Nonnegative Matrix Factorization: Algorithms and Applications

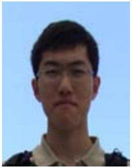
Haesun Park
hpark@cc.gatech.edu

School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, GA, USA

SIAM International Conference on Data Mining, April, 2011

This work was supported in part by the National Science Foundation.

Co-authors

	Jingu Kim	CSE, Georgia Tech
	Yunlong He	Math, Georgia Tech
	Da Kuang	CSE, Georgia Tech

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
- Extensive 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 $k << min(m, n)$, 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 $rank_+(A) > k$,

$$\min_{W_{k+1} \geq 0, H_{k+1} \geq 0} \|A - W_{k+1}H_{k+1}\|_F < \min_{W_k \geq 0, H_k \geq 0} \|A - W_k H_k\|_F,$$

  $W_i \in {\mathbb{R}_+}^{m \times i}$ and $H_i \in {\mathbb{R}_+}^{i \times n}$
- But SVD does better: if $A = U \Sigma V^T$, then $\|A - U_k \Sigma_k V_k^T\|_F \leq min\|A - WH\|_F$, $W \in {\mathbb{R}_+}^{m \times k}$ and $H \in {\mathbb{R}_+}^{k \times n}$
- So Why NMF? Dimension Reduction with Better Interpretation/Lower Dim. Representation for Nonnegative Data.

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 $rank_+(A) = rank(A)$.

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 σ_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...

Block Coordinate Descent (BCD) Method

- A constrained nonlinear problem:

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

  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 $x^{(k+1)} = (x_1^{(k+1)}, \ldots, x_p^{(k+1)})$ 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:

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

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

$$x_i^{(k+1)} = arg \min_{\xi \in X_i} f(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, \xi, x_{i+1}^{(k)}, \ldots, x_p^{(k)}).$$

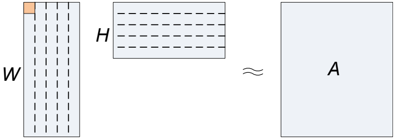
- **Th. (Bertsekas, 99):** Suppose f is continuously differentiable over the Cartesian product of closed, convex sets $X_1, X_2, \ldots, X_p$ and suppose for each i and $x \in X$, the **minimum** for

$$\min_{\xi \in X_i} f(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, \xi, x_{i+1}^{(k)}, \ldots, x_p^{(k)})$$

  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).

BCD with $k(m+n)$ Scalar Blocks



- Minimize functions of w_{ij} or h_{ij} while all other components in W and H are fixed:

$$
\begin{aligned}
w_{ij} &\leftarrow \operatorname*{arg\,min}_{w_{ij}\geq 0} \|(r_i^T - \sum_{k\neq j} w_{ik} h_k^T) - w_{ij} h_j^T\|_2 \\
h_{ij} &\leftarrow \operatorname*{arg\,min}_{h_{ij}\geq 0} \|(a_j - \sum_{k\neq i} w_k h_{kj}) - w_i h_{ij}\|_2
\end{aligned}
$$

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.

BCD with $k(m+n)$ Scalar Blocks

- Lee and Seung (01)'s multiplicative updating (MU) rule

$$w_{ij} \leftarrow w_{ij} \frac{(AH^T)_{ij}}{(WHH^T)_{ij}} \quad , \quad h_{ij} \leftarrow h_{ij} \frac{(W^TA)_{ij}}{(W^TWH)_{ij}}$$

- Derivation based on gradient-descent form:

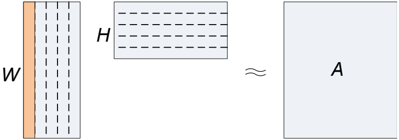
$$\begin{aligned} w_{ij} &\leftarrow & w_{ij} + \frac{w_{ij}}{(WHH^T)_{ij}} \left[(AH^T)_{ij} - (WHH^T)_{ij}\right] \\ h_{ij} &\leftarrow & h_{ij} + \frac{h_{ij}}{(W^TWH)_{ij}} \left[(W^TA)_{ij} - (W^TWH)_{ij}\right] \end{aligned}$$

- Rewriting of the solution of coordinate descent:

$$\begin{aligned} w_{ij} &\leftarrow & \left[w_{ij} + \frac{1}{(HH^T)_{jj}} \left((AH^T)_{ij} - (WHH^T)_{ij}\right)\right]_+ \\ h_{ij} &\leftarrow & \left[h_{ij} + \frac{1}{(W^TW)_{ii}} \left((W^TA)_{ij} - (W^TWH)_{ij}\right)\right]_+ \end{aligned}$$

- In MU, conservative steps are taken to ensure nonnegativity. Bertsekas' Th. on convergence is not applicable to MU.

