Text Mining

using the

Nonnegative Matrix Factorization

Amy Langville

Carl Meyer

Department of Mathematics
North Carolina State University
Raleigh, NC

SIAM-SEAS–Charleston 3/25/2005

Outline

Traditional IR

- Vector Space Model (1960s and 1970s)
- Latent Semantic Indexing (1990s)
- Nonnegative Matrix Factorization (2000)

Vector Space Model (1960s and 1970s)

Gerard Salton's Information Retrieval System

SMART: System for the Mechanical Analysis and Retrieval of Text

(Salton's Magical Automatic Retriever of Text)

- turn n textual documents into n document vectors $\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n$
- create term-by-document matrix $\mathbf{A}_{m \times n} = [\, \mathbf{d}_1 | \mathbf{d}_2 | \cdots | \mathbf{d}_n \,]$
- to retrieve info., create query vector $\mathbf{q}$, which is a pseudo-doc

Vector Space Model (1960s and 1970s)

Gerard Salton's Information Retrieval System

SMART: System for the Mechanical Analysis and Retrieval of Text

(Salton's Magical Automatic Retriever of Text)

- turn n textual documents into n document vectors $\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n$
- create term-by-document matrix $\mathbf{A}_{m \times n} = [\, \mathbf{d}_1 | \mathbf{d}_2 | \cdots | \mathbf{d}_n \,]$
- to retrieve info., create query vector $\mathbf{q}$, which is a pseudo-doc

GOAL: find doc. $\mathbf{d}_i$ closest to $\mathbf{q}$

— angular cosine measure used: $\delta_i = \cos\theta_i = \mathbf{q}^T \mathbf{d}_i / (\|\mathbf{q}\|_2 \|\mathbf{d}_i\|_2)$

Example from Berry's book

Terms	Documents
T1: Bab(y,ies,y's)	D1: Infant & Toddler First Aid
T2: Child(ren's)	D2: Babies & Children's Room (For Your Home)
T3: Guide	D3: Child Safety at Home
T4: Health	D4: Your Baby's Health & Safety : From Infant to Toddler
T5: Home	D5: Baby Proofing Basics
T6: Infant	D6: Your Guide to Easy Rust Proofing
T7: Guide	D7: Beanie Babies Collector's Guide
T8: Safety	
T9: Toddler	

Example from Berry's book

Terms

T1: Bab(y,ies,y's)
T2: Child(ren's)
T3: Guide
T4: Health
T5: Home
T6: Infant
T7: Guide
T8: Safety
T9: Toddler

Documents

D1: Infant & Toddler First Aid
D2: Babies & Children's Room (For Your Home)
D3: Child Safety at Home
D4: Your Baby's Health & Safety : From Infant to Toddler
D5: Baby Proofing Basics
D6: Your Guide to Easy Rust Proofing
D7: Beanie Babies Collector's Guide

$$\mathbf{A} = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \quad \mathbf{q} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \end{bmatrix} = \begin{bmatrix} 0 \\ .5774 \\ 0 \\ .8944 \\ .7071 \\ 0 \\ .7071 \end{bmatrix}$$

Strengths and Weaknesses of VSM

Strengths

- **A** is sparse
- $\mathbf{q}^T\mathbf{A}$ is fast and can be done in parallel
- relevance feedback: $\tilde{\mathbf{q}} = \delta_1\mathbf{d}_1 + \delta_3\mathbf{d}_3 + \delta_7\mathbf{d}_7$

Weaknesses

- synonyms and polysems—noise in **A**
- decent performance
- basis vectors are standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$, which are orthogonal $\Rightarrow$ independence of terms

Latent Semantic Indexing (1990s)

Susan Dumais's improvement to VSM = LSI

Idea: use low-rank approximation to **A** to filter out noise

- Great Idea! 2 patents for Bell/Telcordia
  - Computer information retrieval using latent semantic structure. U.S. Patent No. 4,839,853, June 13, 1989.
  - Computerized cross-language document retrieval using latent semantic indexing. U.S. Patent No. 5,301,109, April 5, 1994.

(Resource: USPTO `http://patft.uspto.gov/netahtml/srchnum.htm`)

SVD

$\mathbf{A}_{m\times n}$: rank r term-by-document matrix

- SVD: $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\,\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$
- LSI: use $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ in place of $\mathbf{A}$
- Why?
  - reduce storage when $k << r$
  - filter out uncertainty, so that performance on text mining tasks (e.g., query processing and clustering) improves

What's Really Happening?

