Sparse NMF via Alternating Non-negativity Constrained Least Squares

Hyunsoo Kim and Haesun Park

{hskim,hpark}@cc.gatech.edu

College of Computing

Georgia Institute of Technology

Atlanta, GA 30332, USA

Nonnegative Matrix Factorization Workshop, NISS

Feb. 24, 2007

Outline

- NMF as a Dimension Reduction/Clustering method
- NMF Algorithm via Alternating Least Squares and Convergence
  - Algorithms for Non-negativity Constraint Least Squares (NLS) Single right hand side vs. Multiple right hand sides
  - Other algorithms
- Sparse NMF via Alternating Least Squares and Convergence
- Applications : Microarray Analaysis ...

Dimension Reduction

- Unsupervised Dimension Reduction
  - SVD (LSI, PCA)
  - Nonnegative Matrix Factorization (NMF)
  - One-sided Nonnegative Matrix Factorization
- Dimension Reduction for Clustered Data
  - Linear Discriminant Analysis (LDA/GSVD)
  - Orthogonal Centroid Method (OCM)
  - Centroid-based Method
  - Nonnegativity constraint Centroid-based Method
  - NMF/initialization with centroid method

Nonnegativity Constraints?

Better Approximation vs. Better Representation/Interpretation
Given $A : m \times n$ and $k < \min(m, n)$

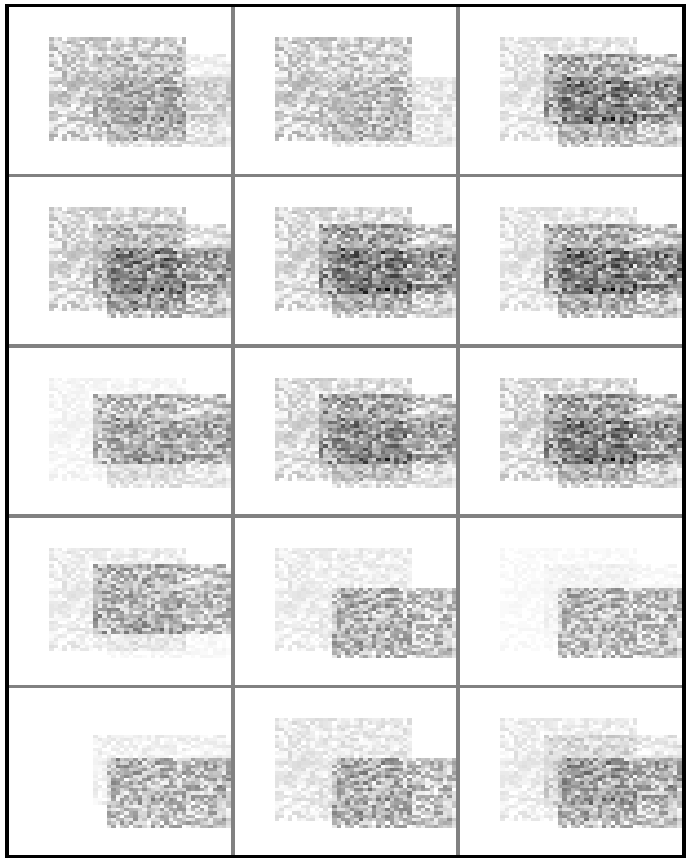
- SVD: Best Approximation
  Find $(W : m \times k)$ and $(H : k \times n)$ s.t. $A \approx WH$
  $\rightarrow \min \|A - WH\|_{2,F}$, $A = U\Sigma V^T$, $A \approx U_k \Sigma_k V_k^T$

- NMF: Better Representation/Interpretation?
  Find $(W : m \times k) \geq 0$ and $(H : k \times n) \geq 0$ s.t. $A \approx WH$
  $\rightarrow \min \|A - WH\|_F$ where $W \geq 0$ and $H \geq 0$

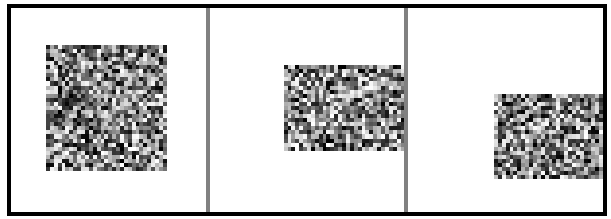
- Non-negative constraints are physically meaningful.
  - Pixels in digital image $\rightarrow$ Biomedical image processing
  - Molecule concentration in bioinformatics (*e.g.* mRNA, protein, miRNA, *etc.*) $\rightarrow$ Microarray data analysis
  - Signal intensities in mass spectrometry $\rightarrow$ Computational Proteomics
- Interpretation of analysis results: non-subtractive combinations of non-negative vectors.

