

# Exploring Nonnegative Matrix Factorization

Holly Jin

LinkedIn Corp

and

Michael Saunders

Systems Optimization Laboratory, Stanford University

MMDS08

Workshop on Algorithms for Modern Massive Data Sets

Stanford University, June 25–28, 2008

# Outline

- 1 Introduction
- 2 SNMF motivation
- 3 Sparse NMF
- 4 Basis Pursuit DeNoising (BPDN)
- 5 SNMF results
- 6 Application examples

# Introduction

Many applications (signal processing, page rank, imaging)  
seek *sparse solutions to square or rectangular systems*

$$Ax \approx b \quad x \text{ sparse}$$

# Introduction

Many applications (signal processing, page rank, imaging) seek *sparse solutions to square or rectangular systems*

$$Ax \approx b \quad x \text{ sparse}$$

Basis Pursuit Denoising (BPDN) seeks sparsity by solving

$$\min \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1$$

# Introduction

Many applications (signal processing, page rank, imaging) seek *sparse solutions to square or rectangular systems*

$$Ax \approx b \quad x \text{ sparse}$$

**Basis Pursuit Denoising (BPDN)** seeks sparsity by solving

$$\min \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1$$

**Sparse Nonnegative Matrix Factorization (SNMF)** involves square or rectangular systems

$$A \approx WH$$

$$W, H \geq 0 \quad \text{low rank and sparse}$$

## Introduction

Many applications (signal processing, page rank, imaging) seek *sparse solutions to square or rectangular systems*

$$Ax \approx b \quad x \text{ sparse}$$

**Basis Pursuit Denoising (BPDN)** seeks sparsity by solving

$$\min \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1$$

**Sparse Nonnegative Matrix Factorization (SNMF)** involves square or rectangular systems

$$A \approx WH$$

$$W, H \geq 0 \quad \text{low rank and sparse}$$

Perhaps **BPDN** can find *very sparse approximate*  $W, H$

# Motivation for Sparse NMF Solver

# An example

- Grouping similar items in a grocery store

## An example

- Grouping similar items in a grocery store
- Customer check-out receipts:

	Flour	Balloon	Beer	Sugar	Chip
Customer1	0	3	8	0	1
Customer2	0	2	8	1	0
Customer3	5	0	1	10	0
Customer4	0	20	40	2	1
Customer5	10	0	1	10	1

## Extract features using SVD

$$U = \begin{pmatrix} -0.1852 & -0.0225 & 0.1457 & 0.9394 & -0.2480 \\ -0.1179 & 0.0282 & -0.1676 & 0.2529 & 0.9451 \\ -0.0338 & 0.6126 & -0.7649 & 0.0793 & -0.1794 \\ -0.9744 & -0.0492 & -0.0030 & -0.2098 & -0.0645 \\ -0.0356 & 0.7881 & 0.6046 & -0.0570 & 0.0945 \end{pmatrix}$$

$$S = \begin{pmatrix} 45.9457 & 0 & 0 & 0 & 0 \\ 0 & 17.7720 & 0 & 0 & 0 \\ 0 & 0 & 2.9418 & 0 & 0 \\ 0 & 0 & 0 & 1.1892 & 0 \\ 0 & 0 & 0 & 0 & 0.2783 \end{pmatrix}$$

$$V = \begin{pmatrix} -0.0114 & 0.6158 & 0.7550 & -0.1455 & 0.1719 \\ -0.4414 & -0.0560 & 0.0144 & -0.7337 & -0.5134 \\ -0.8949 & -0.0341 & 0.0164 & 0.3441 & 0.2818 \\ -0.0601 & 0.7842 & -0.6042 & 0.0478 & -0.1191 \\ -0.0260 & 0.0403 & 0.2540 & 0.5656 & -0.7831 \end{pmatrix}$$

