| S - \tilde{H}^T H \|_F^2 + \beta \| \tilde{H} - H \|_F^2$$

3-block coordinate descent:

- Solve $W$: $\min_{W \geq 0} \| H^T W^T - X^T \|_F$
- Solve $\tilde{H}$: $\min_{\tilde{H} \geq 0} \| \begin{bmatrix} \sqrt{\alpha} H^T \\ \sqrt{\beta} I_k \end{bmatrix} \tilde{H} - \begin{bmatrix} \sqrt{\alpha} S \\ \sqrt{\beta} H \end{bmatrix} \|_F$
- Solve $H$: $\min_{H \geq 0} \| \begin{bmatrix} W \\ \sqrt{\alpha} \tilde{H}^T \\ \sqrt{\beta} I_k \end{bmatrix} H - \begin{bmatrix} X \\ \sqrt{\alpha} S \\ \sqrt{\beta} \tilde{H} \end{bmatrix} \|_F$
Hybrid Clustering: US Patent Data

**Data source:** PatentsView (www.patentsview.org)
- 80,728,766 citations between those patents
- 233,111 ground truth clusters
- We selected 13 subgroups

**F1 score when compared to ground truth:**

\[
F_1 = \frac{1}{2} \left( \frac{1}{k} \sum_{i=1}^{k} \max_{j} F_1(A_i, B_j) + \frac{1}{k'} \sum_{j=1}^{k'} \max_{i} F_1(B_j, A_i) \right) .
\]

<table>
<thead>
<tr>
<th>Class</th>
<th>Joint NMF</th>
<th>NMF</th>
<th>SymNMF</th>
<th>PCL-DC-1</th>
<th>PCL-DC-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A22</td>
<td>0.3730</td>
<td>0.2293</td>
<td>0.3457</td>
<td>0.1351</td>
<td>0.1369</td>
</tr>
<tr>
<td>C06</td>
<td>0.2257</td>
<td>0.1830</td>
<td>0.2004</td>
<td>0.1156</td>
<td>0.1158</td>
</tr>
<tr>
<td>C14</td>
<td>0.3584</td>
<td>0.3191</td>
<td>0.3578</td>
<td>0.2692</td>
<td>0.2659</td>
</tr>
<tr>
<td>D02</td>
<td>0.2990</td>
<td>0.2131</td>
<td>0.2683</td>
<td>0.1756</td>
<td>0.2268</td>
</tr>
<tr>
<td>D10</td>
<td>0.3046</td>
<td>0.2452</td>
<td>0.2783</td>
<td>0.1612</td>
<td>0.2999</td>
</tr>
<tr>
<td>F22</td>
<td>0.3006</td>
<td>0.2211</td>
<td>0.2926</td>
<td>0.1533</td>
<td>0.1388</td>
</tr>
</tbody>
</table>

JointNMF:

$$\min_{W \geq 0, H \geq 0} \|X - WH\|_F^2 + \alpha \|S - H^T H\|_F^2$$

**Note:** The basis $W$ for the content space is computed and the representation (coordinates) of the documents in $H$ reflects their content and linkage information.

**Citation prediction for a new document $x$:**

$$\min_{h \geq 0} \|x - Wh\|_2$$

and then compare $h$ with column vectors in $H$, via inner product or cosine similarity.

**Baseline methods:**

**NMF-1:** $\min_{W \geq 0, H \geq 0} \|X - WH\|_F$

**NMF-2:** $\min_{W \geq 0, H \geq 0, h \geq 0} \|[X, x] - W[H, h]\|_F$

**Naive:** count number of words shared by two documents
Citation Prediction: Tests on cit-HepTh Data Set

Abstract, Cosine Similarity

Title, Cosine Similarity

* Data source: SNAP (http://snap.stanford.edu/data/)
JointNMF for Clustering of Hypergraph with Edge Content

\[
\min_{W \geq 0, H \geq 0} \|X - WH\|_F^2 + \alpha \|S - H^T H\|_F^2
\]

- Hypergraph: an edge can join more than two vertices
- Incidence matrix \( M \): vertices \( \times \) hyperedges in hypergraph
- Dual hypergraph: vertices and hyperedges are interchanged, incidence matrix: \( M^T \)
- JointNMF can be applied as far as one of the dimensions in \( X \) and \( S \) is common.
- In case of email data:
  - ex1. \( X \): term-email and \( S \): email-email relationship
  - ex2. \( X \): term-people and \( S \): people-people relationship
  - Various ways to represent the relationships in \( S \) from a hypergraph
Case of Email Data: Content and Link Info Representations

Email 1
From: CEO
To: Manager 1, Staff 1
...

Email 2
From: Manager 1
To: Staff 1, 2 and 3
...

