

# **Towards Faster Nonnegative Tensor Factorization: A New Active-Set type Algorithm and Comparisons**

Haesun Park

hpark@cc.gatech.edu

College of Computing

Georgia Institute of Technology

Atlanta, GA 30332, USA

Joint work with Krishnakumar Balasubramanian, Hyunsoo Kim, Jingu Kim and Lars Elden

NSF Tensor Workshop, Feb. 20-21, 2009

# Outline

---

- New algorithms for NMF (Nonnegative Matrix Factorization) and NTF(Nonnegative PARAFAC)
  - Algorithms for NMF
  - Block principal pivoting algorithm
  - Comparison results (NMF)
  - Extension to NTF(Nonnegative PARAFAC)
  - results (NTF)
  - Summary

# Alternating Nonnegative Least Squares for NMF

Given  $A \in \mathbb{R}_+^{m \times n}$  and a desired rank  $k$ , find  $W \in \mathbb{R}_+^{m \times k}$  and  $H \in \mathbb{R}_+^{k \times n}$  such that  $A \approx WH \implies \min_{W \geq 0, H \geq 0} \|A - WH\|_F^2$

1. Initialize  $W \geq 0$  (or  $H \geq 0$ )
  2. Iterate the following ANLS until a stopping criteria is satisfied:
    - (a) Solve  $\min_{H \geq 0} \|WH - A\|_F^2$
    - (b) Solve  $\min_{W \geq 0} \|H^T W^T - A^T\|_F^2$
  3. The columns of  $W$  are normalized to unit  $L_2$ -norm
- Convergence : Any limit point of the sequence is a stationary point [Grippo and Sciandrone '00]
  - Alternating Nonnegative Least Squares (ANLS) [Lin '07, Kim et al '07, H. Kim and Park '08]
  - Alternating Least Squares(ALS) [Berry et al '06]: convergence is difficult to analyse, but can solve each sub-problem fast.
  - Multiplicative Updating Rules [Lee and Seung '01]: Simple to implement, but residual non-increasing property may not imply convergence to a stationary point.
  - Other algorithms and variants [Li et al '01, Hoyer '04, Pauca et al '04, Gao and Church '05, Chu and Lin '08]

# NMF/ANLS Algorithms

$$\text{Sub-problem : } \min_{X \geq 0} \|CX - B\|_F^2$$

- Active Set [H. Kim and Park, SIMAX '08]
  - Classical algorithm for NNLS with single right hand side ( $\min_{x \geq 0} \|Cx - b\|_2$ ) [Lawson and Hansen '95]
  - Faster algorithms for multiple right hand side problems [Bro and de Jong, 1997], and [Van Benthem and Keenan '04].

- Projected Gradient [Lin '07]

$$x^{k+1} \leftarrow \mathcal{P}_+(x^k - \alpha_k \nabla f(x^k))$$

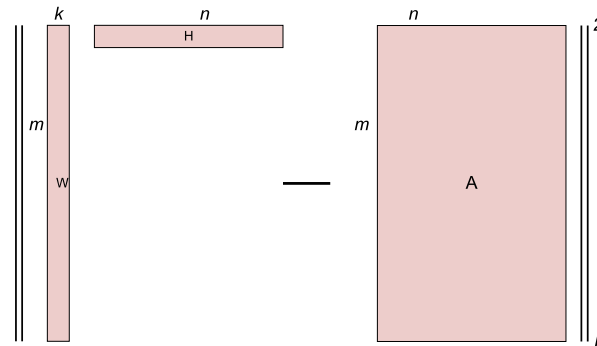
- Improved selection of step constant  $\alpha_k$
- Projected Quasi-Newton [Kim '07]

