Lecture 3: Nonnegative Matrix Factorization: Algorithms and Applications

Haesun Park

School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta GA, U.S.A.

SIAM Gene Golub Summer School, Aussois France,
June 18, 2019

This work was supported in part by
Outline

- Overview of NMF
- Fast algorithms with Frobenius norm
  - Theoretical results on convergence
  - Multiplicative updating
  - Alternating nonnegativity constrained least Squares: Active-set type methods, ...
  - Hierarchical alternating least squares
- Variations/Extensions of NMF: sparse NMF, regularized NMF, nonnegative PARAFAC
- Efficient adaptive NMF algorithms
- Applications of NMF, NMF for Clustering
- Computational results
- Discussions
Nonnegative Matrix Factorization (NMF)

(Lee&Seung 99, Paatero&Tapper 94)

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

- $\min_{W \geq 0, H \geq 0} \| A - WH \|_F$
- Nonconvex
- $W$ and $H$ not unique (e.g. $\hat{W} = WD \geq 0$, $\hat{H} = D^{-1}H \geq 0$)

Notation: $\mathbb{R}_{+}$: nonnegative real numbers
Nonnegative Matrix Factorization (NMF)

(Lee & Seung 99, Paatero & Tapper 94)

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

- $\min_{W \geq 0, H \geq 0} \|A - WH\|_F$
- NMF improves the approximation as $k$ increases:
  - If $\text{rank}_+(A) > k$, 
    \[
    \min_{W_{k+1} \geq 0, H_{k+1} \geq 0} \|A - W_{k+1}H_{k+1}\|_F < \min_{W_k \geq 0, H_k \geq 0} \|A - W_kH_k\|_F,
    \]

  $W_i \in \mathbb{R}_+^{m \times i}$ and $H_i \in \mathbb{R}_+^{i \times n}$
- But SVD does better: if $A = U\Sigma V^T$, then 
  \[
  \|A - U_k\Sigma_k V_k^T\|_F \leq \min \|A - WH\|_F, \quad W \in \mathbb{R}_+^{m \times k} \text{ and } H \in \mathbb{R}_+^{k \times n}
  \]
- $\text{rank}_+(A)$, is the smallest integer $k$ for which there exist $V \in \mathbb{R}_+^{m \times k}$ and $U \in \mathbb{R}_+^{k \times n}$ such that $A = VU$.
- Note: $\text{rank}(A) \leq \text{rank}_+(A) \leq \min(m, n)$
- If $\text{rank}(A) \leq 2$, then $\text{rank}_+(A) = \text{rank}(A)$.
- If either $m \in \{1, 2, 3\}$ or $n \in \{1, 2, 3\}$, then $\text{rank}_+(A) = \text{rank}(A)$. 
Nonnegative Rank of \( A \in \mathbb{R}_+^{m \times n} \) (J. Cohen and U. Rothblum, LAA, 93)

- \( rank_+(A) \), is the smallest integer \( k \) for which there exist \( V \in \mathbb{R}_+^{m \times k} \) and \( U \in \mathbb{R}_+^{k \times n} \) such that \( A = VU \).
  
  Note: \( rank(A) \leq rank_+(A) \leq \min(m, n) \)

- If \( rank(A) \leq 2 \), then \( rank_+(A) = rank(A) \).
- If either \( m \in \{1, 2, 3\} \) or \( n \in \{1, 2, 3\} \), then \( rank_+(A) = rank(A) \).

(Perron-Frobenius) There are nonnegative left and right singular vectors \( u_1 \) and \( v_1 \) of \( A \) associated with the largest singular value \( \sigma_1 \).

- \( rank \) 1 SVD of \( A \) = best rank-one NMF of \( A \)
Applications of NMF

- Text mining
  - Topic model: NMF as an alternative way for PLSI (Gaussier et al., 05; Ding et al., 08)
  - Document clustering (Xu et al., 03; Shahnaz et al., 06)
  - Topic detection and trend tracking, email analysis (Berry et al., 05; Keila et al., 05; Cao et al., 07)
- Image analysis and computer vision
  - Feature representation, sparse coding (Lee et al., 99; Guillamet et al., 01; Hoyer et al., 02; Li et al., 01)
  - Video tracking (Bucak et al., 07)
- Social network
  - Community structure and trend detection (Chi et al., 07; Wang et al., 08)
  - Recommendation system (Zhang et al., 06)
- Bioinformatics-microarray data analysis (Brunet et al., 04, H. Kim and Park, 07)
- Acoustic signal processing, blind source separating (Cichocki et al., 04)
- Financial data (Drakakis et al., 08)
- Chemometrics (Andersson and Bro, 00)
- and SO MANY MORE...
Algorithms for NMF

