Joint NMF for Hybrid Clustering based on Content and Connection Structure

Rundong Du[1], Barry Drake[2,3] and Haesun Park[2]

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

Georgia Institute of Technology, Atlanta, GA

SIAM ALA

HKBU, May, 2018

Supported in part by

Constrained Low Rank Approximations for Scalable Data Analytics

Objectives:

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

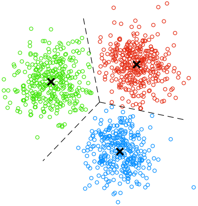
Goal: Orders of magnitude speed improvements over existing data analytics methods and solutions of higher quality

Why CLRA ?

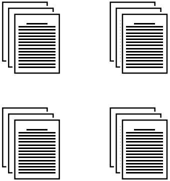
- Utilize advances in numerical linear algebra and optimization
- Exploit software such as BLAS and LAPACK
- Behavior of algorithms easier to analyze
- Facilitates design of MPI based algorithms for scalable solutions
- Can easily be modified for various problem demands, e.g. adaptive methods

Clustering: data clustering, topic modeling, graph clustering, community detection, hybrid clustering...

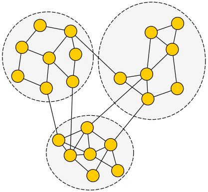
Vector space



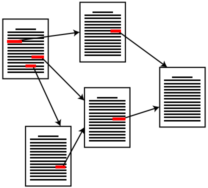
Text (Topic Modeling)

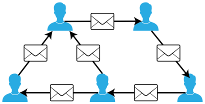


Graph (Community Detection)



Hybrid





Nonnegative Matrix Factorization (NMF)

(Lee&Seung 99, Paatero&Tapper 94)

Given $X \in {\mathbb{R}_+}^{m \times n}$ 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. $X \approx WH$.
Notation: $\mathbb{R}_+$: nonnegative real numbers

- $\min_{W \geq 0, H \geq 0} \|X - WH\|_F$
- Nonconvex

NMF for Clustering?

Objective functions for K-means and NMF may look the same:
$\min \sum_i \|\mathbf{x}_i - \mathbf{w}_{\sigma_i}\|_2^2 = \min \|X - WH\|_F^2$

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

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

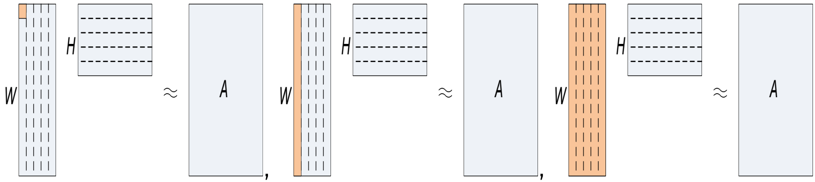
But, the constraints are different:

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

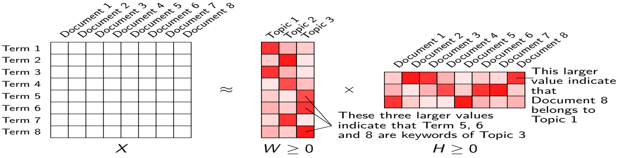
Block Coordinate Descent (BCD) for NMF

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

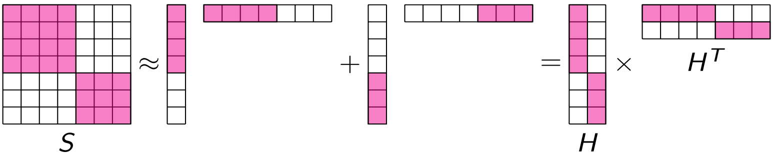
- **BCD** generates $z^{(k+1)} = (z_1^{(k+1)}, \ldots, z_p^{(k+1)})$ by
  $z_i^{(k+1)} = \arg\min_{\xi \in Z_i} f(z_1^{(k+1)}, \ldots, z_{i-1}^{(k+1)}, \xi, z_{i+1}^{(k)}, \ldots, z_p^{(k)})$
- **Th. (Bertsekas, 99):** Suppose f is continuously differentiable over the Cartesian product of closed, convex sets $Z_1, Z_2, \ldots, Z_p$ and for each i, the minimum is uniquely attained. Then every limit point of the sequence generated by the BCD method $\{z^{(k)}\}$ is a stationary point.