A Test on an Artificial Data

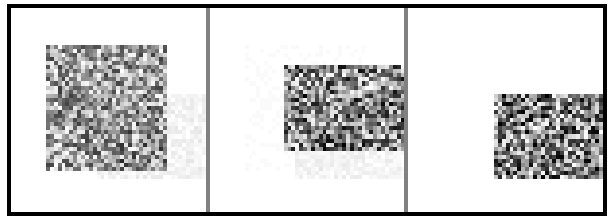
(a) Artificial dataset A



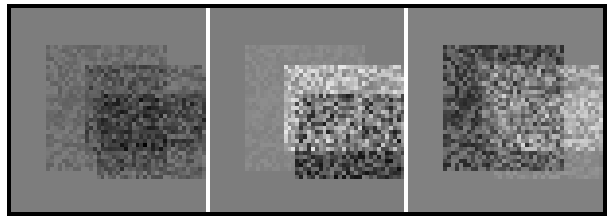
(b) Actual W



(c) W from NMF/ANLS



(d) W from SVD



NMF/ANLS on the artificial dataset $A = WH$.

Zeros: white, Positive values: darker

But in (d), Negative values: black, Zero: gray, Positive values white

Non-negative Matrix Factorization (NMF)

- Given a non-negative matrix $A : m \times n$ and a desired rank k, NMF solves:

$$\min_{W,H} \|A - WH\|_F, \;\; s.t. \;\; W \geq 0 \;\text{ and }\; H \geq 0$$

- $W \in \mathbb{R}^{m \times k}$: basis matrix, related to dimension reducing transformation
  $A \approx WH \rightarrow f(W)A \approx H$
  in SVD, $A \approx U_k \Sigma_k V_k^T \rightarrow U_k^T A \approx \Sigma_k V_k$
  Sparse $\rightarrow$ Parts-based Basis Vectors?
- $H \in \mathbb{R}^{k \times n}$: encoding matrix, non-negative lower dimensional representation
- Sparseness, dimension reduction: computational efficiency (storage, speed)
- W and H not unique

NMF Algorithms (1)

- Multiplicative update rules
  - Lee and Seung, Nature 1999 $\rightarrow$ Brunet, et. al., PNAS 2004 (showed that NMF performs better than HC and SOM)
  - nsNMF, Pascual-Montano et al., IEEE TPAMI 2006
- Alternating Least Squares (ALS)
  - Berry et al., Computational Statistics and Data Analysis, 2006
  - MUR+LS, Pauca et al., SDM 2004 $\rightarrow$ Gao and Church, Bioinformatics 2005
- Gradient Descent
  - Hoyer, JMLR 2004
  - Projected gradient ANLS, C. Lin, tech report, 2005

NMF Algorithms (2)

- Alternating Non-negativity Constrained Least Squares (ANLS)
  - Paatero and Tapper, 1994
  - NMF/ANLS, Kim and Park, ISBRA 2007, to appear
  - SNMF/ANLS, Kim and Park, Bioinformatics 2007, to appear
  - One-sided NMF, Park and Kim, SDM06, Textmining Workshop
- Others
  - Quasi-Newton optimization, Zdunek and Cichocki, ICAISC, 2006
  - Low Dimensional polytope approximation, M. Chu, draft Jan. 2007
  - Improved projected gradient ANLS (Newton+line search), S. Ingram
  - Newton-type ANLS, D. Kim, et al., SDM 2007, to appear

NMF/Alternating Least Squares (NMF/ANLS)

(Paatero and Tapper, Environmetrics, 1994)

1. Initialize $W \in \mathbb{R}^{m \times k}$ (or $H \in \mathbb{R}^{k \times n}$) with non-negative values, and scale the columns of W to unit L_2-norm.
2. Iterate the following ANLS until convergence:
   fixing W, solve $\min_{H \geq 0} \|WH - A\|_F$
   fixing H, solve $\min_{W \geq 0} \|H^T W^T - A^T\|_F$
3. The columns of W are normalized to unit L_2-norm at each iteration.