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

- Gradient scaling only for inactive variables

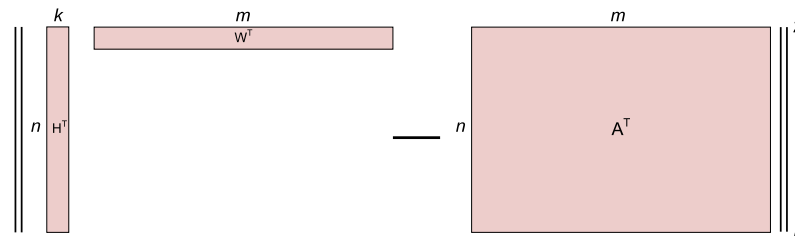
# Structure of Sub-problems in NMF

- Recognizing the structure is important for developing a fast algorithm for NMF :  $k \ll \min(m, n)$
- $\min_{H \geq 0} \|WH - A\|_F^2$



$W \in \mathbb{R}_+^{m \times k}$  is long and thin and  $A \in \mathbb{R}_+^{m \times n}$  has  $n$  right hand sides.

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



$H^T \in \mathbb{R}_+^{n \times k}$  is long and thin and  $A^T \in \mathbb{R}_+^{n \times m}$  has  $m$  right hand sides.

# Block Principal Pivoting Algorithm

- Consider single right-hand side problem [Portugal et al '94]: for  $x \in \mathbb{R}^q$

$$\min_{x \geq 0} \|Cx - b\|_2^2$$

- KKT conditions:

$$y = C^T Cx - C^T b \quad (1a)$$

$$y \geq 0 \quad (1b)$$

$$x \geq 0 \quad (1c)$$

$$x_i y_i = 0, \quad i = 1, \dots, q \quad (1d)$$

- Need to find  $x$  and  $y$  that satisfy KKT conditions.
- Guess two index sets  $F$  and  $G$  that partition  $\{1, \dots, q\}$
- Repeat:
  - Set  $x_G = 0$ .
  - Solve  $x_F = \arg \min_{x_F} \|C_F x_F - b\|_2^2$
  - Set  $y_F = 0$
  - Set  $y_G = C_G^T (C_F x_F - b)$ .
  - If  $x_F \geq 0$  and  $y_G \geq 0$ , solution found. Otherwise, update  $F$  and  $G$ .

# How block principal pivoting works

Update by  $C_F^T C_F x_F = C_F^T b$  and  $y_G = C_G^T C_F x_F - C_G^T b$ .

|   | x | y |
|---|---|---|
| F | + | 0 |
| F | - | 0 |
| F | - | 0 |
| F | + | 0 |
| F | - | 0 |
| G | 0 | - |
| G | 0 | + |
| G | 0 | - |
| G | 0 | + |
| G | 0 | + |

# How block principal pivoting works

Update by  $C_F^T C_F x_F = C_F^T b$  and  $y_G = C_G^T C_F x_F - C_G^T b$ .

|   | x | y |   | x | y |
|---|---|---|---|---|---|
| F | + | 0 | → | F | 0 |
| F | - | 0 |   | G | 0 |
| F | - | 0 |   | G | 0 |
| F | + | 0 |   | F | 0 |
| F | - | 0 |   | G | 0 |
| G | 0 | - |   | F | 0 |
| G | 0 | + |   | G | 0 |
| G | 0 | - |   | F | 0 |
| G | 0 | + |   | G | 0 |
| G | 0 | + |   | G | 0 |