Change of Basis

using truncated SVD $\mathbf{A}_k = \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^T$

- Original Basis: docs represented in Term Space using Standard Basis $S = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m\}$
- New Basis: docs represented in smaller Latent Semantic Space using Basis $B = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ $(k << \min(m,n))$

$$\begin{array}{c} \\ nonneg. \\ entries \end{array} \overset{doc_1}{\begin{pmatrix} \vdots \\ \mathbf{A}_{*1} \\ \vdots \end{pmatrix}}_{m\times 1} \approx \begin{bmatrix} \vdots \\ \mathbf{u}_1 \\ \vdots \end{bmatrix} \sigma_1 v_{11} + \begin{bmatrix} \vdots \\ \mathbf{u}_2 \\ \vdots \end{bmatrix} \sigma_2 v_{12} + \cdots + \begin{bmatrix} \vdots \\ \mathbf{u}_k \\ \vdots \end{bmatrix} \sigma_k v_{1k}$$

What's Really Happening?

Change of Basis

using truncated SVD $\mathbf{A}_k = \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^T$

- Original Basis: docs represented in Term Space using Standard Basis $S = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m\}$
- New Basis: docs represented in smaller Latent Semantic Space using Basis $B = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ (k<<min(m,n))

$$\begin{array}{c} \\ nonneg. \\ entries \end{array} \overset{doc_1}{\begin{pmatrix} \vdots \\ \mathbf{A}_{*1} \\ \vdots \end{pmatrix}}_{m\times 1} \approx \begin{bmatrix} \vdots \\ \mathbf{u}_1 \\ \vdots \end{bmatrix} \sigma_1 v_{11} + \begin{bmatrix} \vdots \\ \mathbf{u}_2 \\ \vdots \end{bmatrix} \sigma_2 v_{12} + \cdots + \begin{bmatrix} \vdots \\ \mathbf{u}_k \\ \vdots \end{bmatrix} \sigma_k v_{1k}$$

- still use angular cosine measure

$$\delta_i = cos\,\theta_i = \mathbf{q}^T\mathbf{d}_i/(\|\mathbf{q}\|_2\|\mathbf{d}_i\|_2) = \mathbf{q}^T\mathbf{A}_k\mathbf{e}_i/(\|\mathbf{q}\|_2\|\mathbf{A}_k\mathbf{e}_i\|_2)$$

$$= \mathbf{q}^T\mathbf{U}_k\boldsymbol{\Sigma}_k\mathbf{V}_k^T\mathbf{e}_i/(\|\mathbf{q}\|_2\|\boldsymbol{\Sigma}_k\mathbf{V}_k^T\mathbf{e}_i\|_2)$$

Properties of SVD

- basis vectors $\mathbf{u}_i$ are orthogonal
- u_{ij}, v_{ij} are mixed in sign

$$\begin{array}{ccccc} \mathbf{A}_k & = & \mathbf{U}_k & \Sigma_k & \mathbf{V}_k^T \\ nonneg & & mixed & nonneg & mixed \end{array}$$

- **U**, **V** are dense
- *uniqueness*—while there are many SVD algorithms, they all create the same (truncated) factorization
- of all rank-k approximations, $\mathbf{A}_k$ is optimal (in Frobenius norm)

$$\|\mathbf{A} - \mathbf{A}_k\|_F = \min_{rank(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_F$$