- Each NLS can be solved by MATLAB's LSQNONNEG (but DO NOT!)
- Lawson and Hanson 74, Active set method, for Single Right Hand Side, $\min_{h \geq 0} \|Wh - a\|_2$
- Faster algorithms exist for Multiple Right Hand Side problems:
  Bro and de Jong 97 (J. of Chemo.), for multi right hand sides
  Van Benthem and Keenan 04 (J. of Chemo.), further improvements

NMF/Multiplicative Update Rules (NMF/NUR)

(Lee and Seung, *Nature*, 1999)

- $\min_{W,H}(f(W,H) = \frac{1}{2}\|A - WH\|_F^2), W, H \geq 0$
  $\nabla_W f(W,H) = (WH - A)H^T, \quad \nabla_H f(W,H) = W^T(WH - A)$
- KKT Conditions:
  $W \geq 0, H \geq 0, \nabla_W f(W,H) \geq 0, \nabla_H f(W,H) \geq 0$
  $W_{il} \cdot \nabla_W f(W,H)_{il} = 0, H_{qj} \cdot \nabla_H f(W,H)_{qj} = 0$
- Alternating multiplicative update rules:

  $$H_{qj} \leftarrow H_{qj} \frac{(W^T A)_{qj}}{(W^T W H)_{qj} + \epsilon}, \quad 1 \leq q \leq k, \quad 1 \leq j \leq n,$$

  $$W_{iq} \leftarrow W_{iq} \frac{(AH^T)_{iq}}{(WHH^T)_{iq} + \epsilon}, \quad 1 \leq i \leq m, \quad 1 \leq q \leq k, 0 < \epsilon \ll 1$$

- If $W_{iq}^{(k+1)} = W_{iq}^{(k)} > 0$ and $(W^{(k)} H^{(k)} H^{(k)^T})_{iq} \neq 0$,
  then $(\nabla_W f(W^{(k)}, H^{(k)})_{iq} = 0$
  $\|A - WH\|_F$ is monotonically non-increasing

Convergence of NMF/ANLS

- Block Coordinate Descent method in Bound-constrained Optimization
- $\min_{W,H} \|A - WH\|_F^2, \quad s.t. \;\; W, H \geq 0$
- Given $A \in \mathbb{R}^{m \times n}$, NMF/ANLS iteratively solves

$$\min(f(W,H) = \|WH - A\|_F^2)$$

  fixing W with constraint $H \geq 0$ and fixing H with constraint $W \geq 0$.

- For k=1,2,...

  $W^{(k+1)} \in arg\min_W f(W, H^{(k)})$

  $H^{(k+1)} \in arg\min_H f(W^{(k+1)}, H)$

- No matter how many blocks, if the sub problems have unique solutions, then the limit point of the sequence is a stationary point (Powell 73, Bertsekas 99)
- For two block problems, any limit point of the sequence is a stationary point (Grippo and Siandrone, 00)

NLS with Multiple Right Hand Side Vectors

- Assume $W : m \times k$ and $A : m \times n$ with $m > k$ are Given.
- LS-S: $\min_h \|Wh - a\|_F$
- LS-M: $\min_H \|WH - A\|_F$
  Extremely inefficient if LS-S is solved n times independently
  W needs to be processed only once (e.g. compute SVD of W only once)
- NLS-S: $\min_{h \geq 0} \|Wh - a\|_F$ (Lawson and Hanson 74)
  Active set method: initially $h = 0$, $S_a = \{1, \cdots, k\}$, $S_p =$ null
  Each step solves $\min \|W^{(p)} h^{(p)} - a\|_2$
- NLS-M: $\min_{H \geq 0} \|WH - A\|_F$
  - Apply NLS-S n times? Inefficient!
  - Bro and de Jong 97:
    Compute $W^T W$ and $W^T A$ only once in $W^T W H = W^T A$
  - Van Benthem and Keenan 04:
    Initialization of active set based on LS-M
    Rearrange computation to be column parallel, e.g., $k = 3, n = 4$
    $S_{P1} = \{\{3\}, \{3\}, \{3\}, \{1\}\}$
    $S_{P2} = \{\{2, 3\}, \{1, 3\}, \{2, 3\}, \{1, 3\}\}$
    $S_{P3} = \{\{2, 3\}, \{1, 3\}, \{1, 2, 3\}, \{1, 2, 3\}\}$