- Multiplicative update rules: Lee and Seung, 99
- Alternating least squares (ALS): Berry et al 06
- Alternating nonnegative least squares (ANLS)
  - Lin, 07, Projected gradient descent
  - D. Kim et al., 07, Quasi-Newton
  - H. Kim and Park, 08, Active-set
  - J. Kim and Park, 08, Block principal pivoting
- Other algorithms and variants
  - Cichocki et al., 07, Hierarchical ALS (HALS)
  - Ho, 08, Rank-one Residue Iteration (RRI)
  - Zdunek, Cichocki, Amari 06, Quasi-Newton
  - Chu and Lin, 07, Low dim polytope approx.
  - Other rank-1 downdating based algorithms (Vavasis,..)
  - C. Ding, T. Li, tri-factor NMF, orthogonal NMF, ...
  - Cichocki, Zdunek, Phan, Amari: NMF and NTF: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, Wiley, 09
  - Andersson and Bro, Nonnegative Tensor Factorization, 00
  - And SO MANY MORE...
A constrained nonlinear problem:

\[
\min f(x) \text{ (e.g., } f(W, H) = \|A - WH\|_F) \\
\text{subject to } x \in X = X_1 \times X_2 \times \cdots \times X_p,
\]

where \( x = (x_1, x_2, \ldots, x_p) \), \( x_i \in X_i \subset \mathbb{R}^{n_i} \), \( i = 1, \ldots, p \).

**Block Coordinate Descent method** generates

\[
x^{(k+1)} = (x_1^{(k+1)}, \ldots, x_p^{(k+1)}) \text{ by }
\]

\[
x_i^{(k+1)} = \arg \min_{\xi \in X_i} f(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, \xi, x_{i+1}^{(k)}, \ldots, x_p^{(k)}).
\]
Block Coordinate Descent (BCD) Method

- A constrained nonlinear problem:
  \[
  \min f(x)(\text{e.g., } f(W, H) = \|A - WH\|_F)
  \]
  subject to \( x \in X = X_1 \times X_2 \times \cdots \times X_p \),
  where \( x = (x_1, x_2, \ldots, x_p), x_i \in X_i \subset \mathbb{R}^{n_i}, i = 1, \ldots, p. \)

- **Block Coordinate Descent method** generates
  \( x^{(k+1)} = (x_1^{(k+1)}, \ldots, x_p^{(k+1)}) \) by
  \[
  x_i^{(k+1)} = \arg\min_{\xi \in X_i} f(x_1^{(k+1)}, \ldots, x_i^{(k+1)}-1, \xi, x_i^{(k)}+1, \ldots, x_p^{(k)}).
  \]

- **Th. (Bertsekas, 99):** Suppose \( f \) is continuously differentiable over the Cartesian product of closed, convex sets \( X_1, X_2, \ldots, X_p \) and suppose for each \( i \) and \( x \in X \), the minimum for
  \[
  \min_{\xi \in X_i} f(x_1^{(k+1)}, \ldots, x_i^{(k+1)}-1, \xi, x_i^{(k)}+1, \ldots, x_p^{(k)})
  \]
  is **uniquely** attained. Then every limit point of the sequence generated by the BCD method \( \{x^{(k)}\} \) is a stationary point.

  **NOTE:** Uniqueness not required when \( p = 2 \) (Grippo and Sciandrone, 00).
Minimize functions of $w_{ij}$ or $h_{ij}$ while all other components in $W$ and $H$ are fixed:

\[
\begin{align*}
    w_{ij} &\leftarrow \arg \min_{w_{ij} \geq 0} \| (r_i^T - \sum_{k \neq j} w_{ik} h_k^T ) - w_{ij} h_j^T \|_2 \\
    h_{ij} &\leftarrow \arg \min_{h_{ij} \geq 0} \| (a_j - \sum_{k \neq i} w_{kj} h_k) - w_i h_{ij} \|_2
\end{align*}
\]

where $W = \begin{pmatrix} w_1 & \cdots & w_k \end{pmatrix}$, $H = \begin{pmatrix} h_1^T \\ \vdots \\ h_k^T \end{pmatrix}$ and $A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} = \begin{pmatrix} r_1^T \\ \vdots \\ r_m^T \end{pmatrix}$

- Scalar quadratic function, closed form solution.
Lee and Seung (01)’s multiplicative updating (MU) rule

\[ w_{ij} \leftarrow w_{ij} \frac{(AH^T)_{ij}}{(WHH^T)_{ij}}, \quad h_{ij} \leftarrow h_{ij} \frac{(W^T A)_{ij}}{(W^T WH)_{ij}} \]

Derivation based on gradient-descent form:

\[ w_{ij} \leftarrow w_{ij} + \frac{w_{ij}}{(WHH^T)_{ij}} \left[ (AH^T)_{ij} - (WHH^T)_{ij} \right] \]
\[ h_{ij} \leftarrow h_{ij} + \frac{h_{ij}}{(W^T WH)_{ij}} \left[ (W^T A)_{ij} - (W^T WH)_{ij} \right] \]

Rewriting of the solution of coordinate descent:

\[ w_{ij} \leftarrow \left[ w_{ij} + \frac{1}{(HH^T)_{jj}} \left( (AH^T)_{ij} - (WHH^T)_{ij} \right) \right]_+ \]
\[ h_{ij} \leftarrow \left[ h_{ij} + \frac{1}{(W^T W)_{ii}} \left( (W^T A)_{ij} - (W^T WH)_{ij} \right) \right]_+ \]