BCD with $2k$ Vector Blocks

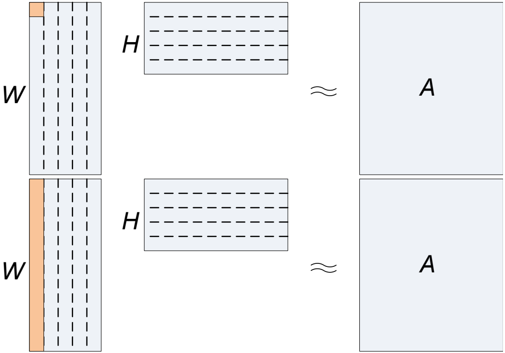


- Minimize functions of w_i or h_i while all other components in W and H are fixed:

$$
\begin{aligned}
\|A - \sum_{j=1}^{k} w_j h_j^T\|_F &= \|(A - \sum_{\substack{j=1 \\ j \neq i}}^{k} w_j h_j^T) - w_i h_i^T\|_F = \|R^{(i)} - w_i h_i^T\|_F \\
w_i &\leftarrow \arg\min_{w_i \geq 0} \|R^{(i)} - w_i h_i^T\|_F \\
h_i &\leftarrow \arg\min_{h_i \geq 0} \|R^{(i)} - w_i h_i^T\|_F
\end{aligned}
$$

- Each subproblem has the form $\min_{x \geq 0} \|cx^T - G\|_F$ and has a closed form solution $x = [\frac{G^T c}{c^T c}]_+$!
  Hierarchical Alternating Least Squares (HALS) (Cichocki et al, 07, 09), (actually HA-NLS)
  Rank-one Residue Iteration (RRI) (Ho, 08)

BCD with Scalar Blocks vs. $2k$ Vector Blocks

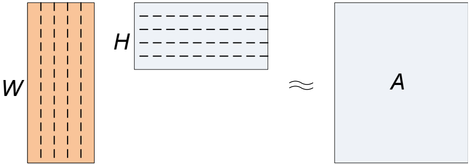


- In scalar BCD, $w_{1j}, w_{2j}, \cdots, w_{mj}$ can be computed independently. Also, $h_{i1}, h_{i2}, \cdots, h_{in}$ can be computed independently.
  $\rightarrow$ scalar BCD $\Leftrightarrow$ $2k$ vector BCD in NMF

Successive Rank-1 Deflation in SVD and NMF

- Successive rank-1 deflation works for SVD but not for NMF
  $A - \sigma_1 u_1 v_1^T \approx \sigma_2 u_2 v_2^T?$ $\qquad$ $A - w_1 h_1^T \approx w_2 h_2^T?$
- $\begin{pmatrix} 4 & 6 & 0 \\ 6 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 10 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{pmatrix}$
- 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
  $$WH = \begin{pmatrix} 4 & 6 & 0 \\ 6 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 6 & 4 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$
- NOTE: $2k$ vector BCD $\neq$ successive rank-1 deflation for NMF

BCD with 2 Matrix Blocks



- Minimize functions of W or H while the other is fixed:

$$\begin{aligned} W &\leftarrow \arg\min_{W \geq 0} \| H^T W^T - A^T \|_F \\ H &\leftarrow \arg\min_{H \geq 0} \| WH - A \|_F \end{aligned}$$

- 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) ??

NLS : $\min_{X \geq 0} \|CX - B\|_F^2 = \sum \min_{x_i} \|Cx_i - b_i\|_2^2$

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 α_k
- Projected Quasi-Newton method (Kim et al., 07)

  $$x^{(k+1)} \leftarrow \begin{bmatrix} y \\ z_k \end{bmatrix} = \begin{bmatrix} \mathcal{P}_+ \left[y^{(k)} - \alpha D^{(k)} \nabla f(y^{(k)}) \right] \\ 0 \end{bmatrix}$$

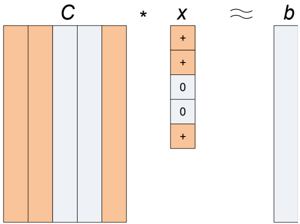
  * Gradient scaling only for nonzero variables
- These do not fully exploit the structure of the NLS problems in NMF
- Active Set method (H. Kim and Park, (08)

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

- KKT conditions: $y = C^TCx - C^Tb$
  $y \geq 0, \quad x \geq 0, \quad x_iy_i = 0, \; i = 1, \cdots, k$
- If we know $P = \{i|x_i > 0\}$ in the solution in advance then we only need to solve $\min\|C_Px_P - b\|_2$, and the rest of $x_i = 0$, where C_P: columns of C with the indices in P



Active-set type Algorithms for $\min_{x\geq 0} \|Cx - b\|_2, C : m \times k$

- KKT conditions: $y = C^TCx - C^Tb$
  $y \geq 0, \quad x \geq 0, \quad x_iy_i = 0, \ i = 1, \cdots, k$
- Active set method (Lawson and Hanson 74)
  - $E = \{1, \cdots, k\}$ (i.e. $x = 0$ initially), $P =$ 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 $\{1, \cdots, k\}$
  - Repeat
    - Let $x_E = 0$ and $x_P = \arg\min_{x_P} \|C_Px_P - b\|_2^2$
      Then $y_E = C_E^T(C_Px_P - b)$ 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