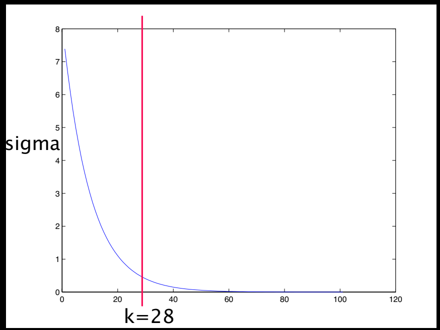
Strengths and Weaknesses of LSI

Strengths

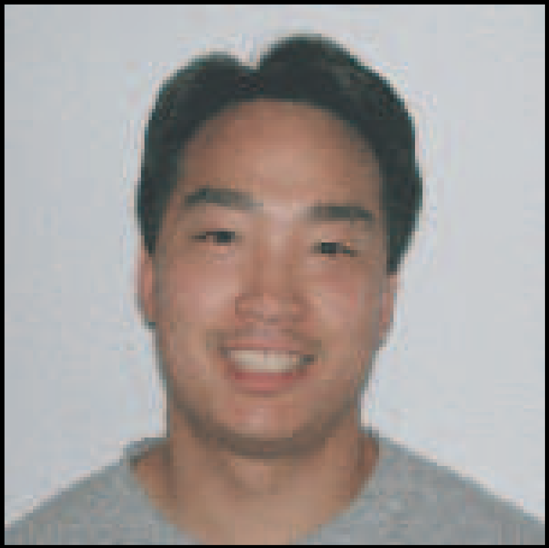
- using $\mathbf{A}_k$ in place of $\mathbf{A}$ gives improved performance
- dimension reduction considers only essential components of term-by-document matrix, filters out noise
- best rank-k approximation

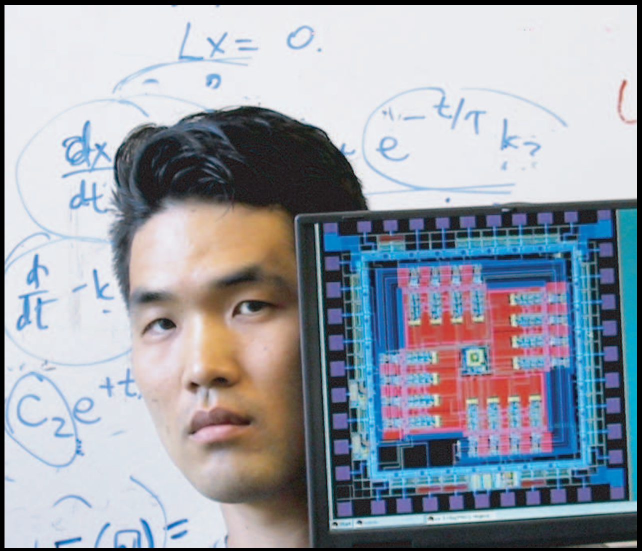
Weaknesses

- storage—$\mathbf{U}_k$ and $\mathbf{V}_k$ are usually completely dense
- interpretation of basis vectors $\mathbf{u}_i$ is impossible due to mixed signs
- good truncation point k is hard to determine
- orthogonality restriction



Nonnegative Matrix Factorization (2000)





Daniel Lee and Sebastian Seung's Nonnegative Matrix Factorization

Idea: use low-rank approximation with nonnegative factors to improve LSI

$$\underset{nonneg}{\mathbf{A}_k} = \underset{mixed}{\mathbf{U}_k} \; \underset{nonneg}{\Sigma_k} \; \underset{mixed}{\mathbf{V}_k^T}$$

$$\underset{nonneg}{\mathbf{A}_k} = \underset{nonneg}{\mathbf{W}_k} \; \underset{nonneg}{\mathbf{H}_k}$$

Better Basis for Text Mining

Change of Basis

using NMF $\mathbf{A}_k = \mathbf{W}_k\mathbf{H}_k$, where $\mathbf{W}_k,\ \mathbf{H}_k \geq \mathbf{0}$

- Use of NMF: replace $\mathbf{A}$ with $\mathbf{A}_k = \mathbf{W}_k\mathbf{H}_k$ $(\mathbf{W}_k = [\,\mathbf{w}_1|\mathbf{w}_2|\ldots|\mathbf{w}_k\,])$
- New Basis: docs represented in smaller Topic Space using Basis $B = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\}$ $(k << \min(m,n))$

$$\begin{array}{cc} & doc_1 \\ \begin{array}{r} nonneg. \\ entries \end{array} & \begin{pmatrix} \vdots \\ \mathbf{A}_{*1} \\ \vdots \end{pmatrix}_{m\times 1} \end{array} \approx \begin{bmatrix} \vdots \\ \mathbf{w}_1 \\ \vdots \end{bmatrix} h_{11} + \begin{bmatrix} \vdots \\ \mathbf{w}_2 \\ \vdots \end{bmatrix} h_{21} + \cdots + \begin{bmatrix} \vdots \\ \mathbf{w}_k \\ \vdots \end{bmatrix} h_{k1}$$

Properties of NMF

- basis vectors $\mathbf{w}_i$ are not $\perp$ $\Rightarrow$ can have overlap of topics
- can restrict $\mathbf{W}$, $\mathbf{H}$ to be sparse
- $\mathbf{W}_k$, $\mathbf{H}_k$ ≥ 0 $\Rightarrow$ immediate interpretation (additive parts-based rep.)