In MU, conservative steps are taken to ensure nonnegativity. Bertsekas’ Th. on convergence is not applicable to MU.
Minimize functions of $w_i$ or $h_i$ while all other components in $W$ and $H$ are fixed:

$$\|A - \sum_{j=1}^{k} w_j h_j^T\|_F = \|(A - \sum_{j=1}^{k} w_j h_j^T) - w_i h_i^T\|_F = \|R^{(i)} - w_i h_i^T\|_F$$

$$w_i \leftarrow \arg\min_{w_i \geq 0} \|R^{(i)} - w_i h_i^T\|_F$$

$$h_i \leftarrow \arg\min_{h_i \geq 0} \|R^{(i)} - w_i h_i^T\|_F$$

Each subproblem has the form $\min_{x \geq 0} \|c x^T - G\|_F$ and has a closed form solution $x = \left[ \frac{G^T c}{c^T c} \right]_+$.

Hierarchical Alternating Least Squares (HALS) (Cichocki et al, 07, 09),

(actually HA-NLS)

Rank-one Residue Iteration (RRI) (Ho, 08)
In scalar BCD, $w_{1j}, w_{2j}, \cdots, w_{mj}$ can be computed independently. Also, $h_{i1}, h_{i2}, \cdots, h_{in}$ can be computed independently.

→ scalar BCD $\iff$ $2k$ vector BCD in NMF
Successive rank-1 deflation works for SVD but not for NMF

\[ A - \sigma_1 u_1 v_1^T \approx \sigma_2 u_2 v_2^T \quad A - w_1 h_1^T \approx w_2 h_2^T? \]

\[
\begin{pmatrix}
4 & 6 & 0 \\
6 & 4 & 0 \\
0 & 0 & 1
\end{pmatrix}
\approx
\begin{pmatrix}
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
10 & 0 & 0 \\
0 & 2 & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

The sum of two successive best rank-1 nonnegative approx. is

\[
\begin{pmatrix}
4 & 6 & 0 \\
6 & 4 & 0 \\
0 & 0 & 1
\end{pmatrix}
\approx
\begin{pmatrix}
5 & 5 & 0 \\
5 & 5 & 0 \\
0 & 0 & 0
\end{pmatrix}
+ \begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

The best rank-2 nonnegative approx. is

\[ WH = \begin{pmatrix}
4 & 6 & 0 \\
6 & 4 & 0 \\
0 & 0 & 0
\end{pmatrix}
= \begin{pmatrix}
4 & 6 \\
6 & 4 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0
\end{pmatrix}
\]

NOTE: 2k vector BCD ≠ successive rank-1 deflation for NMF
Minimize functions of $W$ or $H$ while the other is fixed:

$$W \leftarrow \arg\min_{W \geq 0} \| H^T W^T - A^T \|_F$$

$$H \leftarrow \arg\min_{H \geq 0} \| W H - A \|_F$$

**Alternating Nonnegativity-constrained Least Squares (ANLS)**

- **No closed form solution.**
  - Projected gradient method (Lin, 07)
  - Projected quasi-Newton method (D. Kim et al., 07)
  - Active-set method (H. Kim and Park, 08)
  - Block principal pivoting method (J. Kim and Park, 08)

**ALS** (M. Berry et al. 06) ??
Nonnegativity-constrained Least Squares (NLS) problem

- **Projected Gradient method** (Lin, 07) \( x^{(k+1)} \leftarrow \mathcal{P}_+(x^{(k)} - \alpha_k \nabla f(x^{(k)})) \)
  * \( \mathcal{P}_+(\cdot) \): Projection operator to the nonnegative orthant
  * Back-tracking selection of step \( \alpha_k \)

- **Projected Quasi-Newton method** (Kim et al., 07)
  \[ x^{(k+1)} \leftarrow \begin{bmatrix} y \\ z_k \end{bmatrix} = \mathcal{P}_+ \begin{bmatrix} y^{(k)} - \alpha D^{(k)} \nabla f(y^{(k)}) \\ 0 \end{bmatrix} \]
  * Gradient scaling only for nonzero variables

- These do not fully exploit the structures of the NLS problems in NMF

- **Active Set method** (H. Kim and Park, 08)
  Lawson and Hanson (74), Bro and De Jong (97), Van Benthem and Keenan (04)

- **Block principal pivoting method** (J. Kim and Park, 08)
  linear complementarity problems (LCP) (Judice and Pires, 94)
Active-set type Algorithms for\nn\n\[\min_{x \geq 0} \|Cx - b\|_2, \ C : m \times k\]

• KKT conditions: \( y = C^T Cx - C^T b \)
  \( y \geq 0, \quad x \geq 0, \quad x_i y_i = 0, \ i = 1, \cdots, k \)

• If we know \( P = \{i|x_i > 0\} \) in the solution in advance
  then we only need to solve \( \min \|C_P x_P - b\|_2 \), and the rest of
  \( x_i = 0 \), where \( C_P \): columns of \( C \) with the indices in \( P \)

\[
\begin{array}{c|c|c|c}
C & \ast & x & \approx \\
\hline
+ & + & 0 & 0 \\
\hline
+ & + & + & + \\
\hline
\end{array}
\]