$k = 10$, 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 $y_E = C_E^T (C_P x_P - b)$

	x	y
P		0
P		0
P		0
P		0
P		0
E	0	
E	0	
E	0	
E	0	
E	0	

How block principal pivoting works

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

	x	y
P	+	0
P	-	0
P	-	0
P	+	0
P	-	0
E	0	-
E	0	+
E	0	-
E	0	+
E	0	+

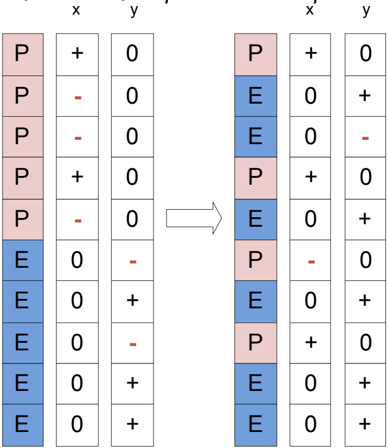
How block principal pivoting works

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

	x	y			x	y
P	+	0		P		0
P	-	0		E	0	
P	-	0		E	0	
P	+	0		P		0
P	-	0	⇒	E	0	
E	0	-		P		0
E	0	+		E	0	
E	0	-		P		0
E	0	+		E	0	
E	0	+		E	0	

How block principal pivoting works

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

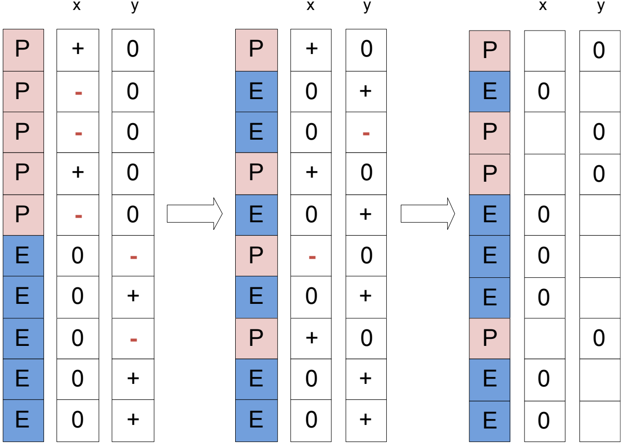


	x	y
P	+	0
P	-	0
P	-	0
P	+	0
P	-	0
E	0	-
E	0	+
E	0	-
E	0	+
E	0	+

	x	y
P	+	0
E	0	+
E	0	-
P	+	0
E	0	+
P	-	0
E	0	+
P	+	0
E	0	+
E	0	+

How block principal pivoting works

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



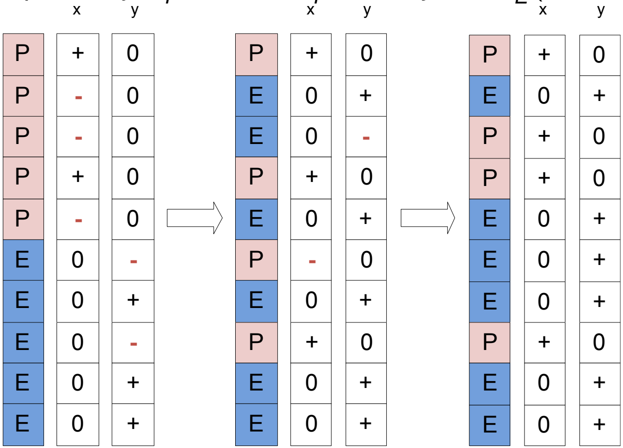
	x	y
P	+	0
P	-	0
P	-	0
P	+	0
P	-	0
E	0	-
E	0	+
E	0	-
E	0	+
E	0	+

	x	y
P	+	0
E	0	+
E	0	-
P	+	0
E	0	+
P	-	0
E	0	+
P	+	0
E	0	+
E	0	+

	x	y
P		0
E	0	
P		0
P		0
E	0	
E	0	
E	0	
P		0
E	0	
E	0	

How block principal pivoting works

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



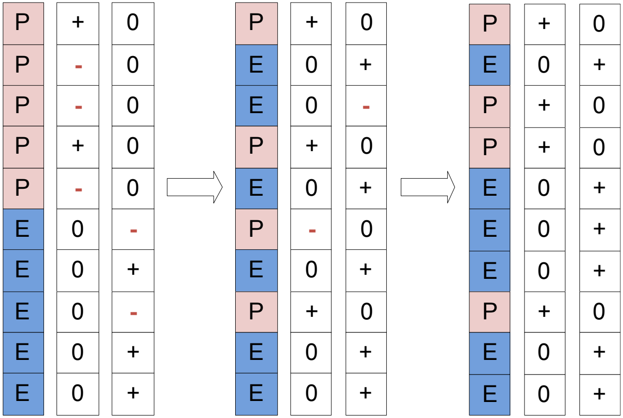
	x	y
P	+	0
P	-	0
P	-	0
P	+	0
P	-	0
E	0	-
E	0	+
E	0	-
E	0	+
E	0	+

	x	y
P	+	0
E	0	+
E	0	-
P	+	0
E	0	+
P	-	0
E	0	+
P	+	0
E	0	+
E	0	+

	x	y
P	+	0
E	0	+
P	+	0
P	+	0
E	0	+
E	0	+
E	0	+
P	+	0
E	0	+
E	0	+

How block principal pivoting works

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



	x	y
P	+	0
P	-	0
P	-	0
P	+	0
P	-	0
E	0	-
E	0	+
E	0	-
E	0	+
E	0	+

	x	y
P	+	0
E	0	+
E	0	-
P	+	0
E	0	+
P	-	0
E	0	+
P	+	0
E	0	+
E	0	+

	x	y
P	+	0
E	0	+
P	+	0
P	+	0
E	0	+
E	0	+
E	0	+
P	+	0
E	0	+
E	0	+

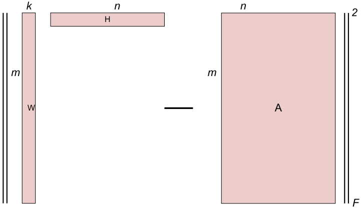
Solved!