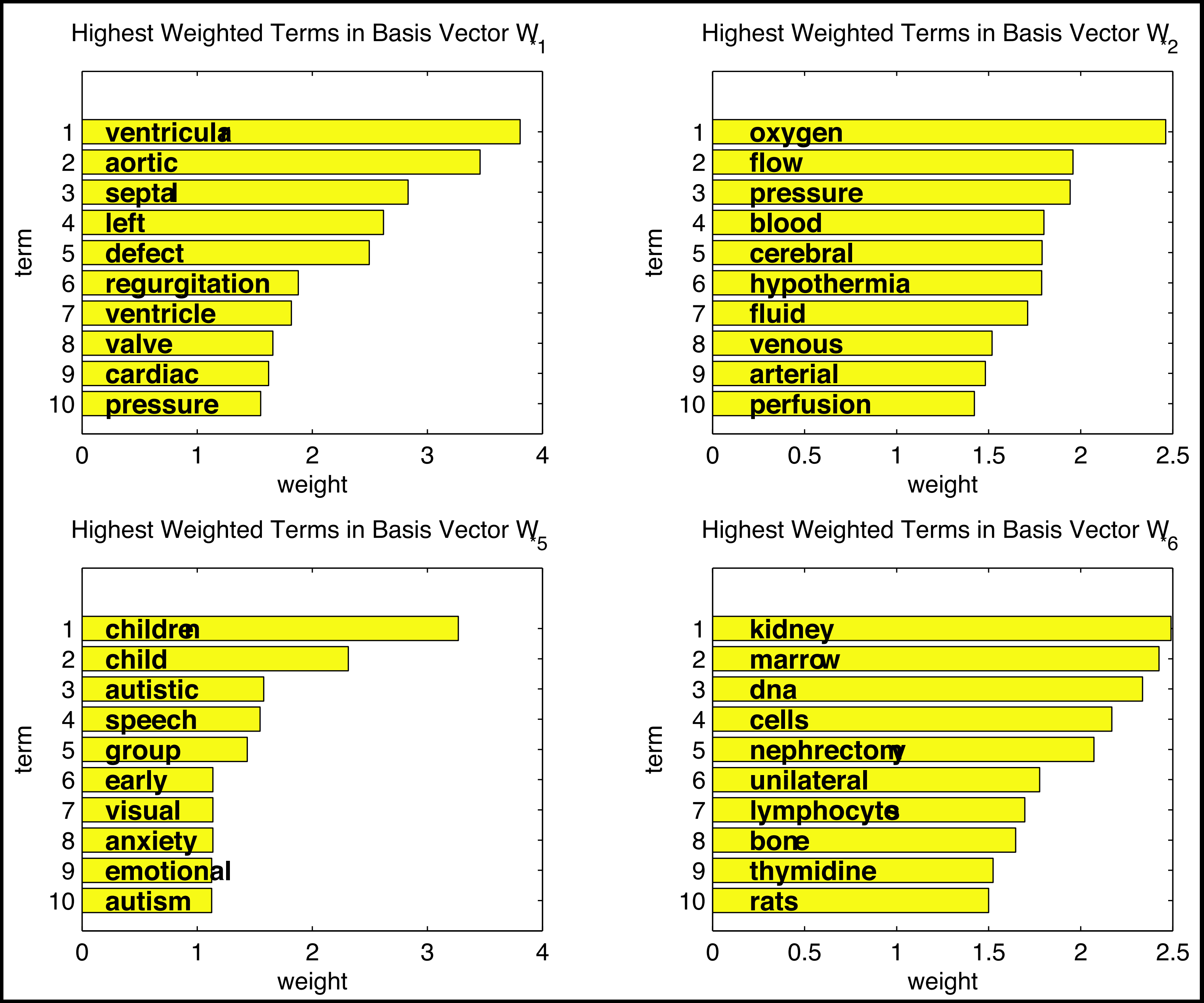
EX: large w_{ij}'s $\Rightarrow$ basis vector $\mathbf{w}_i$ is mostly about terms j