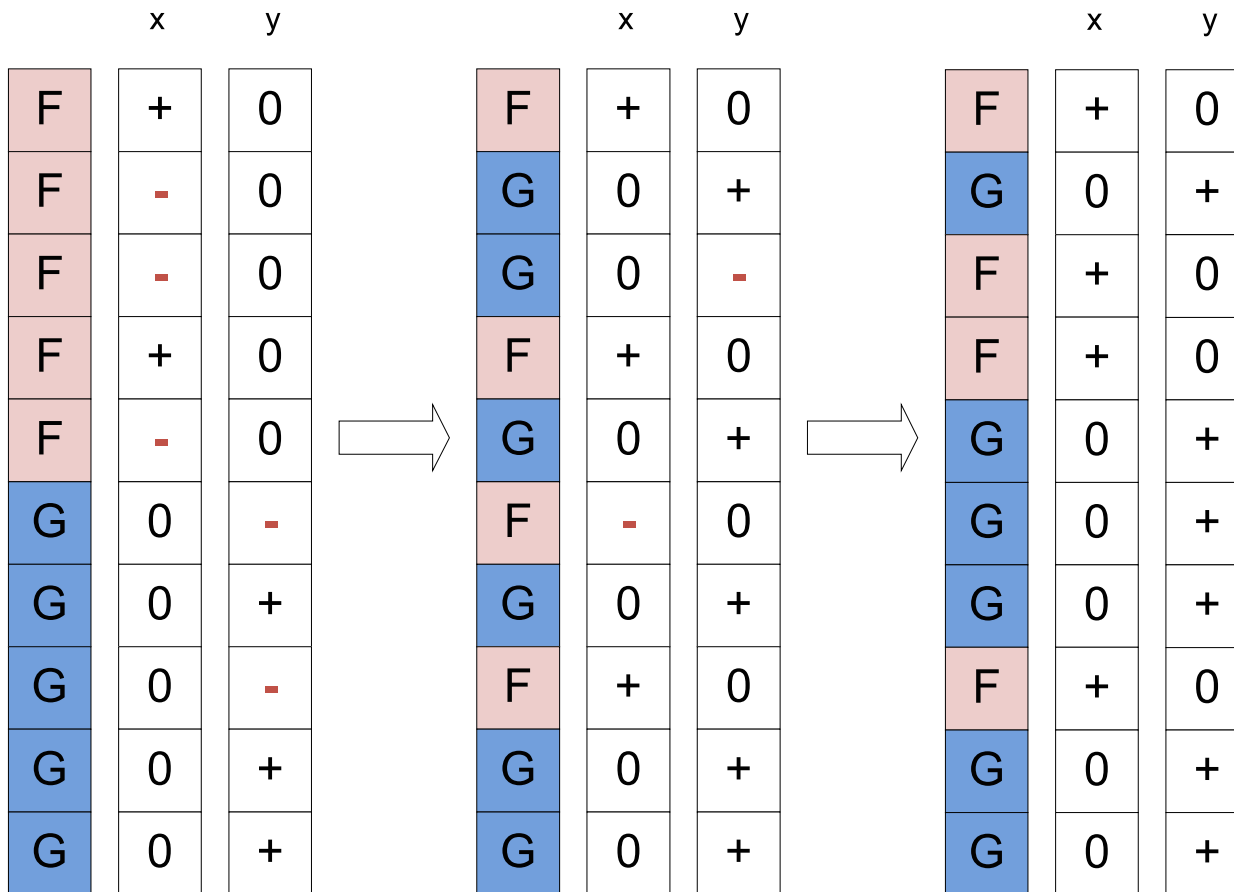
# How block principal pivoting works

Update by  $C_F^T C_F x_F = C_F^T b$  and  $y_G = C_G^T C_F x_F - C_G^T b$ .

|   | x | y |   | x | y |   |
|---|---|---|---|---|---|---|
| F | + | 0 | → | F | + | 0 |
| F | - | 0 |   | G | 0 | + |
| F | - | 0 |   | G | 0 | - |
| F | + | 0 |   | F | + | 0 |
| F | - | 0 |   | G | 0 | + |
| G | 0 | - |   | F | - | 0 |
| G | 0 | + |   | G | 0 | + |
| G | 0 | - |   | F | + | 0 |
| G | 0 | + |   | G | 0 | + |
| G | 0 | + |   | G | 0 | + |

# How block principal pivoting works

Update by  $C_F^T C_F x_F = C_F^T b$  and  $y_G = C_G^T C_F x_F - C_G^T b$ .



Solved!