Refined Exchange Rules

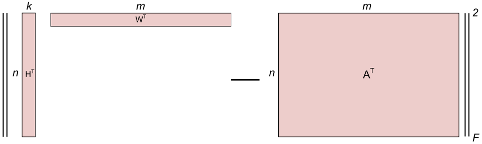
- Active set algorithm is a special instance of single principal pivoting algorithm (H. Kim and Park, SIMAX 08)
- Block exchange rule without modification does not always work.
  - The residual is <u>not</u> guaranteed to monotonically decrease. Block exchange rule may cycle (although rarely).
  - Modification: 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 (J. Kim and Park, 08)

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

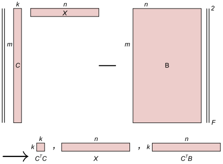


- $\min_{W \geq 0} \|H^T W^T - A^T\|_F^2$

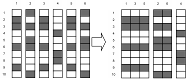


Efficient Algorithm for $\min_{X\geq 0} \|CX - B\|_F^2$ (J. Kim and Park, 08)

- Precompute C^TC and C^TB
  Update x_P and y_E by $C_P^TC_Px_P = C_P^Tb$ and $y_E = C_E^TC_Px_P - C_E^Tb$
  All coefficients can be retrieved from C^TC and C^TB
- C^TC and C^TB 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. → 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

Sparse NMF and Regularized NMF

- 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

$$\min_{H\geq 0}\left\|\begin{pmatrix} W \\ \sqrt{\beta}e_{1\times k}\end{pmatrix}H-\begin{pmatrix} A \\ 0_{1\times n}\end{pmatrix}\right\|_F^2$$

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

- Regularized NMF (Pauca, et al. 06):

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

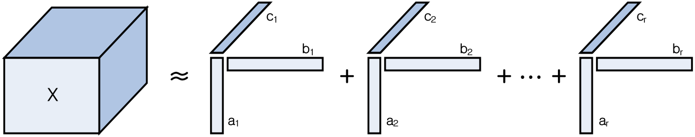
ANLS reformulation : alternate the following

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

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

Nonnegative PARAFAC

- Consider a 3-way Nonnegative Tensor $\underline{\mathbf{T}} \in \mathbb{R}_{+}^{m \times n \times p}$ 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.



Nonnegative PARAFAC (J. Kim and Park, in preparation)

- Iterate until a stopping criteria is satisfied:
  - $\min_{A \geq 0} \left\| Y_{BC} A^T - T_{(1)} \right\|_F$
  - $\min_{B \geq 0} \left\| Y_{AC} B^T - T_{(2)} \right\|_F$
  - $\min_{C \geq 0} \left\| Y_{AB} C^T - T_{(3)} \right\|_F$ where

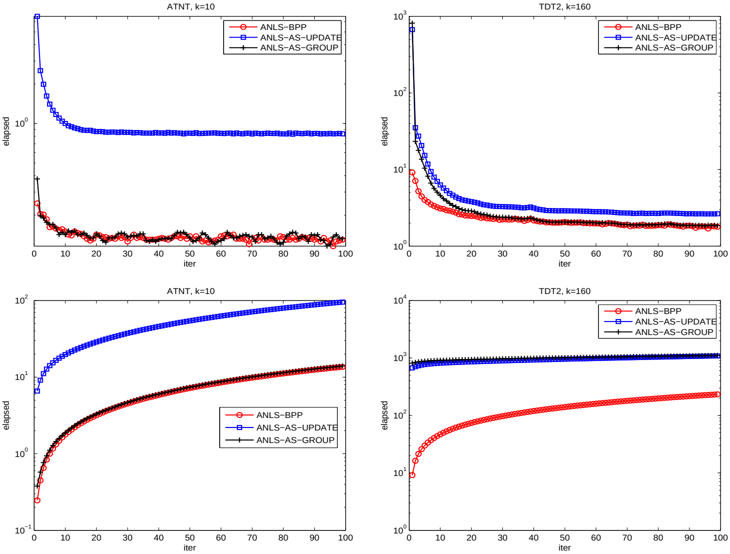
    $Y_{BC} = B \odot C \in \mathbb{R}^{(np) \times k}$, $T_{(1)} \in \mathbb{R}^{(np) \times m}$,
    $Y_{AC} = A \odot C \in \mathbb{R}^{(mp) \times k}$, $T_{(2)} \in \mathbb{R}^{(mp) \times n}$,
    $Y_{AB} = A \odot B \in \mathbb{R}^{(mn) \times k}$, $T_{(3)} \in \mathbb{R}^{(mn) \times p}$ unfolded 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) (J. Kim and Park, 2011)