EX: h_{i1} how much doc_1 is pointing in the "direction" of topic vector $\mathbf{w}_i$

$$\mathbf{A}_k\mathbf{e}_1 = \mathbf{W}_k\mathbf{H}_{*1} = \begin{bmatrix} \vdots \\ \mathbf{w}_1 \\ \vdots \end{bmatrix} h_{11} + \begin{bmatrix} \vdots \\ \mathbf{w}_2 \\ \vdots \end{bmatrix} h_{21} + \cdots + \begin{bmatrix} \vdots \\ \mathbf{w}_k \\ \vdots \end{bmatrix} h_{k1}$$

Interpretation of Basis Vectors

MED dataset ($k = 10$)



Interpretation of Basis Vectors

MED dataset ($k = 10$)

$$\textbf{doc}_5 \approx \begin{array}{c} \mathbf{w}_9 \\ \begin{pmatrix} \text{fatty} \\ \text{glucose} \\ \text{acids} \\ \text{ffa} \\ \text{insulin} \\ \vdots \end{pmatrix} \end{array} .1646 + \begin{array}{c} \mathbf{w}_6 \\ \begin{pmatrix} \text{kidney} \\ \text{marrow} \\ \text{dna} \\ \text{cells} \\ \text{nephr.} \\ \vdots \end{pmatrix} \end{array} .0103 + \begin{array}{c} \mathbf{w}_7 \\ \begin{pmatrix} \text{hormone} \\ \text{growth} \\ \text{hgh} \\ \text{pituitary} \\ \text{mg} \\ \vdots \end{pmatrix} \end{array} .0045 + \cdots$$

NMF Literature

Papers report NMF is

- $\cong$ LSI for query processing

NMF Literature

Papers report NMF is

- $\cong$ LSI for query processing
- $\cong$ LSI for document clustering

NMF Literature

Papers report NMF is

- $\cong$ LSI for query processing
- $\cong$ LSI for document clustering
- $>$ LSI for interpretation of elements of factorization

NMF Literature

Papers report NMF is

- $\cong$ LSI for query processing
- $\cong$ LSI for document clustering
- $>$ LSI for interpretation of elements of factorization
- $>$ LSI potentially in terms of storage (sparse implementations)

NMF Literature

Papers report NMF is

- $\cong$ LSI for query processing
- $\cong$ LSI for document clustering
- $>$ LSI for interpretation of elements of factorization
- $>$ LSI potentially in terms of storage (sparse implementations)
- — most NLP algorithms require $O(kmn)$ computation per iteration

Computation of NMF

(Lee and Seung 2000)

MEAN SQUARED ERROR OBJECTIVE FUNCTION

$$\min \|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 \quad s.t. \quad \mathbf{W}, \mathbf{H} \geq 0$$

Nonlinear Optimization Problem

- convex in **W** or **H**, but not both $\Rightarrow$ can't get global min
- huge # unknowns: mk for **W** and kn for **H**

  (EX: $\mathbf{A}_{70K \times 1K}$ and k=10 topics $\Longrightarrow$ 800K unknowns)
- above objective is one of many possible
- convergence to local min only guaranteed for some algorithms

Computation of NMF

(Lee and Seung 2000)

MEAN SQUARED ERROR OBJECTIVE FUNCTION

$$\min \|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 \quad s.t. \quad \mathbf{W}, \mathbf{H} \geq 0$$

```
W = abs(randn(m,k));
H = abs(randn(k,n));
for i = 1 : maxiter
    H = H .* (W^T A) ./ (W^T WH + 10^-9);
    W = W .* (AH^T) ./ (WHH^T + 10^-9);
end
```

Many parameters affect performance (k, obj. function, sparsity constraints, algorithm, etc.).

— NMF is not unique!