# Active-Set type Algorithms

---

- Active-Set Algorithm:
  - One column is replaced most of the time
  - Residual is guaranteed to monotonically decrease
  - Careful exchange rule requires many iterations
  - Can be faster when the solution is sparse
- Block Principal Pivoting Algorithm:
  - Multiple columns are replaced each time
  - Residual is not guaranteed to decrease
  - Backup exchange rule guarantees BPP to find the solution in a finite number of iterations
  - Can be faster when the solution vector is dense or long

# NNLS with Multiple right-hand side for NMF

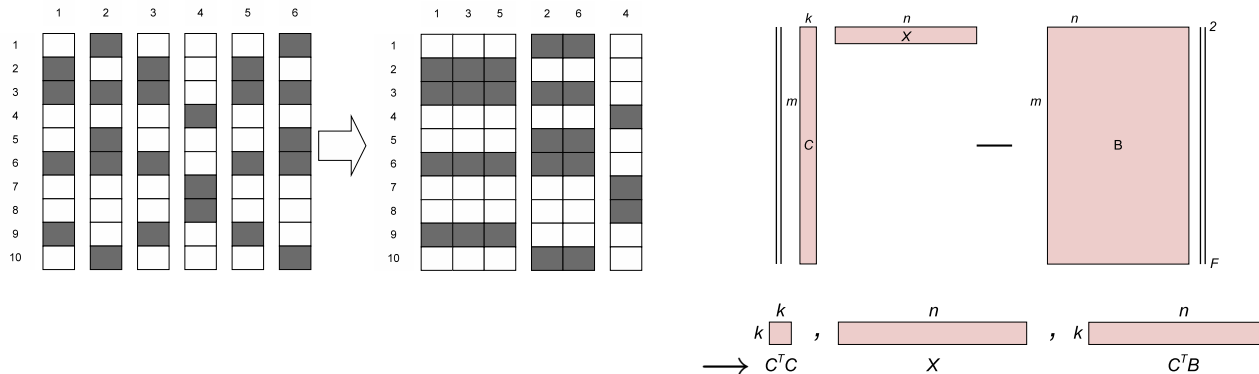
$$\min_{X \geq 0} \|CX - B\|_F^2$$

- Block principal pivoting [Kim and Park '08]
- Exploit long and thin structure
  - Precompute  $C^T C$  and  $C^T B$ : updates of  $x_F$  and  $y_G$  is given by

$$\begin{aligned} C_F^T C_F x_F &= C_F^T b \\ y_G &= C_G^T C_F x_F - C_G^T b. \end{aligned}$$

All coefficients can be directly retrieved from  $C^T C$  and  $C^T B$

- $C^T C$  and  $C^T B$  is small. → Storage is not a problem.
- Exploiting common  $F$  and  $G$  sets.



- $X$  is flat and wide. → More common cases of  $F$  and  $G$  sets.

# Extension to Sparse NMF and Regularized NMF

- Sparse NMF [H. Kim and Park, Bioinformatics '07]

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

subject to  $W, H \geq 0$ .

ANLS reformulation: alternate the following

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

- Similar reformulation for regularized NMF: [Pauca '06]

$$\min_{W, H} \left\{ \|A - WH\|_F^2 + \alpha \|W\|_F^2 + \beta \|H\|_F^2 \right\} \quad (3)$$

subject to  $W, H \geq 0$ .

# Comparison results (NMF)

- Stopping criterion: normalized KKT optimality condition

$$\Delta \leq \epsilon \Delta_0, \text{ where } \Delta = \frac{\delta}{\delta_W + \delta_H}$$

- Data sets:

- Synthetic:  $300 \times 200$ , create sparse  $W$  and  $H$  and produce  $A = WH$  with noise
- Text: Topic Detection and Tracking 2, randomly select 20 topics,  $12617 \times 1491$
- Image: Olivetti Research Laboratory face image,  $10304 \times 400$