\( b \)
Active-set type Algorithms for

\[
\min_{x \geq 0} \|Cx - b\|_2, \ C : m \times k
\]

- **KKT conditions:**
  \[y = C^T Cx - C^T b\]
  \[y \geq 0, \ x \geq 0, \ x_i y_i = 0, \ i = 1, \ldots, k\]

- **Active set method** (Lawson and Hanson 74)
  - \(E = \{1, \ldots, k\}\) (i.e. \(x = 0\) initially), \(P = \text{null}\)
  - Repeat while \(E\) not null and \(y_i < 0\) for some \(i\)
    - Exchange indices between \(E\) and \(P\) while keeping feasibility and reducing the objective function value

- **Block Principal Pivoting method** (Portugal et al. 94 MathComp):
  - Lacks any monotonicity or feasibility but finds a correct active-passive set partitioning.
  - Guess two index sets \(P\) and \(E\) that partition \(\{1, \ldots, k\}\)
  - Repeat
    - Let \(x_E = 0\) and \(x_P = \arg\min_{x_P} \|C_P x_P - b\|_2^2\)
    - Then \(y_E = C_E^T (C_P x_P - b)\) and \(y_P = 0\)
    - If \(x_P \geq 0\) and \(y_E \geq 0\), then optimal values are found.
      Otherwise, update \(P\) and \(E\).
How block principal pivoting works

\( k = 10, \text{Initially } P = \{1, 2, 3, 4, 5\}, \ E = \{6, 7, 8, 9, 10\} \)

Update by \( C_P^T C_P x_P = C_P^T b \), and \( y_E = C_E^T (C_P x_P - b) \)
How block principal pivoting works

Update by $C_P^T C_P x_P = C_P^T b$, and $y_E = C_E^T (C_P x_P - b)$

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>P</th>
<th>P</th>
<th>P</th>
<th>P</th>
<th>E</th>
<th>E</th>
<th>E</th>
<th>E</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
How block principal pivoting works

Update by $C_P^T C_P x_P = C_P^T b$, and $y_E = C_E^T (C_P x_P - b)$
How block principal pivoting works

Update by $C_P^T C_P x_P = C_P^T b$, and $y_E = C_E^T (C_P x_P - b)$

<table>
<thead>
<tr>
<th></th>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>+</td>
</tr>
</tbody>
</table>
How block principal pivoting works

Update by $C_P^T C_P x_P = C_P^T b$, and $y_E = C_E^T (C_P x_P - b)$
How block principal pivoting works

Update by $C_P^T C_P x_P = C_P^T b$, and $y_E = C_E^T (C_P x_P - b)$
How block principal pivoting works

Update by $C_P^T C_P x_P = C_P^T b$, and $y_E = C_E^T (C_P x_P - b)$
Active set algorithm is a special instance of single principal pivoting algorithm (H. Kim and Park, SIMAX 08)

Block exchange rule without modification does not always work.

- The residual is not guaranteed to monotonically decrease. Block exchange rule may cycle (although rarely).
- Modification: if the block exchange rule fails to decrease the number of infeasible variables, use a backup exchange rule
- With this modification, block principal pivoting algorithm finds the solution of NLS in a finite number of iterations.
Structure of NLS problems in NMF

- Matrix is long and thin, solutions vectors short, many right hand side vectors.
- \[ \min_{H \geq 0} \| WH - A \|_F^2 \]

- \[ \min_{W \geq 0} \| H^T W^T - A^T \|_F^2 \]
Efficient Algorithm for $\min_{X \geq 0} \| CX - B \|^2_F$

[Bro and de Jong, 97, Van Benthem and Keenan, 04]

- Precompute $C^T C$ and $C^T B$
  - Update $x_P$ and $y_E$ by $C_P^T C_P x_P = C_P^T b$ and $y_E = C_E^T C_P x_P - C_E^T b$
  - All coefficients can be retrieved from $C^T C$ and $C^T B$

- $C^T C$ and $C^T B$ is small. Storage is not a problem.

- Exploit common $P$ and $E$ sets among col. in $B$ in each iteration.
  - $X$ is flat and wide. $\rightarrow$ More common cases of $P$ and $E$ sets.

- Proposed algorithm for NMF (ANLS/BPP):
  - ANLS framework $+$ Block principal pivoting algorithm for NLS with improvements for multiple right-hand sides

(J. Kim and H. Park, ICDM 08)
What happens when $C$ is rank deficient?