## Truncated SVD

- Choose  $k = 2$  principal features according to matrix  $S$ :

$$U_k = \begin{pmatrix} -0.1852 & 0.0225 \\ -0.1179 & -0.0282 \\ -0.0338 & -0.6126 \\ -0.9744 & 0.0492 \\ -0.0356 & -0.7881 \end{pmatrix}$$

$$S_k = \begin{pmatrix} 45.9457 & 0 \\ 0 & 17.7720 \end{pmatrix}$$

$$V_k = \begin{pmatrix} -0.0114 & -0.6158 \\ -0.4414 & 0.0560 \\ -0.8949 & 0.0341 \\ -0.0601 & -0.7842 \\ -0.0260 & -0.0403 \end{pmatrix}$$

## Truncated SVD

- Choose  $k = 2$  principal features according to matrix  $S$ :

$$U_k = \begin{pmatrix} -0.1852 & 0.0225 \\ -0.1179 & -0.0282 \\ -0.0338 & -0.6126 \\ -0.9744 & 0.0492 \\ -0.0356 & -0.7881 \end{pmatrix}$$

$$S_k = \begin{pmatrix} 45.9457 & 0 \\ 0 & 17.7720 \end{pmatrix}$$

$$V_k = \begin{pmatrix} -0.0114 & -0.6158 \\ -0.4414 & 0.0560 \\ -0.8949 & 0.0341 \\ -0.0601 & -0.7842 \\ -0.0260 & -0.0403 \end{pmatrix}$$

- Error  $\|A - USV^T\| = 1.5e-14$   
vs  $\|A - U_k S_k V_k^T\| = 2.9$  (minimized Frobenius norm)

## Clustering and ranking

- Row and column clustering and rankings:

$$Rr = \begin{pmatrix} (1, 1) & 2 \\ (2, 1) & 3 \\ (4, 1) & 1 \\ (3, 2) & 2 \\ (5, 2) & 1 \end{pmatrix} \quad Rc = \begin{pmatrix} (2, 1) & 2 \\ (3, 1) & 1 \\ (1, 2) & 2 \\ (4, 2) & 1 \\ (5, 2) & 3 \end{pmatrix}$$

## Clustering and ranking

- Row and column clustering and rankings:

$$Rr = \begin{pmatrix} (1, 1) & 2 \\ (2, 1) & 3 \\ (4, 1) & 1 \\ (3, 2) & 2 \\ (5, 2) & 1 \end{pmatrix} \quad Rc = \begin{pmatrix} (2, 1) & 2 \\ (3, 1) & 1 \\ (1, 2) & 2 \\ (4, 2) & 1 \\ (5, 2) & 3 \end{pmatrix}$$

- Order of k preserves the ranking of cluster importance

## Clustering and ranking

- Row and column clustering and rankings:

$$R_r = \begin{pmatrix} (1, 1) & 2 \\ (2, 1) & 3 \\ (4, 1) & 1 \\ (3, 2) & 2 \\ (5, 2) & 1 \end{pmatrix} \quad R_c = \begin{pmatrix} (2, 1) & 2 \\ (3, 1) & 1 \\ (1, 2) & 2 \\ (4, 2) & 1 \\ (5, 2) & 3 \end{pmatrix}$$

- Order of k preserves the ranking of cluster importance
- Cluster example meanings:

	Flour	Balloon	Beer	Sugar	Chip
Customer 1	0	3	8	0	1
Customer 2	0	2	8	1	0
Customer 3	5	0	1	10	0
Customer 4	0	20	40	2	1
Customer 5	10	0	1	10	1

## Clustering and ranking

- Row and column clustering and rankings:

$$R_r = \begin{pmatrix} (1, 1) & 2 \\ (2, 1) & 3 \\ (4, 1) & 1 \\ (3, 2) & 2 \\ (5, 2) & 1 \end{pmatrix} \quad R_c = \begin{pmatrix} (2, 1) & 2 \\ (3, 1) & 1 \\ (1, 2) & 2 \\ (4, 2) & 1 \\ (5, 2) & 3 \end{pmatrix}$$

- Order of k preserves the ranking of cluster importance
- Cluster example meanings:

	Flour	Balloon	Beer	Sugar	Chip
Customer 1	0	3	8	0	1
Customer 2	0	2	8	1	0
Customer 3	5	0	1	10	0
Customer 4	0	20	40	2	1
Customer 5	10	0	1	10	1

- Features extraction:

	Partying	Baking
Customers	1, 2, 4	3, 5
Products	Balloon, Beer	Flour, Sugar, Chip

# NMF

- $A = WH, \quad A \approx W_k H_k$

# NMF

- $A = WH$ ,  $A \approx W_k H_k$
- Factorization not unique: (via Chih-Jen Lin's NMF solver)

$$W_k = \begin{pmatrix} 0 & 1.2850 \\ 0.4711 & 0.8065 \\ 8.4380 & 0.0365 \\ 0.0217 & 6.7563 \\ 10.8476 & 0 \end{pmatrix}$$

$$H_k = \begin{pmatrix} 0.7968 & 0 & 0.0928 & 1.0214 & 0.0567 \\ 0 & 2.9321 & 5.9337 & 0.2885 & 0.1667 \end{pmatrix}$$

# NMF

- $A = WH$ ,  $A \approx W_k H_k$
- Factorization not unique: (via Chih-Jen Lin's NMF solver)

$$W_k = \begin{pmatrix} 0 & 1.2850 \\ 0.4711 & 0.8065 \\ 8.4380 & 0.0365 \\ 0.0217 & 6.7563 \\ 10.8476 & 0 \end{pmatrix}$$

$$H_k = \begin{pmatrix} 0.7968 & 0 & 0.0928 & 1.0214 & 0.0567 \\ 0 & 2.9321 & 5.9337 & 0.2885 & 0.1667 \end{pmatrix}$$

- Clustering and Ranking:

$$R_r = \begin{pmatrix} (1, 2) & 2 \\ (2, 2) & 3 \\ (4, 2) & 1 \\ (3, 1) & 2 \\ (5, 1) & 1 \end{pmatrix} \quad R_c = \begin{pmatrix} (2, 2) & 2 \\ (3, 2) & 1 \\ (5, 2) & 3 \\ (1, 1) & 2 \\ (4, 1) & 1 \end{pmatrix}$$

# NMF

- $A = WH$ ,  $A \approx W_k H_k$
- Factorization not unique: (via Chih-Jen Lin's NMF solver)

$$W_k = \begin{pmatrix} 0 & 1.2850 \\ 0.4711 & 0.8065 \\ 8.4380 & 0.0365 \\ 0.0217 & 6.7563 \\ 10.8476 & 0 \end{pmatrix}$$

$$H_k = \begin{pmatrix} 0.7968 & 0 & 0.0928 & 1.0214 & 0.0567 \\ 0 & 2.9321 & 5.9337 & 0.2885 & 0.1667 \end{pmatrix}$$

- Clustering and Ranking:

$$R_r = \begin{pmatrix} (1, 2) & 2 \\ (2, 2) & 3 \\ (4, 2) & 1 \\ (3, 1) & 2 \\ (5, 1) & 1 \end{pmatrix} \quad R_c = \begin{pmatrix} (2, 2) & 2 \\ (3, 2) & 1 \\ (5, 2) & 3 \\ (1, 1) & 2 \\ (4, 1) & 1 \end{pmatrix}$$

- Order of  $k$  does not preserve the ranking of cluster importance

# Sparse Nonnegative Factorization

# Sparse NMF

$$A \approx WH, \quad W, H \geq 0, \quad \text{low rank and sparse}$$

Kim and Park (2007)

$$\min_{W, H \geq 0} \frac{1}{2} \|A - WH\|_F^2 + \eta \|W\|_F^2 + \beta \sum \|h_j\|_1^2$$

# Sparse NMF

$$A \approx WH, \quad W, H \geq 0, \quad \text{low rank and sparse}$$

Kim and Park (2007)

$$\min_{W, H \geq 0} \frac{1}{2} \|A - WH\|_F^2 + \eta \|W\|_F^2 + \beta \sum \|h_j\|_1^2$$