- Compared algorithms

- (**mult**) Lee and Seung's multiplicative updating algorithm['01]
- (**als**) Berry et al.'s alternating least squares algorithm ['07]
- (**lsqnonneg**) ANLS with Lawson and Hanson's active set algorithm for single right hand side ['95]
- (**projnewton**) ANLS with Kim et al.'s projected quasi-Newton algorithm ['07]
- (**projgrad**) ANLS with Lin's projected gradient algorithm ['07]
- (**activeset**) ANLS with Kim and Park's active set algorithm for multiple right hand sides ['07 Bioinformatics, '08 SIMAX]
- (**blockpivot**) Kim and Park's ANLS with block principal pivoting algorithm ['08 ICDM]

# Stopping Criterion

- KKT condition:

$$\begin{aligned}
 W &\geq 0 & H &\geq 0 \\
 \partial f(W, H)/\partial W &\geq 0 & \partial f(W, H)/\partial H &\geq 0 \\
 W. * (\partial f(W, H)/\partial W) &= 0 & H. * (\partial f(W, H)/\partial H) &= 0
 \end{aligned}$$

- These conditions can be simplified as

$$\min (W, \partial f(W, H)/\partial W) = 0 \quad (4a)$$

$$\min (H, \partial f(W, H)/\partial H) = 0 \quad (4b)$$

where the minimum is taken componentwise.

- Normalized KKT residual:

$$\Delta = \frac{\delta}{\delta_W + \delta_H} \quad (5)$$

where

$$\begin{aligned}
 \delta &= \sum_{i=1}^m \sum_{q=1}^k \left| \min(W_{iq}, (\partial f(W, H)/\partial W)_{iq}) \right| \\
 &\quad + \sum_{q=1}^k \sum_{j=1}^n \left| \min(H_{qj}, (\partial f(W, H)/\partial H)_{qj}) \right|
 \end{aligned} \quad (6)$$

$$\delta_W = \# (\min(W, (\partial f(W, H)/\partial W) \neq 0) \quad (7)$$

$$\delta_H = \# (\min(H, (\partial f(W, H)/\partial H) \neq 0). \quad (8)$$

# Synthetic data set

|            | $k$ | multi   | als     | lsqnonneg | projnewton | projgrad | activeset | blockpivot    |
|------------|-----|---------|---------|-----------|------------|----------|-----------|---------------|
| time (sec) | 5   | 35.336  | 36.697  | 23.188    | 5.756      | 0.976    | 0.262     | <b>0.252</b>  |
|            | 10  | 47.132  | 52.325  | 82.619    | 13.43      | 4.157    | 0.848     | <b>0.786</b>  |
|            | 20  | 72.888  | 83.232  |           | 45.007     | 9.32     | 4.41      | <b>4.004</b>  |
|            | 30  |         |         |           | 127.33     | 62.317   | 17.252    | <b>14.384</b> |
|            | 40  |         |         |           |            | 81.445   | 22.246    | <b>16.132</b> |
|            | 60  |         |         |           |            | 128.76   | 37.376    | <b>21.368</b> |
|            | 80  |         |         |           |            | 276.29   | 65.566    | <b>30.055</b> |
| iterations | 5   | 9784.2  | 10000   | 25.6      | 25.8       | 30       | 26.4      | 26.4          |
|            | 10  | 10000   | 10000   | 34.8      | 35.2       | 45       | 35.2      | 35.2          |
|            | 20  | 10000   | 10000   |           | 70.8       | 104      | 69.8      | 69.8          |
|            | 30  |         |         |           | 166        | 205.2    | 166.6     | 166.6         |
|            | 40  |         |         |           |            | 234.8    | 118       | 117.8         |
|            | 60  |         |         |           |            | 157.8    | 84.2      | 84.2          |
|            | 80  |         |         |           |            | 131.8    | 67.2      | 67.2          |
| residual   | 5   | 0.04035 | 0.04043 | 0.04035   | 0.04035    | 0.04035  | 0.04035   | 0.04035       |
|            | 10  | 0.04345 | 0.04379 | 0.04343   | 0.04343    | 0.04344  | 0.04343   | 0.04343       |
|            | 20  | 0.04603 | 0.04556 |           | 0.04412    | 0.04414  | 0.04412   | 0.04412       |
|            | 30  |         |         |           | 0.04313    | 0.04316  | 0.04327   | 0.04327       |
|            | 40  |         |         |           |            | 0.04944  | 0.04943   | 0.04944       |