- Active set method: if the first matrix with passive set is of full rank, the method never runs into rank deficient subproblems
  [Drake et.al., Info Fusion 10]

- Block principal pivoting: subproblems may become rank deficient
Sparse NMF (for sparse $H$) (H. Kim and Park, Bioinformatics, 07)

$$\min_{W,H} \left\{ \|A - WH\|_F^2 + \eta \|W\|_F^2 + \beta \sum_{j=1}^{n} \|H(:,j)\|_1^2 \right\}, \forall ij, W_{ij}, H_{ij} \geq 0$$

ANLS reformulation (H. Kim and Park, 07) : alternate the following

$$\min_{H \geq 0} \left\| \left( \frac{W}{\sqrt{\beta} e_1 \times k} \right) H - \left( \begin{array}{c} A \\ 0_{1 \times n} \end{array} \right) \right\|_F^2$$

$$\min_{W \geq 0} \left\| \left( \frac{H^T}{\sqrt{\eta} I_k} \right) W^T - \left( \begin{array}{c} A^T \\ 0_{k \times m} \end{array} \right) \right\|_F^2$$

Regularized NMF (Pauca, et al. 06) :

$$\min_{W,H} \left\{ \|A - WH\|_F^2 + \eta \|W\|_F^2 + \beta \|H\|_F^2 \right\}, \forall ij, W_{ij}, H_{ij} \geq 0.$$

ANLS reformulation : alternate the following

$$\min_{H \geq 0} \left\| \left( \frac{W}{\sqrt{\beta} I_k} \right) H - \left( \begin{array}{c} A \\ 0_{k \times n} \end{array} \right) \right\|_F^2$$

$$\min_{W \geq 0} \left\| \left( \frac{H^T}{\sqrt{\eta} I_k} \right) W^T - \left( \begin{array}{c} A^T \\ 0_{k \times m} \end{array} \right) \right\|_F^2$$
Consider a 3-way Nonnegative Tensor $\mathbf{T} \in \mathbb{R}^{m \times n \times p}_{+}$ and its PARAFAC
\[ \min_{A,B,C \geq 0} \| \mathbf{T} - [ABC] \|_F^2 \]
where $A \in \mathbb{R}^{m \times k}_{+}$, $B \in \mathbb{R}^{n \times k}_{+}$, $C \in \mathbb{R}^{p \times k}_{+}$.

The loading matrices ($A,B,$ and $C$) can be iteratively estimated by an NLS algorithm such as block principal pivoting method.
Iterate until a stopping criteria is satisfied:

- \( \min_{A \geq 0} \| Y_{BC} A^T - T(1) \|_F \)
- \( \min_{B \geq 0} \| Y_{AC} B^T - T(2) \|_F \)
- \( \min_{C \geq 0} \| Y_{AB} C^T - T(3) \|_F \) where

\[
Y_{BC} = B \odot C \in \mathbb{R}^{(np) \times k}, \quad T(1) \in \mathbb{R}^{(np) \times m},
\]
\[
Y_{AC} = A \odot C \in \mathbb{R}^{(mp) \times k}, \quad T(2) \in \mathbb{R}^{(mp) \times n},
\]
\[
Y_{AB} = A \odot B \in \mathbb{R}^{(mn) \times k}, \quad T(3) \in \mathbb{R}^{(mn) \times p}
\]

unfolded matrices, and \( F \odot G_{(mn) \times k} = [f_1 \otimes g_1 \quad f_2 \otimes g_2 \quad \cdots \quad f_k \otimes g_k] \) is the Khatri-Rao product of \( F \in \mathbb{R}^{m \times k} \) and \( G \in \mathbb{R}^{n \times k} \).

Matrices are longer and thinner, ideal for ANLS/BPP.
Can be similarly extended to higher order tensors.
# Experimental Results (NMF)

## NMF Algorithms Compared

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANLS-BPP</td>
<td>ANLS / block principal pivoting</td>
<td>J. Kim and HP 08</td>
</tr>
<tr>
<td>ANLS-AS</td>
<td>ANLS / active set</td>
<td>H. Kim and HP 08</td>
</tr>
<tr>
<td>ANLS-PGRAD</td>
<td>ANLS / projected gradient</td>
<td>Lin 07</td>
</tr>
<tr>
<td>ANLS-PQN</td>
<td>ANLS / projected quasi-Newton</td>
<td>D. Kim et al. 07</td>
</tr>
<tr>
<td>HALS</td>
<td>Hierarchical ALS</td>
<td>Cichocki et al. 07</td>
</tr>
<tr>
<td>MU</td>
<td>Multiplicative updating</td>
<td>Lee and Seung 01</td>
</tr>
<tr>
<td>ALS</td>
<td>Alternating least squares</td>
<td>Berry et al. 06</td>
</tr>
</tbody>
</table>
Residual vs. Execution time

TDT2, k=10

TDT2, k=160

TDT2 text data: $19,009 \times 3,087$, $k = 10$ and $k = 160$
Residual vs. Execution time

ATNT, $k=10$

20 Newsgroups, $k=160$

ATNT image data: $10, 304 \times 400, \ k = 10$ and
20 Newsgroups text data: $26, 214 \times 11, 314, \ k = 160$
Residual vs. Execution time

PIE 64, k=80

PIE 64, k=160

PIE 64 image data: 4,096 \times 11,554, k = 80 and k = 160
Given \((W, H)\) with \(k\), how to compute \((\tilde{W}, \tilde{H})\) with \(\tilde{k}\) fast?