NMF Algorithms Compared

Name	Description	Author
ANLS-BPP	ANLS / block principal pivoting	J. Kim and HP 08
ANLS-AS	ANLS / active set	H. Kim and HP 08
ANLS-PGRAD	ANLS / projected gradient	Lin 07
ANLS-PQN	ANLS / projected quasi-Newton	D. Kim et al. 07
HALS	Hierarchical ALS	Cichocki et al. 07
MU	Multiplicative updating	Lee and Seung 01
ALS	Alternating least squares	Berry et al. 06

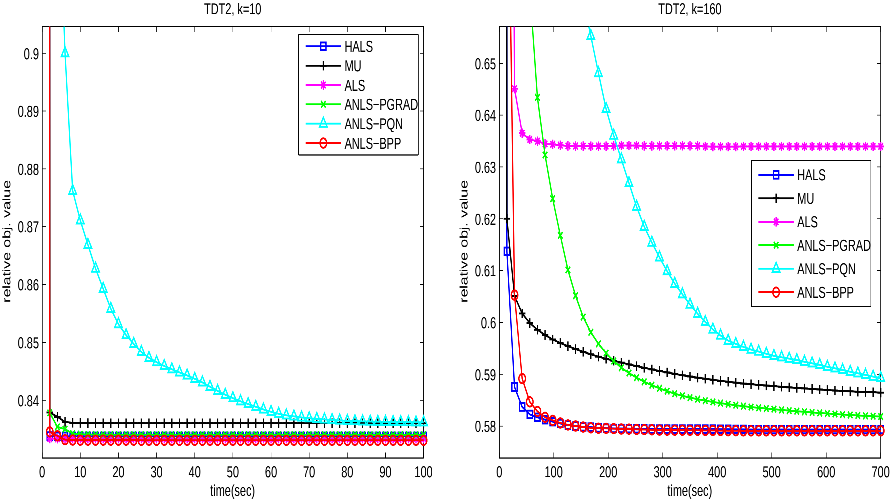
Active-set vs. Block principal pivoting (J. Kim and Park, 2011)



ATNT image data: $10,304 \times 400$, $k = 10$, TDT2 text data:$19,009 \times 3,087$, $k = 160$

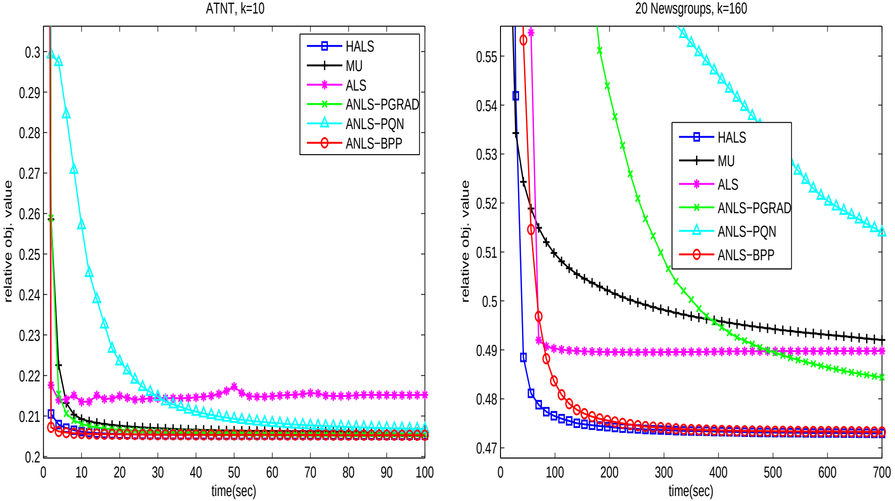
Top:time per iteration, bottom:cumulative time

Residual vs. Execution time (J. Kim and Park, 2011)



TDT2 text data: $19,009 \times 3,087$, $k = 10$ and $k = 160$

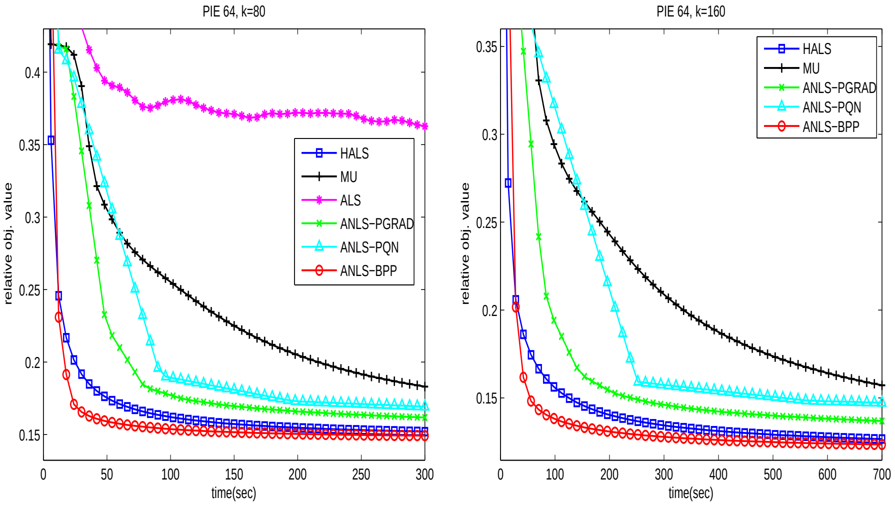
Residual vs. Execution time (J. Kim and Park, 2011)