Constrained NMF (CNMF) for Sparse NMF

(Pauca *et al., LAA*, 2006, Pauca *et al.*, *SDM*, 2004; Gao and Church, *Bioinformatics*, 2005)

- $\min_{W,H}\{\|A - WH\|_F^2 + \alpha\|W\|_F^2 + \beta\|H\|_F^2\}, \;\; s.t. \;\; W, H \geq 0$
- Multiplicative updating rules:

  $$H_{qj} \leftarrow H_{qj} \frac{(W^T A)_{qj} - \beta H_{qj}}{(W^T W H)_{qj} + \epsilon}, \quad 1 \leq q \leq k, \quad 1 \leq j \leq n,$$

  $$W_{iq} \leftarrow W_{iq} \frac{(AH^T)_{iq} - \alpha W_{iq}}{(WHH^T)_{iq} + \epsilon}, \quad 1 \leq i \leq m, \quad 1 \leq q \leq k,$$

  $\alpha \geq 0$ and $\beta \geq 0$ balance between approximation and sparseness

- Set negative values to zero for imposing non-negativity $\rightarrow$ not LS sol.
- L_1-norm based formulations recommended to control sparsity (Tibshirani, *J. Roy. Statist. Soc. B,* 1996)

Sparse NMF using L_1-norm (SNMF/R)

(Kim and Park 2007, Bioinformatics)

- SNMF/L (sparse W) and SNMF/R (sparse H)
- $\min_{W,H}(\|A - WH\|_F^2 + \eta\|W\|_F^2 + \beta\sum_{j=1}^{n}\|H(:,j)\|_1^2), W, H \geq 0$
  $\min_{W,H}(\|A - WH\|_F^2 + \eta\|W\|_F^2 + \beta\sum_{j=1}^{n}(\sum_{i=1}^{k}H(i,j))^2), W, H \geq 0$
- Initialize W with nonnegative values
- Iterate the following ANLS until convergence:

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

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

- $\beta > 0$ and $\eta > 0$ balance between
  accuracy of approximation and sparseness of H.
- Two-block coordinate-descent method.
  Any limit point is a stationary point.

Performance Comparison on Leukemia Data Set

Leukemia Data:$5,000 \times 38$, 3 clusters

Algorithms	NMF/NUR	NMF/ANLS
#($W = 0$) (%)	2.72%*	2.71%
#($H = 0$) (%)	17.28%*	18.42%
Purity	0.974	0.974
Entropy	0.095	0.095
# of iterations	3806	91.5
Computing time	159.2 sec.	7.1 sec.

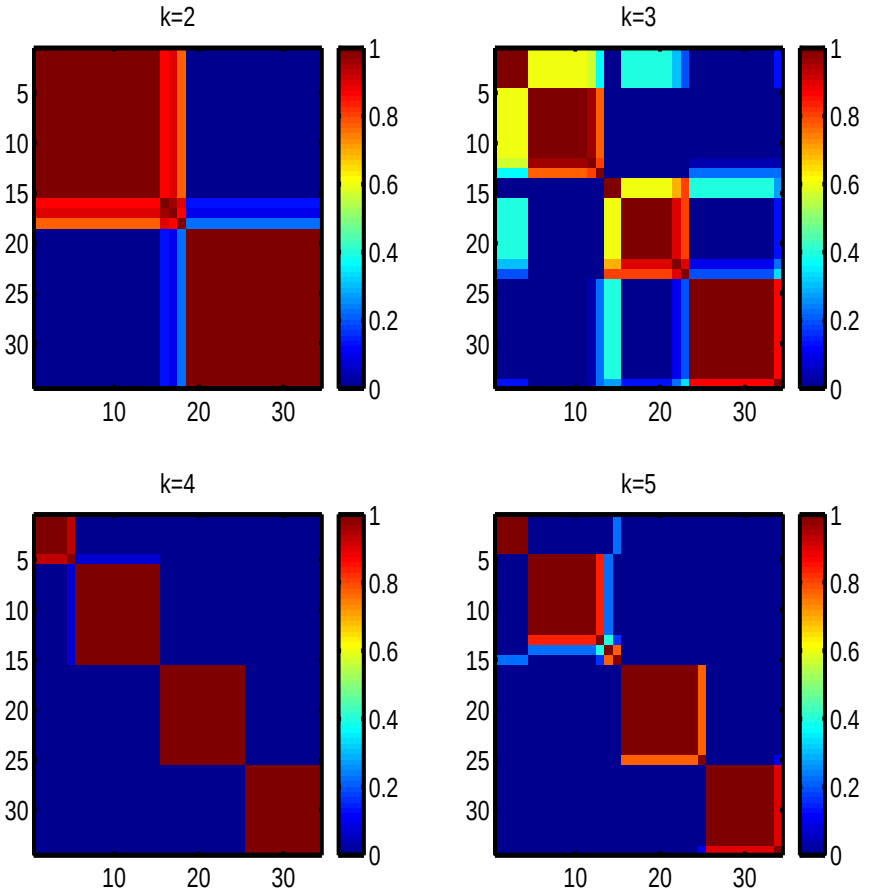
$k = 3$, average of 30 runs. Purity and entropy computed from H with the lowest approximation error. *The average percentages non-negative elements that are smaller than 10^{-8} in magnitude.