E.g., model selection for NMF clustering

**AdaNMF**

- Initialize \(\tilde{W}\) and \(\tilde{H}\) using \(W\) and \(H\)
  - If \(\tilde{k} > k\), compute NMF for \(A - WH \approx \Delta W \Delta H\). Set \(\tilde{W} = [W \Delta W]\) and \(\tilde{H} = [H; \Delta H]\)
  - If \(\tilde{k} < k\), initialize \(\tilde{W}\) and \(\tilde{H}\) with \(\tilde{k}\) pairs of \((w_i, h_i)\) with largest \(\|w_i h_i^T\|_F = \|w_i\|_2 \|h_i\|_2\)
- Update \(\tilde{W}\) and \(\tilde{H}\) using HALS algorithm.
Consensus matrix based on $A \approx WH$:

$$C^t_{ij} = \begin{cases} 
0 & \max(H(:, i)) = \max(H(:, j)), \\
1 & \max(H(:, i)) \neq \max(H(:, j)), 
\end{cases} , \ t = 1, \ldots, l$$

Dispersion coefficient $\rho(k) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 4(C_{ij} - \frac{1}{2})^2$, where $C = \frac{1}{l} \sum C^t$

Clustering results on MNIST digit images ($784 \times 2000$) by AdaNMF with $k = 3, 4, 5$ and 6. Averaged consensus matrices, dispersion coefficient, execution time
Relative error vs. exec. time of AdaNMF and “recompute”. Given an NMF of $600 \times 600$ synthetic matrix with $k = 60$, compute NMF with $\tilde{k} = 50, 80$. 
Theorem: For $A \in \mathbb{R}^{m \times n}_+$, if $\text{rank}_+(A) > k$, then
$$\min \| A - W^{(k+1)} H^{(k+1)} \|_F < \min \| A - W^{(k)} H^{(k)} \|_F,$$
where $W^{(i)} \in \mathbb{R}^{m \times i}_+$ and $H^{(i)} \in \mathbb{R}^{i \times n}_+$.

Rank path on synthetic data set: relative residual vs. $k$
ORL Face image (10304 × 400) classification errors (by LMNN) on training and testing set vs. $k$.

$k$-dim rep. $H_T$ of training data $T$ by BPP $\min_{H_T \geq 0} \| W H_T - T \|_F$
Given an NMF \((W, H)\) for \(A = [\delta A \  \hat{A}]\), how to compute NMF \((\tilde{W}, \tilde{H})\) for \(\tilde{A} = [\hat{A} \ \Delta A]\) fast? (Updating and Downdating)

**DynNMF** (Sliding Window NMF)

- Initialize \(\tilde{H}\) as follows:
  - Let \(\hat{H}\) be the remaining columns of \(H\).
  - Solve \(\min_{\Delta H \geq 0} \|W \Delta H - \Delta A\|^2_F\) using block principal pivoting
  - Set \(\tilde{H} = [\hat{H} \ \Delta H]\)
- Run HALS on \(\tilde{A}\) with initial factors \(\tilde{W} = W\) and \(\tilde{H}\)
PET2001 data with 3064 images from a surveillance video. DynNMF on 110, 592 × 400 data matrix each time, with 100 new columns and 100 obsolete columns. The residual images track the moving vehicle in the video.
Find $k$ clusters (topics) in a text corpus represented in $X$

$$x_i \approx w_1 h_{1i} + w_2 h_{2i} + w_3 h_{3i}$$

- Nonnegative $w_1, w_2, w_3$: cluster representatives or topics: dist. of keywords
- Nonnegative $h_{1i}, h_{2i}, h_{3i}$ – soft clustering assignment of $x_i$
- In NMF, $w_i$’s have equal roles among them unlike in SVD. Successive rank-1 deflation does not work in NMF.
Clustering and Lower Rank Approximation are related. NMF for Clustering: Document (Xu et al. SIGIR 03), Image (Cai et al. ICDM 08), Microarray (Kim & Park, Bio 07), etc.

Objective functions for K-means and NMF may look the same:
\[ \sum_i \| x_i - w_{\sigma_i} \|_2^2 = \| X - WH \|_F^2 \]  
(Ding et al. SDM 05; Kim & Park, TR 08)

\[ \sigma_i = j \text{ when } i\text{-th point is assigned to } j\text{-th cluster (} j \in \{1, \ldots, k\} \text{). However, constraints are different:} \]

- **K-means:** $H \in \{0, 1\}^{k \times n}$, $1_k^T H = 1_n^T$
- **NMF:** $W \geq 0$, $H \geq 0$

Paths to solution:
- **K-means:** Expectation-Maximization
- **NMF:** Relax the condition on $H$ to $H \geq 0$ with orthogonal rows or $H \geq 0$ with sparse columns - soft clustering
**NMF vs K-means**

**K-means:** $W$: $k$ cluster centroids, $h_i$: cluster membership indicator

**NMF:** $W$: basis vectors for rank-$k$ approx., $H$: $k$-dim rep. of $X$

**Sparse NMF (SNMF)** (H. Kim & H. Park, Bioinformatics, 07)

Clustering accuracy on TDT2 text data: (aver. among 100 runs)