# Text data set

|            | $k$ | projgrad | activeset     | blockpivot     |
|------------|-----|----------|---------------|----------------|
| time (sec) | 5   | 107.24   | <b>81.476</b> | 82.954         |
|            | 10  | 131.12   | <b>87.012</b> | 88.728         |
|            | 20  | 161.56   | 154.1         | <b>144.77</b>  |
|            | 30  | 355.28   | 314.78        | <b>234.61</b>  |
|            | 40  | 618.1    | 753.92        | <b>479.49</b>  |
|            | 50  | 1299.6   | 1333.4        | <b>741.7</b>   |
|            | 60  | 1616.05  | 2405.76       | <b>1041.78</b> |
| iterations | 5   | 66.2     | 60.6          | 60.6           |
|            | 10  | 51.8     | 42            | 42             |
|            | 20  | 45.8     | 44.6          | 44.6           |
|            | 30  | 100.6    | 67.2          | 67.2           |
|            | 40  | 118      | 103.2         | 103.2          |
|            | 50  | 120.4    | 126.4         | 126.4          |
|            | 60  | 154.2    | 171.4         | 172.6          |
| residual   | 5   | 0.9547   | 0.9547        | 0.9547         |
|            | 10  | 0.9233   | 0.9229        | 0.9229         |
|            | 20  | 0.8898   | 0.8899        | 0.8899         |
|            | 30  | 0.8724   | 0.8727        | 0.8727         |
|            | 40  | 0.8600   | 0.8597        | 0.8597         |

# Image data set

|            | $k$ | projgrad | activeset     | blockpivot    |
|------------|-----|----------|---------------|---------------|
| time (sec) | 16  | 68.529   | <b>11.751</b> | 11.998        |
|            | 25  | 124.05   | 25.675        | <b>22.305</b> |
|            | 36  | 109.1    | 53.528        | <b>35.249</b> |
|            | 49  | 150.49   | 115.54        | <b>57.85</b>  |
|            | 64  | 169.7    | 270.64        | <b>91.035</b> |
|            | 81  | 249.45   | 545.94        | <b>146.76</b> |
| iterations | 16  | 26.8     | 16.4          | 16.4          |
|            | 25  | 20.6     | 15            | 15            |
|            | 36  | 17.6     | 13.4          | 13.4          |
|            | 49  | 16.2     | 12.4          | 12.4          |
|            | 64  | 16.6     | 13.2          | 13.2          |
|            | 81  | 16.8     | 14.4          | 14.4          |
| residual   | 16  | 0.1905   | 0.1907        | 0.1907        |
|            | 25  | 0.1757   | 0.1751        | 0.1751        |
|            | 36  | 0.1630   | 0.1622        | 0.1622        |
|            | 49  | 0.1524   | 0.1514        | 0.1514        |
|            | 64  | 0.1429   | 0.1417        | 0.1417        |
|            | 81  | 0.1343   | 0.1329        | 0.1329        |

size  $10304 \times 400$ ,  $\epsilon = 5 \times 10^{-4}$ . Average of 10 executions with different initial values.