NMF for text clustering: (J. Kim and HP, SISC 11; J.Kim, Y. He, and HP, JOGO 14)

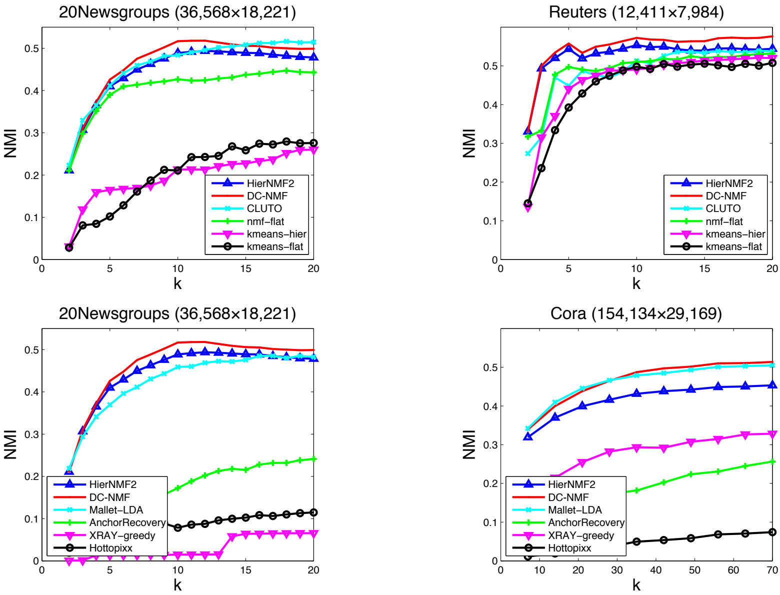


SymNMF for graph clustering: (D. Kuang, S. Yun, and HP, JOGO 15)



Input	Eigenbasis	Nonnegative basis
Feature-Data matrix	**SVD/PCA**	**NMF/Affine NMF**
Data-Data matrix	**Spectral clustering**	**SymNMF**

NMF Performance for Clustering and Topic Modeling



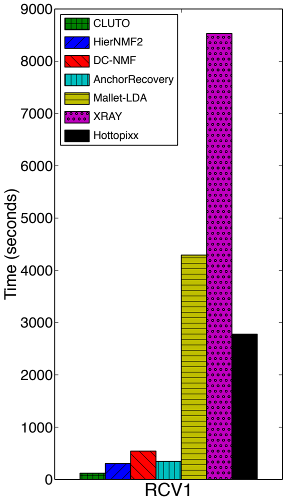
Source: R. Du, D. Kuang, B. Drake, HP, to appear in JOGO

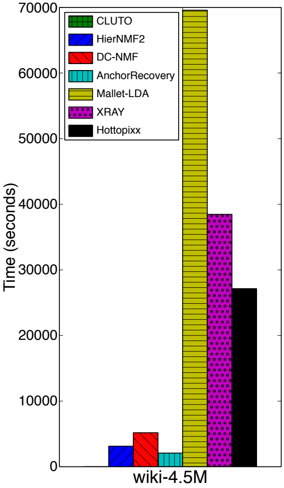
Methods Compared:

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

Data size (# of topics):

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





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

SmallK http://smallk.github.io

JointNMF from NMF and SymNMF

$\min_{W\geq 0, H\geq 0} \|X - WH\|_F$
NMF: content/text clustering
$X \in \mathbf{R}_+^{m\times n}$: term $\times$ doc

$\min_{H\geq 0} \|S - H^T H\|_F$
SymNMF: graph clustering
$S \in \mathbf{R}_+^{n\times n}$: doc $\times$ doc, $S^T = S$

$W \in \mathbf{R}_+^{m\times k}$, $H \in \mathbf{R}_+^{k\times n}$, $k << min\{m, n\}$

JointNMF for Hybrid Clustering:
$\min_{W\geq 0, H\geq 0} \alpha_1 \|X - WH\|_F^2 + \alpha_2 \|S - H^T H\|_F^2$

JointNMF and Block Coordinate Descent (BCD)

Formulation:

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

Recast for the BCD framework:

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

3-block coordinate descent:

- Solve W: $\min_{W\geq 0} \|H^T W^T - X^T\|_F$
- Solve $\tilde{H}$: $\min_{\tilde{H}\geq 0} \left\| \begin{bmatrix} \sqrt{\alpha} H^T \\ \sqrt{\beta} I_k \end{bmatrix} \tilde{H} - \begin{bmatrix} \sqrt{\alpha} S \\ \sqrt{\beta} H \end{bmatrix} \right\|_F$
- Solve H: $\min_{H\geq 0} \left\| \begin{bmatrix} W \\ \sqrt{\alpha}\widetilde{H}^T \\ \sqrt{\beta} I_k \end{bmatrix} H - \begin{bmatrix} X \\ \sqrt{\alpha} S \\ \sqrt{\beta}\widetilde{H} \end{bmatrix} \right\|_F$

Data source: PatentsView (`www.patentsview.org`)

- Full text of the claims of 5,915,134 granted US patents (1976-2016).
- 80,728,766 citations between those patents
- 233,111 ground truth clusters
- We selected 13 subgroups

F1 score when compared to ground truth:

F_1 score for comparing clusters $\{A_1, \ldots, A_k\}$ and $\{B_1, \ldots, B_{k'}\}$:

$$F_1 = \frac{1}{2}\left(\frac{1}{k}\sum_{i=1}^{k}\max_j F_1(A_i, B_j) + \frac{1}{k'}\sum_{j=1}^{k'}\max_i F_1(B_j, A_i)\right).$$

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

PCL-DC-1 and PCL-DC-2: hybrid clustering method Yang, Jin, Chi, Zhu, KDD 2009.

JointNMF:

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

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

Citation prediction for a new document $\boldsymbol{x}$:

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

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

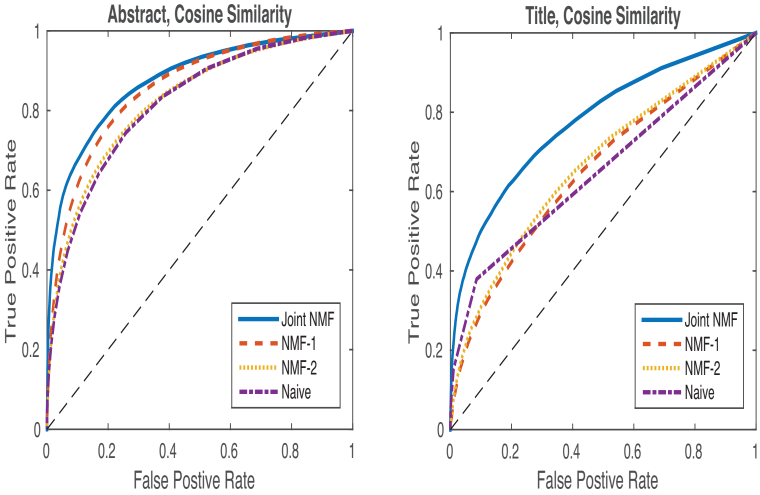
Baseline methods:

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

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

Naive: count number of words shared by two documents

Citation Prediction: Tests on cit-HepTh Data Set*



* Data source: SNAP (`http://snap.stanford.edu/data/`)

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

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

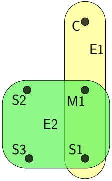
Case of Email Data: Content and Link Info Representations

Email 1

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

Email 2

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



$$\begin{array}{c} \\ C \\ M1 \\ S1 \\ S2 \\ S3 \\ \vdots \end{array}\begin{array}{c} \begin{array}{ccc} E1 & E2 & \cdots \end{array} \\ \begin{bmatrix} 1 & 0 & \cdots \\ 1 & 1 & \cdots \\ 1 & 1 & \cdots \\ 0 & 1 & \cdots \\ 0 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}\end{array}$$

M: Incidence matrix
$S = D_e^{-1/2} M^T D_v^{-1} M D_e^{-1/2}$?
D_v and D_e: vertex and hyperedge degrees

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

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

Other representation:

- Keep the incidence matrix M (person-email relation)
- Construct similarity matrix for email-email relationship using email content and construct corresponding normalized graph laplacian L.
- Solve $\min_{W,H} \|M - WH\|_F^2 + \lambda \operatorname{tr}(HLH^\intercal)$

Case Study: Enron Email Data Set

Frequency of number of memberships

#memberships	1	2	3	4	5	6	7	11
#employees	1069	149	45	17	8	7	1	1

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

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

Topic keywords of clusters

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

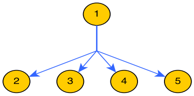
Data source: a subset of 1702 emails from the Enron Email data set, extracted by a group from SIMS, UC Berkeley.