NMF Algorithm: Berry et al. 2004

Gradient Descent–Constrained Least Squares

W = abs(randn(m,k)); (scale cols of **W** to unit norm)

H = zeros(k,n);

for i = 1 : maxiter

CLS for j = 1 : $\#docs$, solve

$$\min_{\mathbf{H}_{*j}} \|\mathbf{A}_{*j} - \mathbf{W}\mathbf{H}_{*j}\|_2^2 + \lambda\|\mathbf{H}_{*j}\|_2^2$$

$$\text{s.t. } \mathbf{H}_{*j} \geq \mathbf{0}$$

GD $\mathbf{W} = \mathbf{W}$.* $(\mathbf{A}\mathbf{H}^T)$ ./ $(\mathbf{W}\mathbf{H}\mathbf{H}^T + 10^{-9})$; (scale cols of **W**)

end

NMF Algorithm: Berry et al. 2004

GRADIENT DESCENT–CONSTRAINED LEAST SQUARES

W = abs(randn(m,k)); (scale cols of **W** to unit norm)

H = zeros(k,n);

for i = 1 : maxiter

CLS for j = 1 : $\#docs$, solve

$$\min_{\mathbf{H}_{*j}} \|\mathbf{A}_{*j} - \mathbf{W}\mathbf{H}_{*j}\|_2^2 + \lambda\|\mathbf{H}_{*j}\|_2^2$$

$$\text{s.t. } \mathbf{H}_{*j} \geq 0$$

solve $(\mathbf{W}^T\mathbf{W} + \lambda\,\mathbf{I})\,\mathbf{H} = \mathbf{W}^T\mathbf{A}$ for **H** (small $k \times k$ system solve)

GD $\mathbf{W} = \mathbf{W}$.* $(\mathbf{A}\mathbf{H}^T)$ ./ $(\mathbf{W}\mathbf{H}\mathbf{H}^T + 10^{-9})$; (scale cols of **W**)

end

— convergence to local min not guaranteed, but works well in practice

— objective function tails off after 15-30 iterations

Strengths and Weaknesses of NMF

Strengths

- Great Interpretability
- Performance for query processing/clustering comparable to LSI
- Sparsity of factorization allows for significant storage savings
- Scalability good as k, m, n increase
- possibly faster computation time than SVD

Weaknesses

- Factorization is not unique $\Rightarrow$ dependency on algorithm and parameters
- Convergence, when guaranteed, only to local min

Basis Vectors & Random Initialization

(gd-cls $\lambda = 2$, 50 iter. on REUTERS10)

W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
MIN	A=22658	seed=59							
+tonne	+billion	+share	stg	mln-mln	gulf	+dollar	+oil	+loss	+trade
+wheat	+year	+offer	+bank	cts	iran	+rate	opec	+profit	japan
+grain	+earn	+company	money	mln	+attack	+curr.	+barrel	oper	japanese
+crop	+qrtr	+stock	+bill	shr	+iranian	+bank	bpd	+exclude	+tariff
corn	+rise	+sharehol.	+market	+net	+ship	yen	crude	+net	+import
agricul.	pct	+common	england	avg	+tanker	monetary	+price	dlrs	reagan

Basis Vectors & Random Initialization

(gd-cls λ = 2, 50 iter. on REUTERS10)

W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
MIN	A=22658	seed=59							
+tonne	+billion	+share	stg	mln-mln	gulf	+dollar	+oil	+loss	+trade
+wheat	+year	+offer	+bank	cts	iran	+rate	opec	+profit	japan
+grain	+earn	+company	money	mln	+attack	+curr.	+barrel	oper	japanese
+crop	+qrtr	+stock	+bill	shr	+iranian	+bank	bpd	+exclude	+tariff
corn	+rise	+sharehol.	+market	+net	+ship	yen	crude	+net	+import
agricul.	pct	+common	england	avg	+tanker	monetary	+price	dlrs	reagan
AVER	A=22688	seed=1							
+tonne	+billion	+share	stg	+rate	analy.	+dollar	+oil	+loss	+trade
+wheat	+quarter	+offer	+bank	+bank	+market	+curr.	+barrel	cts	japan
+grain	+earn	+stock	money	+econom.	+sell	yen	opec	mln	japanese
+crop	+year	+company	+bill	+fed	+firm	+paris	bpd	+net	+tariff
corn	+rise	+common	london	+cut	+business	japan	crude	shr	+import
usda	dlrs	+sharehol.	england	+pct	+wall	+exhch.	+price	mln2	u.s.a

Basis Vectors & Random Initialization

(gd-cls λ = 2, 50 iter. on REUTERS10)