ATNT image data: $10,304 \times 400$, $k = 10$ and

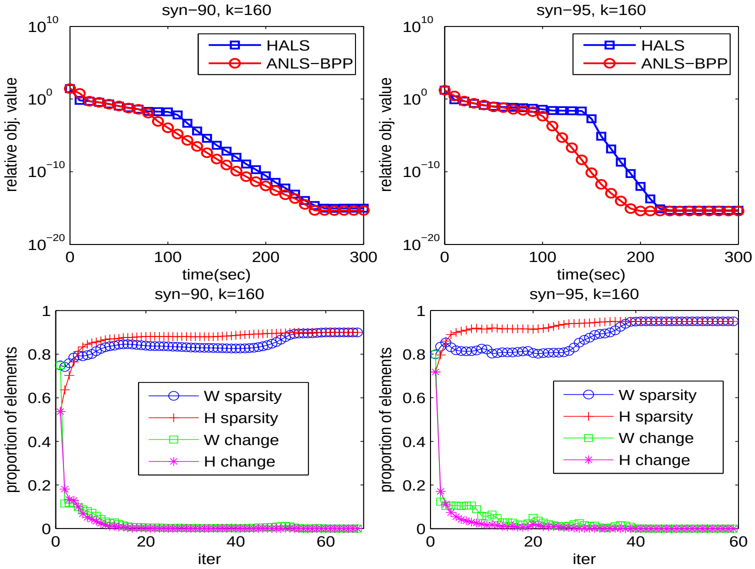
20 Newsgroups text data: $26,214 \times 11,314$, $k = 160$

Residual vs. Execution time (J. Kim and Park, 2011)



PIE 64 image data: $4,096 \times 11,554$, $k = 80$ and $k = 160$

BPP vs HALS : Influence of Sparsity (J. Kim and Park, 2011)



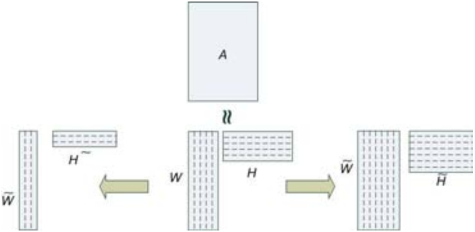
Synthetic data $10,000 \times 2,000$ created by factors with different sparsities

Left: 90% sparsity, Right: 95% sparsity

Adaptive NMF for Varying Reduced Rank $k \longrightarrow \tilde{k}$

(He, Kim, Cichocki, and Park, in preparation)

- Given (W, H) with k, how to compute $(\tilde{W}, \tilde{H})$ with $\tilde{k}$ fast?
- E.g., model selection for NMF clustering



AdaNMF

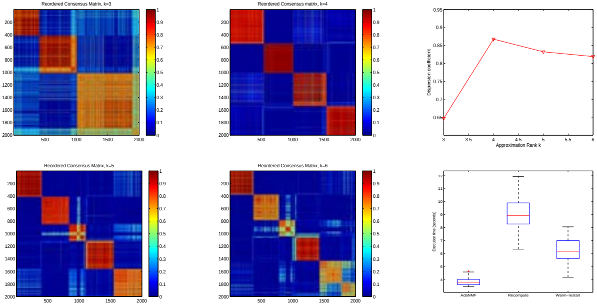
- Initialize $\tilde{W}$ and $\tilde{H}$ using W and H
  - If $\tilde{k} > k$, compute NMF for $A - WH \approx \Delta W \Delta H$. Set $\tilde{W} = [W \; \Delta W]$ and $\tilde{H} = [H; \Delta H]$
  - If $\tilde{k} < k$, initialize $\tilde{W}$ and $\tilde{H}$ with $\tilde{k}$ pairs of (w_i, h_i) with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

Model Selection in NMF Clustering (He, Kim, Cichocki, and Park, in preparation)

- Consensus matrix based on $A \approx WH$:

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

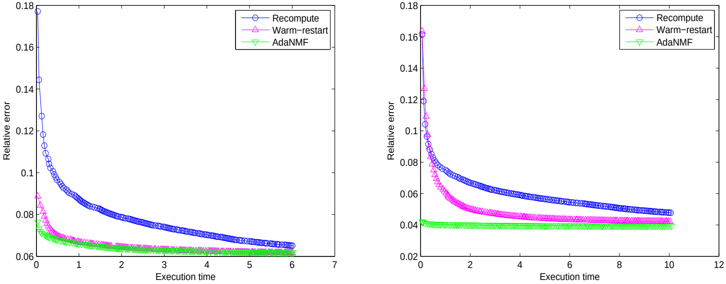
- Dispersion coefficient $\rho(k) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 4(C_{ij} - \frac{1}{2})^2$, where $C = \frac{1}{l} \sum C^t$