<table>
<thead>
<tr>
<th># clusters</th>
<th>2</th>
<th>6</th>
<th>10</th>
<th>14</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>0.8099</td>
<td>0.7295</td>
<td>0.7015</td>
<td>0.6675</td>
<td>0.6675</td>
</tr>
<tr>
<td>NMF/ANLS</td>
<td>0.9990</td>
<td>0.8717</td>
<td>0.7436</td>
<td>0.7021</td>
<td>0.7160</td>
</tr>
<tr>
<td>SNMF/ANLS</td>
<td>0.9991</td>
<td>0.8770</td>
<td>0.7512</td>
<td>0.7269</td>
<td>0.7278</td>
</tr>
</tbody>
</table>

Sparsity constraint improves clustering result (J. Kim and Park, 08):

$$\min_{W \geq 0, H \geq 0} \| A - WH \|_F^2 + \eta \| W \|_F^2 + \beta \sum_{j=1}^{n} \| H(:, j) \|_1^2$$

# of times achieving optimal assignment

(a synthetic data set, with a clear cluster structure):

<table>
<thead>
<tr>
<th>$k$</th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>12</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF</td>
<td>69</td>
<td>65</td>
<td>74</td>
<td>68</td>
<td>44</td>
</tr>
<tr>
<td>SNMF</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>97</td>
</tr>
</tbody>
</table>

NMF and SNMF much better than k-means in general.
Equivalence of objective functions is not enough to explain the clustering capability of NMF:

- Ex. $k = 2$

  K-means

  NMF

- NMF is more related to spherical k-means, than to k-means
  \[\rightarrow\] NMF shown to work well in text data clustering
Symmetric NMF and Spectral Clustering

Symmetric NMF: \( \min_{S \geq 0} \| A - SS^T \|_F \), \( A \in \mathbb{R}^{n \times n}_{+} \): affinity matrix

- Spectral clustering \( \rightarrow \) Eigenvectors (Ng et al. 01), A normalized if needed, Laplacian,...
- Symmetric NMF (Ding et al.) \( \rightarrow \) can handle nonlinear structure, and \( S \geq 0 \) naturally captures a cluster structure in \( S \) (a new PGD based Alg.)

Eigenvalues of similarity matrix:

Ex. Well separated, one loose and the other tight clusters, with
\( \exp(-\|x_i - x_j\|^2/2\sigma^2) \)

Spectral clustering:

Eigenvectors vs \( S \) from SymNMF:
Large text collections such as Wikipedia and Twitter data contain many topics of very wide range

Only a subset of data items is related to a specific question and interest, e.g. documents related to particular event, subject such as sustainability, brand, or product

Direct keyword match or keyword filtering does not work well

Want high recall than precision: retrieving as many relevant documents vs. a small number of most relevant documents

How to find the relevant parts of the documents?
Iterative application of NMF with topic refinement each step

1. Apply initial NMF to find $k$ clusters.
2. Check top keywords of each cluster and determine relevant topics. Assume first $k_1$ topics are relevant and denote the first $k_1$ columns of matrix $W$ as $W_1$.
3. Refine $W_1$ by setting its small components to zero.
4. Solve the following NMF
   \[
   \min_{W_2 \geq 0, H \geq 0} f(W_2, H) = \| (W_1 \ W_2) H - X \|_F
   \]
5. Find more relevant topics and attach their corresponding columns to $W_1$.
6. Repeat from Step 3 until getting satisfactory results.
7. Delete unrelated topics and docs, and redo topic modeling.

Advantage: easy to interpret the result and adjust intermediate results for interactive computing
Targeted Topic Modeling for Event Detection in Document Data

Application Flow

- LEAN and SmallK to perform document clustering
- Web application to leverage SME insights
- Iterative topic refinement, by removing unwanted clusters

Web Application

- Translate topics, key documents, and keywords
- Show stemmed terms and full terms for analysis

Diagram:

1. LEAN (a) → A
2. SmallK (b) → W
3. Post processing (b) → H
4. SmallK (b) → A’
5. Remove key documents (a)
6. Noisy? (Yes → Remove key documents (a), No → End)
Targeted Topic Modeling: Ex. Yemen Ceasefire Violation Detection in Arabic Text

- Interface for selecting clusters to remove from the dataset
- Key documents identified within the clusters (using a threshold parameter)
- Reapplication of SmallK to produce refined clustering results
Case Study for Sustainability in Twitter: Keyword Expansion and Retrieval of Relevant Data Items

Direct match on Twitter with 50 specialized keywords related to sustainability such as new urbanism, rain water harvesting, green roof, walkable community, and decentralized energy, retrieved only a very small number of tweets.