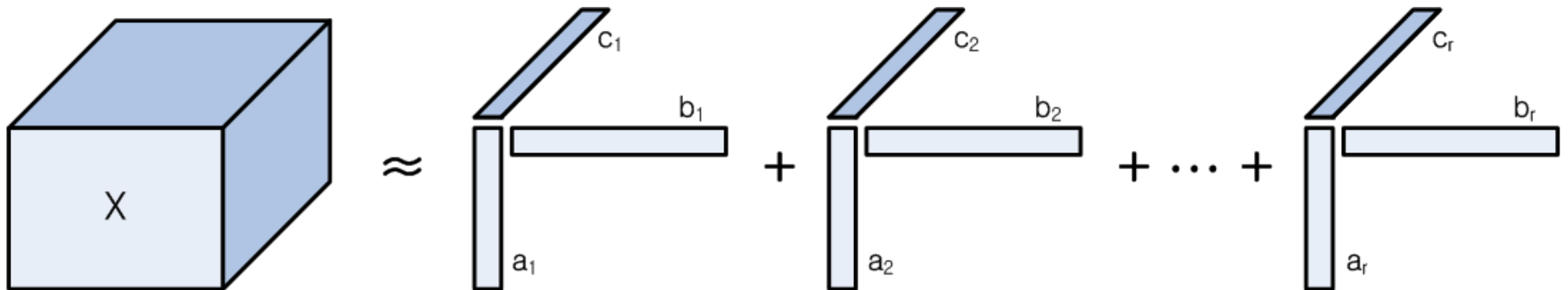
# Nonnegative Tensor Factorization (Nonnegative PARAFAC)

- For a three-way Nonnegative Tensor  $\mathbf{X} \in \mathbb{R}_+^{m \times n \times p}$  and an integer  $r$  we want

$$\min_{A, B, C \geq 0} \|\mathbf{X} - \llbracket ABC \rrbracket\|_F^2 = \min \sum_{i, j, z} \left( x_{ijz} - \sum_{q=1}^r a_{iq} b_{jq} c_{zq} \right)^2$$

where  $\llbracket ABC \rrbracket = \sum_{q=1}^r a_q \circ b_q \circ c_q$ ,  $A \in \mathbb{R}_+^{m \times r}$ ,  $B \in \mathbb{R}_+^{n \times r}$ ,  $C \in \mathbb{R}_+^{p \times r}$  and  $\circ$  represents vector outer product

- The loading matrices ( $A, B$  and  $C$ ) can be iteratively estimated by ANLS framework.
- The unfolding operation which facilitates this alternate formulation makes the matrices long and thin, which immediately makes the block-pivoting method efficient in solving it.



# Nonnegative Tensor Factorization

$$\min_{A,B,C \geq 0} \|\mathbf{X} - \llbracket ABC \rrbracket\|_F^2$$

1. Initialize  $B \in \mathbb{R}_+^{n \times r}$  and  $C \in \mathbb{R}_+^{p \times r}$

2. Iterate the following alternating until a stopping criteria is satisfied:



$$\min_{A \geq 0} \left\| Y_{BC} A^T - \mathbf{X}_{(1)} \right\|_F^2$$

where  $Y_{BC} = B \odot C$  and  $\mathbf{X}_{(1)}$  is the  $(np) \times m$  unfolded matrix.



$$\min_{B \geq 0} \left\| Y_{AC} B^T - \mathbf{X}_{(2)} \right\|_F^2$$

where  $Y_{AC} = A \odot C$  and  $\mathbf{X}_{(2)}$  is the  $(mp) \times n$  unfolded matrix, and



$$\min_{C \geq 0} \left\| Y_{AB} C^T - \mathbf{X}_{(3)} \right\|_F^2$$

where  $Y_{AB} = A \odot B$  and  $\mathbf{X}_{(3)}$  is the  $(mn) \times p$  unfolded matrix.

# Sparse Nonnegative Tensor Factorization

- This framework can be further extended to obtain Sparse NTF (e.g. sparse  $A$ ):

$$\min_{A, B, C \geq 0} \left\{ \|\mathbf{X} - \llbracket ABC \rrbracket\|_F^2 + \alpha \sum_{j=1}^r \|A(:, j)\|_1^2 + \beta \|B\|_F^2 + \gamma \|C\|_F^2 \right\}$$