Clustering results on MNIST digit images (784 × 2000) by AdaNMF with $k = 3, 4, 5$ and 6. Averaged consensus matrices, dispersion coefficient, execution time

Adaptive NMF for Varying Reduced Rank

(He, Kim, Cichocki, and Park, in preparation)

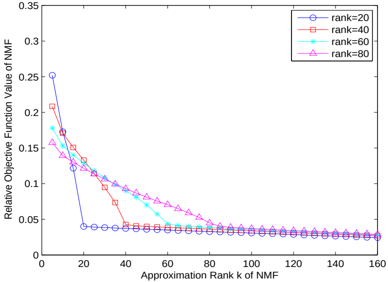


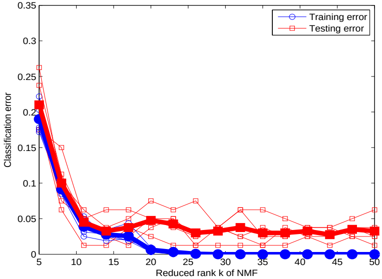
Relative error vs. exec. time of AdaNMF and "recompute". Given an NMF of 600 $\times$ 600 synthetic matrix with $k = 60$, compute NMF with $\tilde{k} = 50, 80$.

Adaptive NMF for Varying Reduced Rank

(He, Kim, Cichocki, and Park, in preparation)

Theorem: For $A \in \mathbb{R}_+^{m\times n}$, If $rank_+(A) > k$, 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 $H^{(i)} \in \mathbb{R}_+^{i\times n}$.





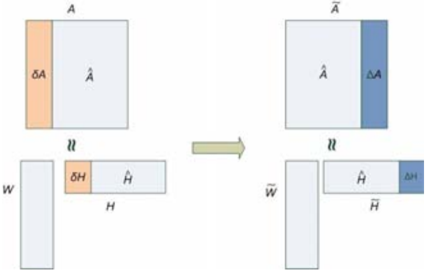
Rank path on synthetic data set: relative residual vs. k

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 $\min_{H_T \geq 0} \|WH_T - T\|_F$

NMF for Dynamic Data (DynNMF) (He, Kim, Cichocki, and Park, in preparation)

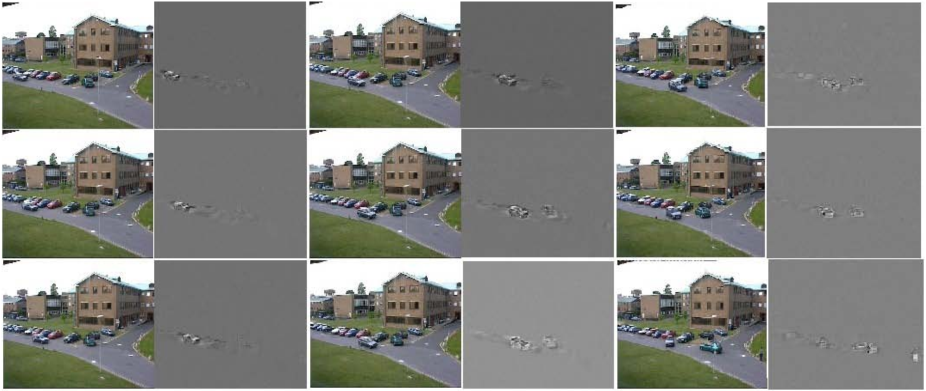
- Given an NMF (W, H) for $A = [\delta A\ \hat{A}]$, how to compute NMF $(\tilde{W}, \tilde{H})$ for $\tilde{A} = [\hat{A}\ \Delta A]$ fast ?
  (Updating and Downdating)



DynNMF (Sliding Window NMF)

- Initialize $\tilde{H}$ as follows:
  - Let $\hat{H}$ be the remaining columns of H.
  - Solve $\min_{\Delta H \geq 0} \|W\Delta H - \Delta A\|_F^2$ using block principal pivoting
  - Set $\tilde{H} = [\hat{H}\ \ \Delta H]$
- Run HALS on $\tilde{A}$ with initial factors $\tilde{W} = W$ and $\tilde{H}$

DynNMF for Dynamic Data (He, Kim, Cichocki, and Park, in preparation)



PET2001 data with 3064 images from a surveillance video. DynNMF on $110,592 \times 400$ data matrix each time, with 100 new columns and 100 obsolete columns. The residual images track the moving vehicle in the video.

NMF as a Clustering Method (Kuang and Park, in preparation)

- Clustering and Lower Rank Approximation are related.

  NMF for Clustering: Document (Xu et al. 03), Image (Cai et al. 08), Microarray (Kim & Park 07), etc.

- Equivalence of objective functions between k-means and NMF:

  (Ding, et al., 05; Kim & Park, 08)

$$\min \sum_{i=1}^{n} \|a_i - w_{S_i}\|_2^2 = \min \|A - WH\|_F^2$$

$S_i = j$ when i-th point is assigned to j-th cluster ($j \in \{1, \cdots, k\}$)
k-means: W: k cluster centroids, $H \in \mathbf{E}$
NMF: W : basis vectors for rank-k approximation,
H: representation of A in W space
($\mathbf{E}$: matrices whose columns are columns of an identity matrix)
NOTE: The equivalence of obj. functions holds when $H \in \mathbf{E}$, $A \geq 0$.