Alternating nonnegative least squares (NLS) on

$$\min_{W \geq 0} \left\| \begin{pmatrix} H^T \\ \sqrt{\eta} I \end{pmatrix} W^T - \begin{pmatrix} A^T \\ 0 \end{pmatrix} \right\|_F^2, \quad \min_{H \geq 0} \left\| \begin{pmatrix} W \\ \sqrt{\beta} e^T \end{pmatrix} H - \begin{pmatrix} A \\ 0 \end{pmatrix} \right\|_F^2$$

Sparse  $H$

# Sparse NMF via BPDN

$$\min_{W, H \geq 0} \frac{1}{2} \|A - WH\|_F^2 + \beta \sum \|w_i\|_1 + \eta \sum \|h_j\|_1$$

# Sparse NMF via BPDN

$$\min_{W, H \geq 0} \frac{1}{2} \|A - WH\|_F^2 + \beta \sum \|w_i\|_1 + \eta \sum \|h_j\|_1$$

Alternating BPDN on

$$\min_{W \geq 0} \frac{1}{2} \|H^T W^T - A^T\|^2 + \eta \sum \|w_i\|_1$$

$$\min_{H \geq 0} \frac{1}{2} \|WH - A\|^2 + \beta \sum \|h_j\|_1$$

Sparse  $W$  and  $H$

# L1<sup>2</sup> or L1?

- Kim and Park (2007):

$$\min_{x \geq 0} \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1^2$$

$x \rightarrow 0$  only as  $\lambda \rightarrow \infty$

(Nevertheless, Kim and Park report sparse solutions with moderate  $\lambda$ )

# L1<sup>2</sup> or L1?

- Kim and Park (2007):

$$\min_{x \geq 0} \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1^2$$

$x \rightarrow 0$  only as  $\lambda \rightarrow \infty$

(Nevertheless, Kim and Park report sparse solutions with moderate  $\lambda$ )

- BPDN:

$$\min_{x \geq 0} \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1$$

$x = 0$  for  $\lambda \geq \|A^T b\|_\infty$

$x$  very sparse for  $\lambda = 0.9 \|A^T b\|_\infty$  say

Easy to control the sparsity of  $x$

## SNMF via BPDN implementation

$$\min_{U,D,V \geq 0} \frac{1}{2} \|A - UDV^T\|_F^2 + \sum \beta_i \|Du_i\|_1 + \sum \eta_j \|Dv_j\|_1$$

Alternating BPDN on

$$\min_{v_j \geq 0} \frac{1}{2} \|Uv_j - a_j\|^2 + \eta_j \|v_j\|_1, \quad \text{normalize } V \rightarrow VD$$

$$\min_{u_i \geq 0} \frac{1}{2} \|Vu_i - a_i\|^2 + \beta_i \|u_i\|_1, \quad \text{normalize } U \rightarrow UD$$

# SNMF via BPDN implementation

$$\min_{U, D, V \geq 0} \frac{1}{2} \|A - UDV^T\|_F^2 + \sum \beta_i \|Du_i\|_1 + \sum \eta_j \|Dv_j\|_1$$

Alternating BPDN on

$$\min_{v_j \geq 0} \frac{1}{2} \|Uv_j - a_j\|^2 + \eta_j \|v_j\|_1, \quad \text{normalize } V \rightarrow VD$$

$$\min_{u_i \geq 0} \frac{1}{2} \|Vu_i - a_i\|^2 + \beta_i \|u_i\|_1, \quad \text{normalize } U \rightarrow UD$$

- $\eta_j \leq \sigma \|U^T a_j\|_\infty$   
 $\beta_i \leq \sigma \|V^T a_i\|_\infty$

$\sigma$  = "sparsity" input parameter  
 = 0.9 or 0.8 say

# SNMF via BPDN implementation

$$\min_{U,D,V \geq 0} \frac{1}{2} \|A - UDV^T\|_F^2 + \sum \beta_i \|Du_i\|_1 + \sum \eta_j \|Dv_j\|_1$$