$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	$\mathbf{W}_5$	$\mathbf{W}_6$	$\mathbf{W}_7$	$\mathbf{W}_8$	$\mathbf{W}_9$	$\mathbf{W}_{10}$
MIN	A=22658	seed=59							
+tonne	+billion	+share	stg	mln-mln	gulf	+dollar	+oil	+loss	+trade
+wheat	+year	+offer	+bank	cts	iran	+rate	opec	+profit	japan
+grain	+earn	+company	money	mln	+attack	+curr.	+barrel	oper	japanese
+crop	+qrtr	+stock	+bill	shr	+iranian	+bank	bpd	+exclude	+tariff
corn	+rise	+sharehol.	+market	+net	+ship	yen	crude	+net	+import
agricul.	pct	+common	england	avg	+tanker	monetary	+price	dlrs	reagan
AVER	A=22688	seed=1							
+tonne	+billion	+share	stg	+rate	analy.	+dollar	+oil	+loss	+trade
+wheat	+quarter	+offer	+bank	+bank	+market	+curr.	+barrel	cts	japan
+grain	+earn	+stock	money	econom.	+sell	yen	opec	mln	japanese
+crop	+year	+company	+bill	+fed	+firm	+paris	bpd	+net	+tariff
corn	+rise	+common	london	+cut	+business	japan	crude	shr	+import
usda	dlrs	+sharehol.	england	+pct	+wall	+exhch.	+price	mln-mln	u.s.a
MAX	A=22727	seed=58							
+tonne	+bank	+share	japanes	+rate	gulf	+dollar	+oil	+loss	+trade
+wheat	brazil	+offer	japan	pct	iran	+curr.	+barrel	mln	+import
+grain	+strike	+company	semicon.	+rise	+iranian	yen	opec	cts	+country
+crop	+loan	+stock	tokyo	money	+attack	+central	bpd	+net	+surplus
corn	+billion	dlrs	+chip	econom.	+ship	paris	crude	shr	+deficit
usda	seaman	+sharehol.	+official	+bank	+missile	+bank	+price	+profit	reagan

SVD Acc = 22656 vs. NMF Acc = 22658

Basis Vectors & SVD Initialization

- NMF algorithm gd-cls only needs to initialize **W**.
- Since Text Miner builds SVD basis vectors **U** (from $\mathbf{A}_k = \mathbf{U}\Sigma\mathbf{V}^T$), and **U** is optimal basis in some sense . . .

can we use **U** to initialize **W**?

- — Does this improve convergence rate?
- — Does this improve accuracy, i.e., does gd-cls converge to better local min?

Basis Vectors & SVD Initialization

- NMF algorithm gd-cls only needs to initialize **W**.
- Since Text Miner builds SVD basis vectors **U** (from $\mathbf{A}_k = \mathbf{U}\Sigma\mathbf{V}^T$), and **U** is optimal basis in some sense . . .

can we use **U** to initialize **W**?

- — Does this improve convergence rate? No, on aver., 30 iter.
- — Does this improve accuracy, i.e., does gd-cls converge to better local min? No

Basis Vectors & SVD Initialization

How should we use **U** to initialize **W**?

- Column i of **U** contains +, –, 0 values. Maybe this means that basis vector i is positively and negatively correlated with terms.
  - $\mathbf{W}_0 = \mathbf{U} > 0$ (initialize basis vectors to terms with + correlation)
  - $\mathbf{W}_0 = \mathbf{U} < 0$ (initialize basis vectors to terms with – correlation)
  - $\mathbf{W}_0 = abs(\mathbf{U} > .001)$ (initialize basis vectors to terms with any large correlation)

Basis Vectors & SVD Initialization

How should we use **U** to initialize **W**?

- Maybe +, – signs in column i of **U** connote positive and negative correlation with terms.

  — $\mathbf{W}_0 = \mathbf{U} > 0$ (initialize basis vectors to terms with + correlation)
  Acc=22725

  — $\mathbf{W}_0 = \mathbf{U} < 0$ (initialize basis vectors to terms with – correlation)
  Acc=22765

  — $\mathbf{W}_0 = abs(\mathbf{U} > .001)$ (initialize basis vectors to terms with any large correlation)

  Acc=22688

  (Recall: Best Acc=22658)

Basis Vectors & SVD Initialization

How should we use **U** to initialize **W**?