NMF and K-means

$\min \|A - WH\|_F^2$ s.t. $H \in \mathbf{E}$

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

TDT2 text data set: (clustering accuracy aver. among 100 runs)

# clusters	2	6	10	14	18
K-means	0.8099	0.7295	0.7015	0.6675	0.6675
NMF/ANLS	0.9990	0.8717	0.7436	0.7021	0.7160

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):

k	3	6	9	12	15
NMF	69	65	74	68	44
SNMF	100	100	100	100	97

NMF and SNMF much better than k-means in general.

NMF, K-means, and Spectral Clustering (Kuang and Park, in preparation)

Equivalence of objective functions is not enough to explain the clustering capability of NMF:

- NMF is more related to spherical k-means, than to k-means $\longrightarrow$ NMF shown to work well in text data clustering

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

- Spectral clustering → Eigenvectors (Ng et al. 01), A normalized if needed, Laplacian,...
- Symmetric NMF (Ding et al.) → can handle nonlinear structure, and $S \geq 0$ natually captures a cluster structure in S

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
- Extensive computational comparisons

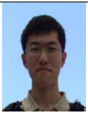
Summary/Discussions

- 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
- Extensive computational comparisons

- NMF for clustering and semi-supervised clustering
- NMF and probability related methods
- NMF and geometric understanding
- NMF algorithms for large scale problems, parallel implementation? GPU?
- Fast NMF with other divergences (Bregman and Csiszar divergences)
- NMF for blind source separation? Uniqueness?
- More theoretical study on NMF especially for foundations for computational methods

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

Collaborators

Today's talk		
		
Jingu Kim	Yunlong He	Da Kuang
CSE, Georgia Tech	Math, Georgia Tech	CSE, Georgia Tech

Krishnakumar Balasubramanian	CSE, Georgia Tech
Prof. Michael Berry	EECS, Univ. of Tennessee
Prof. Moody Chu	Math, North Carolina State Univ.
Dr. Andrzej Cichocki	Brain Science Institute, RIKEN, Japan
Prof. Chris Ding	CSE, UT Arlington
Prof. Lars Elden	Math, Linkoping Univ., Sweden
Dr. Mariya Ishteva	CSE, Georgia Tech
Dr. Hyunsoo Kim	Wistar Inst.
Anoop Korattikara	CS, UC Irvine
Prof. Guy Lebanon	CSE, Georgia Tech
Liangda Li	CSE, Georgia Tech
Prof. Tao Li	CS, Florida International Univ.
Prof. Robert Plemmons	CS, Wake Forest Univ.
Andrey Puretskiy	EECS, Univ. of Tennessee
Prof. Max Welling	CS, UC Irvine
Dr. Stan Young	NISS

Thank you!

Related Papers by H. Park's Group

- H. Kim and H. Park, Sparse Non-negative Matrix Factorizations via Alternating Non-negativity-constrained Least Squares for Microarray Data Analysis. Bioinformatics, 23(12):1495-1502, 2007.
- H. Kim, H. Park, and L. Eldén. Non-negative Tensor Factorization Based on Alternating Large-scale Non-negativity-constrained Least Squares. Proc. of IEEE 7th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 1147-1151, 2007.
- H. Kim and H. Park, Nonnegative Matrix Factorization Based on Alternating Non-negativity-constrained Least Squares and the Active Set Method. SIAM Journal on Matrix Analysis and Applications, 30(2):713-730, 2008.
- J. Kim and H. Park, Sparse Nonnegative Matrix Factorization for Clustering, Georgia Tech Technical Report GT-CSE-08-01, 2008.
- J. Kim and H. Park. Toward Faster Nonnegative Matrix Factorization: A New Algorithm and Comparisons. Proc. of the 8th IEEE International Conference on Data Mining (ICDM), pp. 353-362, 2008.
- B. Drake, J. Kim, M. Mallick, and H. Park, Supervised Raman Spectra Estimation Based on Nonnegative Rank Deficient Least Squares. In Proceedings of the 13th International Conference on Information Fusion, Edinburgh, UK, 2010.

Related Papers by H. Park's Group

- A. Korattikara, L. Boyles, M. Welling, J. Kim, and H. Park, Statistical Optimization of Non-Negative Matrix Factorization. Proc. of The Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR: W&CP 15, 2011.
- J. Kim and H. Park, Fast Nonnegative Matrix Factorization: an Active-set-like Method and Comparisons, Submitted for review, 2011.
- J. Kim and H. Park, Fast Nonnegative Tensor Factorization with an Active-set-like Method, In High-Performance Scientific Computing: Algorithms and Applications, Springer, in preparation.
- Y. He, J. Kim, A. Cichocki, and H. Park, Fast Adaptive NMF Algorithms for Varying Reduced Rank and Dynamic Data, in preparation.
- L. Li, G. Lebanon, and H. Park, Fast Algorithm for Non-Negative Matrix Factorization with Bregman and Csiszar Divergences, in preparation.
- D. Kuang and H. Park, Nonnegative Matrix Factorization for Spherical and Spectral Clustering, in preparation.