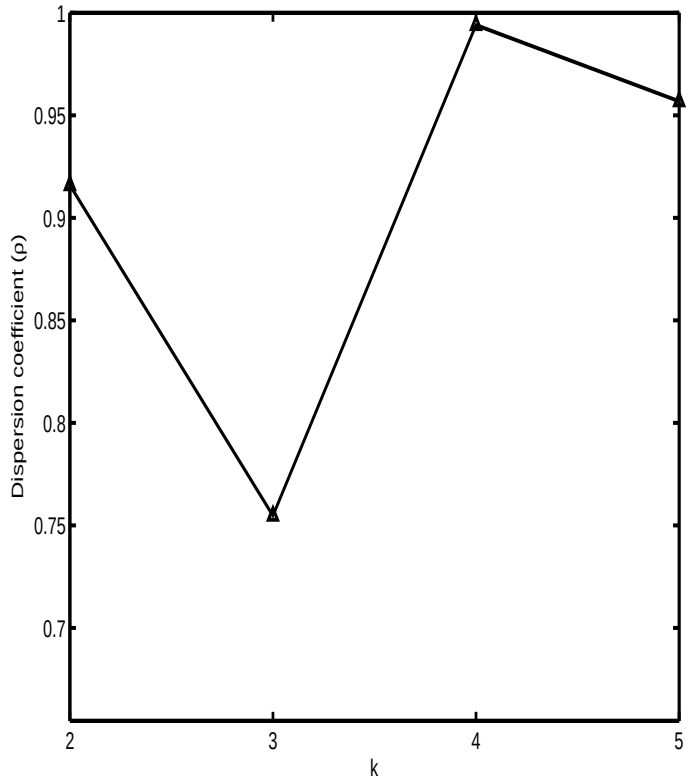
Performance Comparison on CNS Tumor Data Set

Algorithm	NMF/NUR		
k	3	4	5
$\#(W = 0)$ (%)	8.77%*	9.07%*	12.60%*
$\#(H = 0)$ (%)	16.99%*	24.14%*	25.43%*
# of iterations	11151	13770	16717
Computing time	563.5 sec.	836.4 sec.	1334.9 sec.
Algorithm	NMF/ANLS		
k	3	4	5
$\#(W = 0)$ (%)	8.69%	9.03%	12.54%
$\#(H = 0)$ (%)	18.63%	25.00%	26.88%
# of iterations	105.2	100.3	130.5
Computing time	9.8 sec.	12.1 sec.	20.3 sec.

Average of 30 runs. Central Nerve System tumors: four distinct morphologies: 10 classic medulloblastomas, 10 malignant gliomas, 10 rhabdoids and 4 normals. (Brunet *et al.*, *PNAS*, 2004. [2]Pomeroy *et al.*, *Nature*, 2002.)