Alternating BPDN on

$$\min_{v_j \geq 0} \frac{1}{2} \|Uv_j - a_j\|^2 + \eta_j \|v_j\|_1, \quad \text{normalize } V \rightarrow VD$$

$$\min_{u_i \geq 0} \frac{1}{2} \|Vu_i - a_i\|^2 + \beta_i \|u_i\|_1, \quad \text{normalize } U \rightarrow UD$$

- $\eta_j \leq \sigma \|U^T a_j\|_\infty$   
 $\beta_i \leq \sigma \|V^T a_i\|_\infty$
- $\sigma =$  "sparsity" input parameter  
 $= 0.9$  or  $0.8$  say
- At some point, freeze  $D$  (Also  $\eta_j \beta_i$  stop changing)

# BPDN solvers

# BPDN solvers

$$\min \lambda \|x\|_1 + \frac{1}{2} \|Ax - b\|^2$$

OMP	Davis, Mallat et al 1997	Greedy
BPDN-interior	Chen, Donoho & S, 1998, 2001	Interior, <b>CG</b>
PDSCO, PDCO	Saunders 1997, 2002	Interior, <b>LSQR</b>
BCR	Sardy, Bruce & Tseng 2000	Orthogonal blocks
Homotopy	Osborne et al 2000	Active-set, all $\lambda$
LARS	Efron, Hastie, Tibshirani 2004	Active-set, all $\lambda$
STOMP	Donoho, Tsaig, et al 2006	Double greedy
l1_ls	Kim, Koh, Lustig, Boyd et al 2007	Primal barrier, <b>PCG</b>
GPSR	Figueiredo, Nowak & Wright 2007	Gradient Projection
SPGL1	van den Berg & Friedlander 2007	Spectral GP, all $\lambda$
BPdual	Friedlander & Saunders 2007	Active-set on dual
LPdual	Friedlander & Saunders 2007	Active-set on dual, $x \geq 0$
lsNMF, lsNTF	Friedlander & Hatz 2007	Sparse NMF <i>and</i> NTF (BCLS subproblem solver)

## LPdual solver

Active-set method for *dual* of regularized LP:

$$\min_{x,y} e^T x + \frac{1}{2} \lambda \|y\|^2 \quad \text{st} \quad Ax + \lambda y = b, \quad x \geq 0$$

$$\min_y -b^T y + \frac{1}{2} \lambda \|y\|^2 \quad \text{st} \quad A^T y \leq e$$

## LPdual solver

Active-set method for *dual* of regularized LP:

$$\min_{x,y} e^T x + \frac{1}{2} \lambda \|y\|^2 \quad \text{st} \quad Ax + \lambda y = b, \quad x \geq 0$$

$$\min_y -b^T y + \frac{1}{2} \lambda \|y\|^2 \quad \text{st} \quad A^T y \leq e$$

$B \equiv$  columns of  $A$  for active constraints ( $B^T y = e$ )

Initially  $y = 0$ ,  $B$  empty

Selects columns of  $B$  in mostly *greedy* manner

## LPdual solver

Active-set method for *dual* of regularized LP:

$$\min_{x,y} e^T x + \frac{1}{2} \lambda \|y\|^2 \quad \text{st} \quad Ax + \lambda y = b, \quad x \geq 0$$

$$\min_y -b^T y + \frac{1}{2} \lambda \|y\|^2 \quad \text{st} \quad A^T y \leq e$$

$B \equiv$  columns of  $A$  for active constraints ( $B^T y = e$ )

Initially  $y = 0$ ,  $B$  empty

Selects columns of  $B$  in mostly *greedy* manner

Main work per iteration:

Solve  $\min \|Bx - g\|$

Form  $dy = (g - Bx)/\lambda$

Form  $dz = A^T dy$

Add or delete a column of  $B$

# SNMF Results

## Sparse NMF example

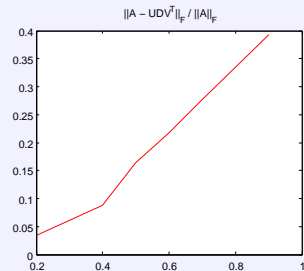
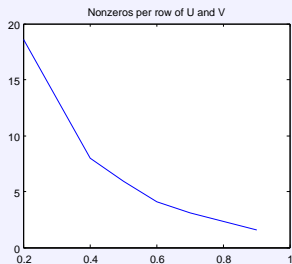
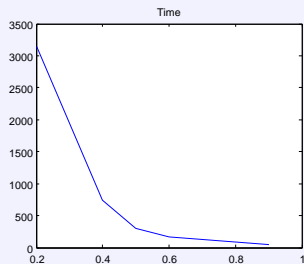
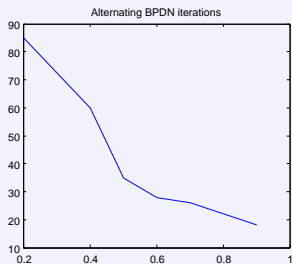
- Sparse solution:  $k = 2$

$$U_k = \begin{pmatrix} 0.1859 & & & \\ 0.1170 & & & \\ & 0.6146 & & \\ 0.9756 & & & \\ & & 0.7889 & \end{pmatrix} \quad V_k = \begin{pmatrix} & 0.6153 & & \\ 0.4428 & & & \\ 0.8963 & & & \\ & 0.7877 & & \\ 0.0253 & 0.0303 & & \end{pmatrix}$$
$$D_k = \begin{pmatrix} 27.51 & & & \\ & 10.69 & & \end{pmatrix}$$





$m = n = 450$ ,  $k = 200$ , increasing sparsity



# Real Application Examples

# Keyword clusterings

- About 8000 stem terms

# Keyword clusterings

- About 8000 stem terms
- Create term similarity matrix  $A$

## Keyword clusterings

- About 8000 stem terms
- Create term similarity matrix  $A$
- Sample clusters:

googladword

c++

adword

cc++

googl

java

googlanalyt

c++java

yahoo

c++program

searchmarket

c++unix

omnitur

pascal

msn

c++develop

webtrend

c++programm

adbrit

javaprogram

## User input standardization

- Field-of-study user input, about 400k unique entries
- Cluster user inputs automatically
- Sample clusters:
  - Abbreviations, variation of the same word or typos
    - hr, human resources, hrm
    - film production, film, theatre, acting, theater
  - New words
    - physical therapy, kinesiology
  - Similar disciplines
    - materials science and engineering, materials science, materials engineering
  - Foreign language
    - business economics, bedrijfseconomie
    - bedrijfskundige informatica, business informatics, informatica
  - Noise elimination, or crowded cluster
    - business administration, business, mba, project management, master in business administration, business administration, master of business administration, technology, business admin, general education

# Thanks

- Michael Friedlander (BP solvers)

# Thanks

- Michael Friedlander (BP solvers)
- Sou-Cheng Choi  
Lek-Heng Lim

# Thanks

- Michael Friedlander (BP solvers)
- Sou-Cheng Choi  
Lek-Heng Lim
- Jay Kreps  
Jonathan Goldman  
Huitao Luo

# Thanks

- Michael Friedlander (BP solvers)
- Sou-Cheng Choi  
Lek-Heng Lim
- Jay Kreps  
Jonathan Goldman  
Huitao Luo
- Understanding Complex Datasets  
Data Mining with Matrix Decompositions  
(useful book by David Skillicorn)

# Thanks

- Michael Friedlander (BP solvers)
- Sou-Cheng Choi  
Lek-Heng Lim
- Jay Kreps  
Jonathan Goldman  
Huitao Luo
- Understanding Complex Datasets  
Data Mining with Matrix Decompositions  
(useful book by David Skillicorn)
- Michael Saunders