- Discover the casual terms that are used to express sustainability-related topics in social networks in addition to specialized keywords
- Extract the data items that have relevance to sustainability. These data items may not contain technical or specialized keywords.
Applying targeted topic modeling to the 60K Wikipedia pages obtained by keyword refinement, we obtained 5,387 pages that are clearly relevant to sustainability. Applying HierNMF2, we discovered 20 topics such as solar energy, water energy, efficient electricity usage, sustainable development, renewable energy, and sustainable manufacturing. A subset of the results is illustrated:
DARPA XDATA Challenge Problem: Tracking IED Drop
Visualizing our analysis results

Visualization of our analytics fused with GPS and communications data can “guide” an analyst to important “events” and “actors”

- Track messages of common topics occurring within a time window
- Find areas of interest when perpetrators within the same cluster often converge
- Flag persons of interest by detecting messages that occur within a topic of interest
Associate device movement with classes of communication activity

### Topic Modeling Results (NMF)

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
<th>Topic 7</th>
<th>Topic 8</th>
<th>Topic 9</th>
<th>Topic 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>like</td>
<td>picture</td>
<td>get</td>
<td>think</td>
<td>just</td>
<td>good</td>
<td>line</td>
<td>creative</td>
<td>text</td>
<td>video</td>
</tr>
<tr>
<td>really</td>
<td>took</td>
<td>did</td>
<td>know</td>
<td>want</td>
<td>answer</td>
<td>green</td>
<td>full</td>
<td>terrorist</td>
<td>took</td>
</tr>
<tr>
<td>better</td>
<td>ied</td>
<td>time</td>
<td>haha</td>
<td>say</td>
<td>sing</td>
<td>red</td>
<td>text</td>
<td>pic</td>
<td>sent</td>
</tr>
<tr>
<td>sounds</td>
<td>sending</td>
<td>back</td>
<td>getting</td>
<td>got</td>
<td>hey</td>
<td>jabal</td>
<td>little</td>
<td>call</td>
<td>snd</td>
</tr>
<tr>
<td>say</td>
<td>script</td>
<td>need</td>
<td>got</td>
<td>see</td>
<td>did</td>
<td>medina</td>
<td>data</td>
<td>sending</td>
<td>sim</td>
</tr>
<tr>
<td>music</td>
<td>site</td>
<td>call</td>
<td>better</td>
<td>guess</td>
<td>new</td>
<td>ghazi</td>
<td>collection</td>
<td>script</td>
<td>wait</td>
</tr>
<tr>
<td>feel</td>
<td>hole</td>
<td>see</td>
<td>car</td>
<td>send</td>
<td>morning</td>
<td>amber</td>
<td>did</td>
<td>back</td>
<td>package</td>
</tr>
<tr>
<td>sure</td>
<td>denver?</td>
<td>thanks</td>
<td>ready</td>
<td>wait</td>
<td>years</td>
<td>send</td>
<td>long</td>
<td>sent</td>
<td>checkpoint</td>
</tr>
<tr>
<td>done</td>
<td>taken</td>
<td>days</td>
<td>dont</td>
<td>day</td>
<td>name</td>
<td>vehicles</td>
<td>step</td>
<td>ied</td>
<td>sending</td>
</tr>
<tr>
<td>did</td>
<td>moving</td>
<td>able</td>
<td>need</td>
<td>life</td>
<td>music</td>
<td>phone</td>
<td>house</td>
<td>site</td>
<td>area</td>
</tr>
<tr>
<td>people</td>
<td>drop</td>
<td>contact</td>
<td>want</td>
<td>going</td>
<td>man</td>
<td>time</td>
<td>thing</td>
<td>please</td>
<td>convoy</td>
</tr>
<tr>
<td>gaga</td>
<td>frame</td>
<td>meeting</td>
<td>lol</td>
<td>make</td>
<td>tell</td>
<td>life</td>
<td>maybe</td>
<td>meet</td>
<td>send</td>
</tr>
<tr>
<td>haha</td>
<td>police</td>
<td>report</td>
<td>really</td>
<td>lot</td>
<td>songs</td>
<td>chapter</td>
<td>addicted</td>
<td>phone</td>
<td>mcc</td>
</tr>
<tr>
<td>paul</td>
<td>province</td>
<td>ready</td>
<td>days</td>
<td>time</td>
<td>hear</td>
<td>guess</td>
<td>gonna</td>
<td>day</td>
<td>king</td>
</tr>
<tr>
<td>does</td>
<td>end</td>
<td>sir</td>
<td>start</td>
<td>things</td>
<td>job</td>
<td>blue</td>
<td>sons</td>
<td>short</td>
<td>got</td>
</tr>
</tbody>
</table>

**Topic 2, 9, 10:**

The majority of the conversations were about planning, meeting, observing, and exchanging.

**Topic 3:**

Suspicious meetings mixed with civilian chatter

**Topic 7:**

Interesting. Could be code for organizing the simulation?

**Other Topics:**

Separate out civilian conversations

**Pictures taken for planning and observing IED bomb site and for the aftermath.**

**Videos taken of checkpoints and convoy areas.**

**Package exchange and drop off confirmations.**
1. Package dropped
Scene 2 of 4

2. Left area
Scene 3 of 4

3. Nice fireworks
4. Leader identified
### Summary/Discussions

- Overview of NMF with Frobenius norm and algorithms
- Fast algorithms and convergence via BCD framework
- Adaptive NMF algorithms
- Variations/Extensions of NMF: nonnegative PARAFAC and sparse NMF
- NMF for clustering
- Computational comparisons

NMF Matlab codes and papers available at http://www.cc.gatech.edu/~hpark and