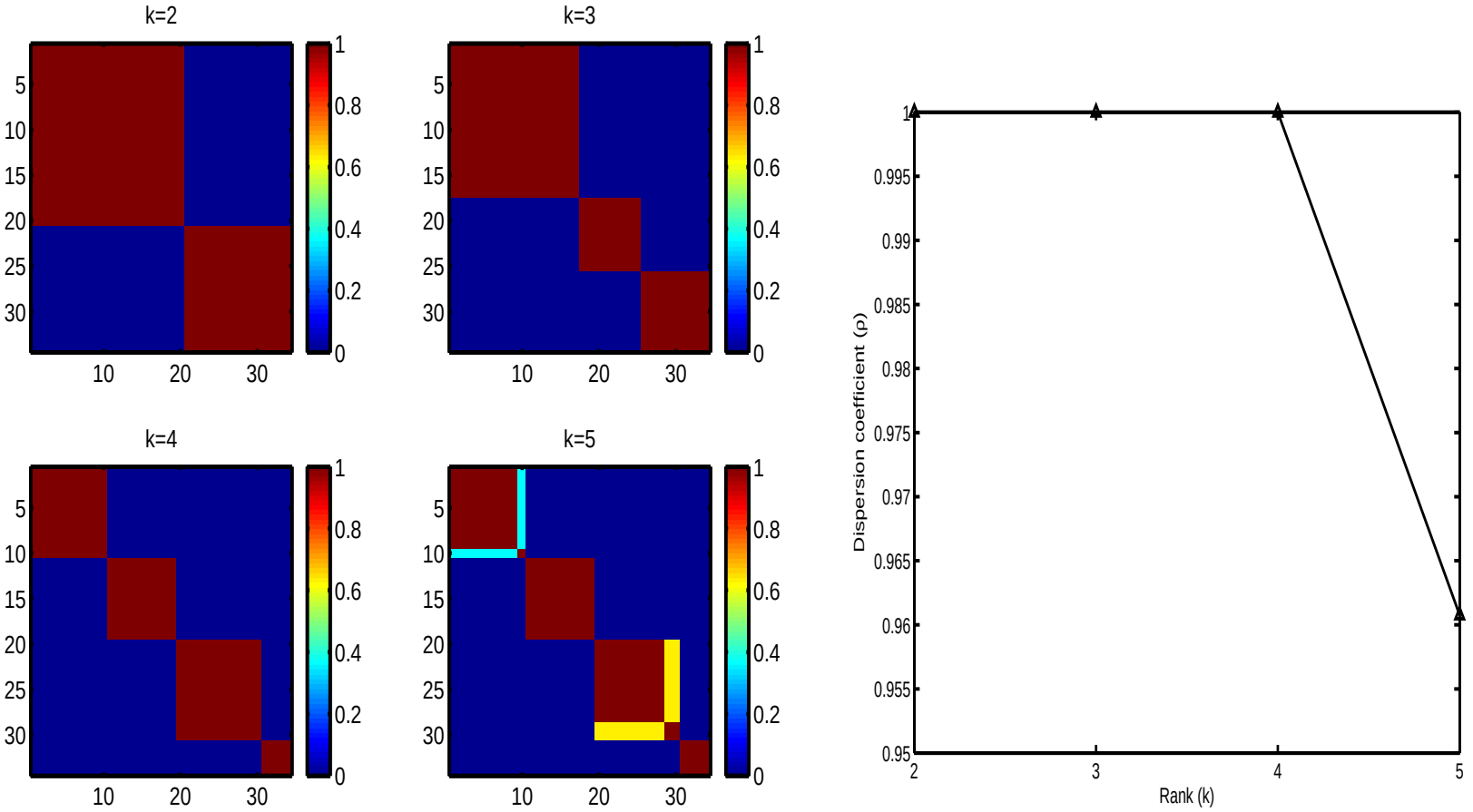
CNS Tumors Clustering by NMF/DUR





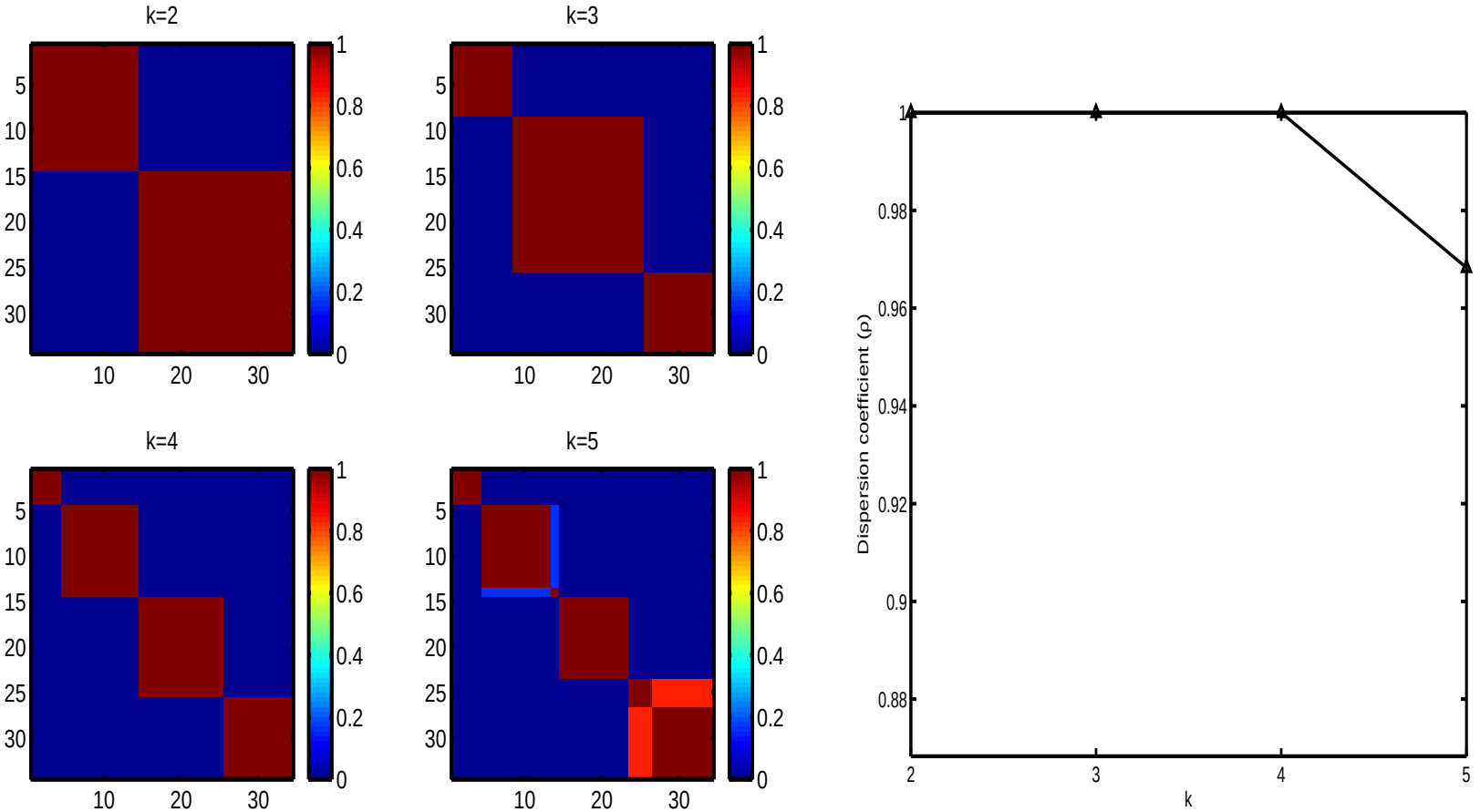
Reordered consensus matrices on the CNS dataset and the corresponding dispersion coefficients $\rho = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 4(C_{ij} - \frac{1}{2})^2$

CNS Tumors Clustering by NMF/ANLS



Reordered consensus matrices on the CNS dataset and the corresponding dispersion coefficients.

CNS Tumors Clustering by SNMF/R



Reordered consensus matrices on the CNS dataset and the corresponding dispersion coefficients.

SNMF/R ($k = 3$) on leukemia data: $5,000 \times 38$

Leukemia	NMF/DUR	SNMF/R			
β	-	0.001	0.01	0.1	0.5
#($W = 0$) (%)	0.10%*	2.43%	2.17%	1.57%	1.09%
#($H = 0$) (%)	0.00%*	24.56%	30.70%	44.74%	51.75%
Purity	0.953	0.974	0.974	0.947	0.921
Entropy	0.141	0.095	0.095	0.158	0.210
# of iterations	502.0	328.0	139.0	77.0	95.0
Computing time	53.6	40.1	17.0	9.4	10.9