- Maybe +, – signs in column i of **U** connote positive and negative correlation with terms.

  - $\mathbf{W}_0 = \mathbf{U} > 0$ (initialize basis vectors to terms with + correlation)
    Acc=22725
  - $\mathbf{W}_0 = \mathbf{U} < 0$ (initialize basis vectors to terms with – correlation)
    Acc=22765
  - $\mathbf{W}_0 = abs(\mathbf{U} > .001)$ (initialize basis vectors to terms with any large correlation)
    Acc=22680

    (Recall: Best Acc=22658)

Mixed signs in **U** make correspondence with **W** impossible. They are completely different bases built from completely different philosophies.

Basis Vectors & SVD Initialization

- Wilds has shown Concept/Centroid Decomposition makes for good initialization. (unfortunately, too expensive: 26 sec., which is > gd-cls)

Can we use SVD output to form cheap centroid basis vectors?

Basis Vectors & SVD Initialization

- Wilds has shown Concept/Centroid Decomposition makes for good initialization.

Can we use SVD output to form cheap centroid basis vectors?

Yes. Use low dimension $\mathbf{V}^T$ to cluster documents.

- Run clustering algorithm on $\mathbf{V}_{n\times k}$. (EX: k-means on $\mathbf{V}_{9,248\times 10}$)
- Locate documents (cols of **A**) corresponding to clusters of **V**. (EX: cluster 1 = [$\mathbf{A}_1$,$\mathbf{A}_5$,$\mathbf{A}_9$], etc.)
- Compute centroid of these document clusters.

(EX: $\mathbf{C}_1$=$\mathbf{A}_1$+$\mathbf{A}_5$+$\mathbf{A}_9$)

Basis Vectors & SVD Initialization

- Wilds has shown Concept/Centroid Decomposition makes for good initialization.

Can we use SVD output to form cheap centroid basis vectors?

Yes. Use low dimension $\mathbf{V}^T$ to cluster documents.

- Run clustering algorithm on $\mathbf{V}_{n\times k}$. (EX: k-means on $\mathbf{V}_{9,248\times 10}$)
- Locate documents (cols of $\mathbf{A}$) corresponding to clusters of $\mathbf{V}$. (EX: cluster 1 = $[\mathbf{A}_1, \mathbf{A}_5, \mathbf{A}_9]$, etc.)
- Compute centroid of these document clusters. (EX: $\mathbf{C}_1 = \mathbf{A}_1 + \mathbf{A}_5 + \mathbf{A}_9$)

Results when $\mathbf{W}_0 = [\,\mathbf{C}_1 | \cdots | \mathbf{C}_k\,]$

- Time: clustering on $\mathbf{V}^T$ about 1 sec. + 15 sec. for NMF gd-cls.
- Acc: 22666, slightly better than average random $\mathbf{W}_0$ case.

Basis Vectors & Centroid Initialization

(gd-cls $\lambda = 2$, 50 iter. on REUTERS10)

W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
centroids									
+tonne	+billion	+share	+comp	cts	iran	+bank	+oil	+loss	+trade
+wheat	+earn	+offer	pct	shr	+gulf	+rate	+barrel	oper	japan
+grain	+qrtr	+company	+bank	mln	+attack	money	opec	+profit	japanese
corn	+year	+stock	dlrs	+net	+iranian	+market	bpd	cts	+offic.
+crop	dlrs	pct	+type	mln2	+missile	+dollar	crude	mln	+tariff
agricul.	+rise	+common	inc	+rev	+ship	central	+price	+net	+import
MIN	A=22658	seed=59							
+tonne	+billion	+share	stg	mln2	gulf	+dollar	+oil	+loss	+trade
+wheat	+year	+offer	+bank	cts	iran	+rate	opec	+profit	japan
+grain	+earn	+company	money	mln	+attack	+curr.	+barrel	oper	japanese
+crop	+qrtr	+stock	+bill	shr	+iranian	+bank	bpd	+exclude	+tariff
corn	+rise	+sharehol.	+market	+net	+ship	yen	crude	+net	+import
agricul.	pct	+common	england	avg	+tanker	monetary	+price	dlrs	reagan

Future Work

- Other algorithms: quasi-Newton methods
- New NLP objective: pseudo NMF, discrete NMF