- Here we iterate the following ANLS until convergence :

$$\min_{A \geq 0} \left\| \begin{pmatrix} Y_{BC} \\ \sqrt{\alpha} e_{1 \times r} \end{pmatrix} A^T - \begin{pmatrix} \mathbf{X}_{(1)} \\ 0_{1 \times m} \end{pmatrix} \right\|_F^2$$

$$\min_{B \geq 0} \left\| \begin{pmatrix} Y_{AC} \\ \sqrt{\beta} I_{r \times r} \end{pmatrix} B^T - \begin{pmatrix} \mathbf{X}_{(2)} \\ 0_{r \times n} \end{pmatrix} \right\|_F^2$$

$$\min_{C \geq 0} \left\| \begin{pmatrix} Y_{AB} \\ \sqrt{\gamma} I_{r \times r} \end{pmatrix} C^T - \begin{pmatrix} \mathbf{X}_{(3)} \\ 0_{r \times p} \end{pmatrix} \right\|_F^2$$

# Comparison results (NTF)

| Algo                           | r  | NTF.blockpivot | NTF.activeset | AB-PARAFAC-NC | NTF.mupdates |
|--------------------------------|----|----------------|---------------|---------------|--------------|
| Time(sec)                      | 5  | 0.6558         | 3.0233        | 16.7876       | 78.5518      |
|                                | 30 | 2.1932         | 11.0865       | 46.4766       | 171.7668     |
|                                | 50 | 6.9089         | 24.9563       | 76.4766       |              |
| $SSR = \sum_{i,j,z} e_{ijz}^2$ | 5  | 270.67         | 270.67        | 322.55        | 452.50       |
|                                | 20 | 270.31         | 270.31        | 320.56        | 352.68       |
|                                | 50 | 250.75         | 250.75        | 278.55        |              |

$\mathbf{X} \in \mathbb{R}_+^{50 \times 201 \times 61}$  is a randomly generated tensor. No. of Iterations was 26

| Algo                           | r  | NTF.blockpivot | NTF.activeset | AB-PARAFAC-NC | NTF.mupdates |
|--------------------------------|----|----------------|---------------|---------------|--------------|
| Time(sec)                      | 9  | 1.0558         | 1.9237        | 2.7651        | 308.5518     |
|                                | 50 | 8.1932         | 19.0865       | 32.0012       |              |
|                                | 90 | 40.9811        | 87.9563       | 132.5542      |              |
| $SSR = \sum_{i,j,z} e_{ijz}^2$ | 9  | 1890.67        | 1865.67       | 2321.02       | 3452.50      |
|                                | 50 | 1344.33        | 1344.78       | 2012.43       |              |
|                                | 90 | 1266.75        | 1268.75       | 1122.43       |              |

$\mathbf{X} \in \mathbb{R}_+^{100 \times 433 \times 200}$  is a randomly generated tensor. No. of Iterations was 15

# Comparison results (NTF)

| Algo      | r  | NTF.blockpivot | NTF.activeset |
|-----------|----|----------------|---------------|
| Time(sec) | 3  | 2.0558         | 3.9237        |
|           | 10 | 18.1932        | 40.0865       |

$\mathbf{X} \in \mathbb{R}_+^{1000 \times 234 \times 654}$  is a randomly generated tensor. No. of Iterations was 15

| Algo      | r   | SparseNTF.blockpivot | SparseNTF.activeset |
|-----------|-----|----------------------|---------------------|
| Time(sec) | 10  | 1.4868               | 2.9211              |
|           | 50  | 10.0558              | 21.9914             |
|           | 100 | 58.1854              | 90.3214             |

Sparse NTF -  $\mathbf{X} \in \mathbb{R}_+^{173 \times 234 \times 854}$  is a randomly generated tensor. No. of Iterations was 20

# Summary

---

- A new algorithm for NMF and its extension to NTF is proposed:  
ANLS framework + Block principal pivoting algorithm with improvements for multiple right-hand sides
- Utilize: long and thin structure
- Extensions for sparse/regularized NMF and NTF
- Outperform other algorithms in computational experiments
- Some NMF codes are available at
  - <http://www.cc.gatech.edu/~hpark/softwareNMF.html>
  - <http://www.cc.gatech.edu/~jingu/nmf/index.html>