Representation of a Hypergraph

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

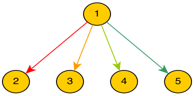
hypergraph | graph

directed



hyper-edge, node

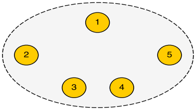
$$M = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \end{array}\begin{bmatrix} -1 & \cdots \\ 1 & \cdots \\ 1 & \cdots \\ 1 & \cdots \\ 1 & \cdots \\ \vdots & \ddots \end{bmatrix}$$



$M_- M_+^\intercal$

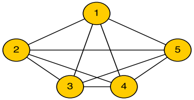
$$(M_-)_{ij} = [M_{ij} < 0]$$
$$(M_+)_{ij} = [M_{ij} > 0]$$
$$[\mathcal{P}] = \begin{cases} 1, & \text{if statement } \mathcal{P} \text{ is true;} \\ 0, & \text{otherwise.} \end{cases}$$

undirected



hyper-edge, node

$$|M| = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \end{array}\begin{bmatrix} 1 & \cdots \\ 1 & \cdots \\ 1 & \cdots \\ 1 & \cdots \\ 1 & \cdots \\ \vdots & \ddots \end{bmatrix}$$



$|M||M|^\intercal$

smallk.github.io: our XDATA Open Catalog Software

- Goals: Develop fast and effective software for the variants of NMF with usability and extensibility as key design features
- Application to real-world large-scale data analytics problems

Implementation

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

Documentation and Tutorials

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

Summary

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

References

- H. Kim and HP, Sparse NMF via Alternating Non-negativity-constrained Least Squares for Microarray Data Analysis. Bioinfo., 23(12):1495-1502, 2007.
- H. Kim and HP, NMF Based on Alter. NLS and the Active Set Method. SIAM Jour. on Matrix Analysis and Applic. (SIMAX), 30(2):713-730, 2008.
- J. Kim and HP, Fast NMF: an Active-set-like Method and Comparisons, SIAM Journal on Scientific Computing (SISC), 33(6):3261-3281, 2011.
- J. Kim, Y. He, and HP, Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework, Journal of Global Optimization (JOGO), 58(2):285-319, 2014.
- D. Kuang and HP, Fast rank-2 NMF for hierarchical document clustering, Proc. of the 19th ACM SIGKDD (KDD), pp.739-747, 2013.
- J. Choo, C. Lee, C. Reddy, and HP, UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization, IEEE Trans. on Vis. and Computer Graphics (TVCG), 19-12:1992-2001, 2013
- J. Choo and HP, Screen space- and perception-based framework for efficient computational algorithms in large-scale visual analytics, IEEE Computer Graphics and Applications, Special Issue - Big Data Vis., 33-4:22-28, 2013.
- D. Kuang, S. Yun, and HP, *SymNMF: Nonnegative low-rank approximation of a similarity matrix for graph clustering*, Journal of Global Optimization (JOGO), 62(3):545-574, 2015.
- N. Gillis, D. Kuang, and HP, *Hierarchical Clustering of Hyperspectral Images Using Rank-Two Nonnegative Matrix Factorization*, IEEE Transactions on Geoscience and Remote Sensing (TGRS), 53(4):2066-2078, 2015.
- R. Du, D. Kuang, B. Drake, and HP, *DC-NMF: nonnegative matrix factorization based on divide-and-conquer for fast clustering and topic modeling*, Journal of Global Optimization (JOGO), 68(4):777-798, 2017.
- R. Du, D. Kuang, B. Drake, and HP, *Hierarchical Community Detection via Rank-2 Symmetric Nonnegative Matrix Factorization*, to appear.

Thank you!