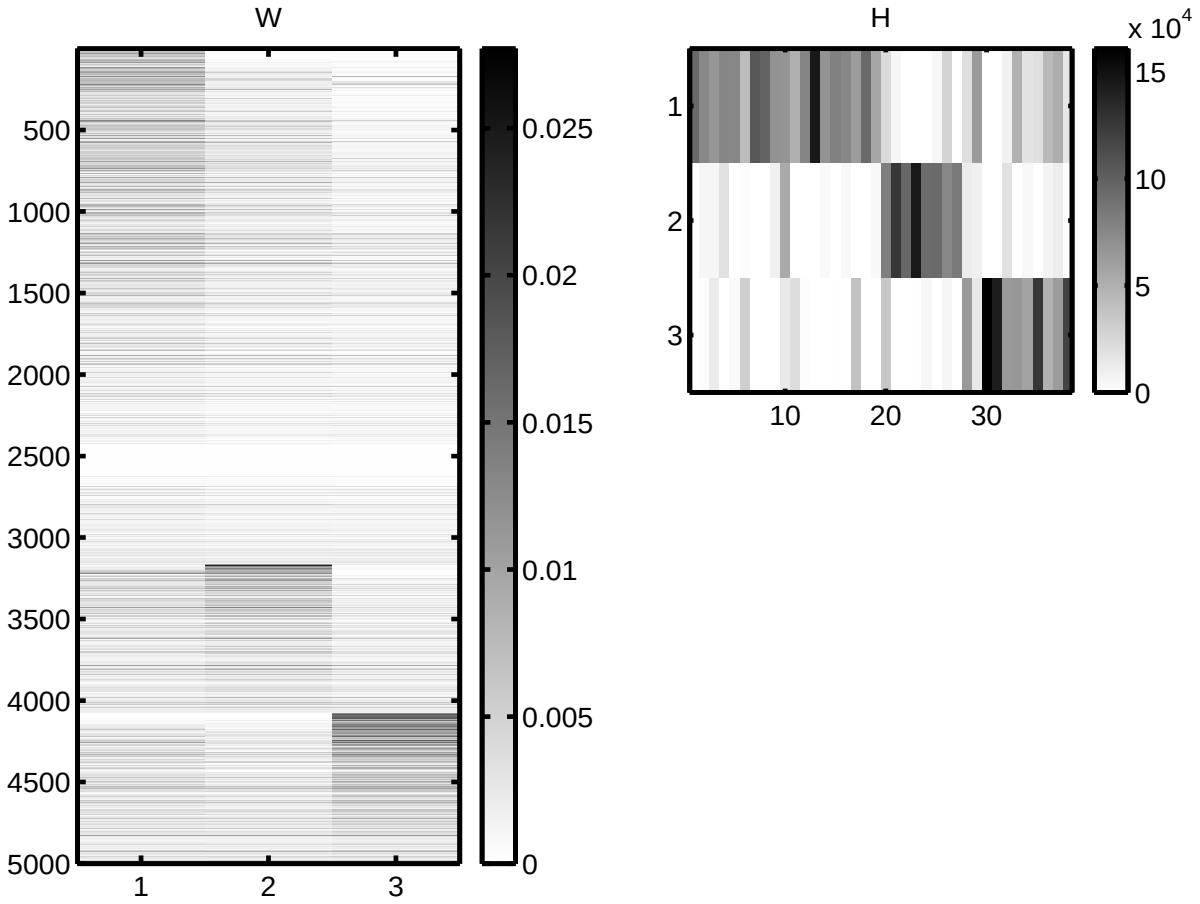
*For NMF using divergence-based multiplicative update rules (NMF/DUR) the average percentages the non-negative elements that are smaller than 10^{-8} in magnitude.

SNMF/R ($k = 4$) on CNS tumors data: $5,597 \times 34$

CNS tumors	NMF/DUR	SNMF/R			
β	-	0.01	0.1	1.0	2.0
#($W = 0$) (%)	1.65%*	8.45%	7.45%	5.06%	4.31%
#($H = 0$) (%)	1.47%*	25.74%	28.68%	36.76%	41.91%
Purity	0.941	0.971	0.971	0.971	0.941
Entropy	0.122	0.071	0.071	0.071	0.144
# of iterations	566.0	319.0	174.0	134.0	103.0
Computing time	63.4	51.6	29.5	20.9	16.0

*For NMF using divergence-based multiplicative update rules (NMF/DUR) the average percentages non-negative elements smaller than 10^{-8} in W and H.

W and H from SNMF/R



Leukemia dataset: $5,000 \times 38$, (38 samples: 19 ALL-B, 8 ALL-T, 11 AML)

Summary

- NMF as a Dimension Reduction/Clustering method
- NMF Algorithm via Alternating Least Squares and Convergence
  - Algorithms for Non-negativity Constraint Least Squares (NLS) Single right hand side vs. Multiple right hand sides
  - Other algorithms
- Sparse NMF via Alternating Least Squares and Convergence
- Imposing Constraints only on One Factor: Sparsity, Nonnegativity
- Applications : Gene clustering ...

Thank you!