- Email content in a term-email matrix $X$
- Email-email relationship $S$ from the dual hypergraph based on the incidence matrix $M^T$
- $\min_{H \geq 0} \| S - H^T H \|_F$ is a relaxation of minimizing the normalized hypergraph cut

clusters of emails $\xrightarrow{\text{people involved}}$ clusters of people

Other representation:
- Keep the incidence matrix $M$ (person-email relation)
- Construct similarity matrix for email-email relationship using email content and construct corresponding normalized graph laplacian $L$.
- Solve $\min_{W, H} \| M - WH \|_F^2 + \lambda \text{tr} (HLH^T)$
Case Study: Enron Email Data Set

Frequency of number of memberships

<table>
<thead>
<tr>
<th>#memberships</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>#employees</td>
<td>1069</td>
<td>149</td>
<td>45</td>
<td>17</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

People with $j$ memberships ($j \geq 6$)

<table>
<thead>
<tr>
<th>$j$</th>
<th>Name</th>
<th>Position in Enron</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Steven Kean</td>
<td>Chief of staff</td>
</tr>
<tr>
<td>7</td>
<td>Jeff Dasovich</td>
<td>Governmental affairs executive</td>
</tr>
<tr>
<td></td>
<td>Susan Mara</td>
<td>California director of Regulatory Affairs</td>
</tr>
<tr>
<td></td>
<td>Richard Shapiro</td>
<td>VP of regulatory affairs</td>
</tr>
<tr>
<td></td>
<td>Paul Kaufman</td>
<td>VP of Government Affairs</td>
</tr>
<tr>
<td>6</td>
<td>James Steffes</td>
<td>VP of Government Affairs</td>
</tr>
<tr>
<td></td>
<td>Tim Belden</td>
<td>Head of trading</td>
</tr>
<tr>
<td></td>
<td>Richard Sanders</td>
<td>VP of Enron Whole Sale Services</td>
</tr>
<tr>
<td></td>
<td>Joe Hartsoe</td>
<td>VP of Federal Regulatory Affairs</td>
</tr>
</tbody>
</table>

Topic keywords of clusters

<table>
<thead>
<tr>
<th># Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 ubs, warburg, forecast, confidential, win</td>
</tr>
<tr>
<td>1 blackberry, handheld, wireless</td>
</tr>
<tr>
<td>2 california, power, confidential, tariff, pursuant</td>
</tr>
<tr>
<td>3 caiso, refund, ferc, proceedings</td>
</tr>
<tr>
<td>4 burrito, peace, things, price, market, board, california</td>
</tr>
<tr>
<td>5 document, fax, tonight, sign, back, attach, thanks</td>
</tr>
<tr>
<td>6 wholesale, policy, compliance, receipt, legal, service</td>
</tr>
<tr>
<td>7 enron, please, know, meeting, contact, call, any, time</td>
</tr>
<tr>
<td>8 london, conference, meeting, next, week</td>
</tr>
<tr>
<td>9 handheld, blackberry, wireless, agreement, confidential</td>
</tr>
<tr>
<td>10 testify, witness, fault, burden, cut, budget</td>
</tr>
<tr>
<td>11 california, electricity, energy, price, market, rate, bill</td>
</tr>
<tr>
<td>12 recommendation, template, participant, management</td>
</tr>
<tr>
<td>13 passcode, please, effective, confidential, change</td>
</tr>
<tr>
<td>14 stanford, university, expert, try, best, mail, california</td>
</tr>
<tr>
<td>15 account, invoice, trust, fund, transfer</td>
</tr>
<tr>
<td>16 expense, report, employee, name, approve, amount</td>
</tr>
<tr>
<td>17 folder, audit, access, apollo, email, sensitivity, server</td>
</tr>
<tr>
<td>18 sent, talk, presentation, infrastructure, amendment</td>
</tr>
<tr>
<td>19 hpl, aep, agreement, compete, deal, arrangement</td>
</tr>
</tbody>
</table>

Data source: a subset of 1702 emails from the Enron Email data set, extracted by a group from SIMS, UC Berkeley.
Representation of a Hypergraph with Content

Representation of a Hypergraph
- Symmetrize into an adjacency matrix?
- Leave incidence matrix as it is?
- Directed hypergraph for sender/receiver relationships?

\[
\begin{bmatrix}
-1 & \cdots \\
1 & \cdots \\
1 & \cdots \\
1 & \cdots \\
\vdots & \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
M_{ij} < 0 \\
M_{ij} > 0 \\
\end{bmatrix}
\]

\[
[P] = \begin{cases} 
1, & \text{if statement } P \text{ is true;} \\
0, & \text{otherwise.}
\end{cases}
\]
Goals: Develop fast and effective software for the variants of NMF with usability and extensibility as key design features

Application to real-world large-scale data analytics problems

Implementation

- C++ codes: fast NMF based dimension reduction, hierarchical and flat linear/nonlinear clustering/topic modeling
- High level Python driver code in addition to command line interface
- Linux and Mac OS X supported. Will expand to Windows
- Currently based on Elemental: numerically robust, distributed matrix computations
- Virtual Machine (platform-agnostic) installation option: Vagrant installation based on Ubuntu minimal installation

Documentation and Tutorials

- Step-by-step procedures for installation and execution
- Test case inputs and outputs documented for comparison
Summary

- CLRA for Efficient and Effective Clustering
- Objective function level fusion possible with CLRA for utilizing content and network structure in clustering: for better clustering, link prediction, and new discoveries
- Best representations of feature-data and data-data relationships, especially for hypergraphs relationships?