The

Nonnegative Matrix Factorization

a tutorial

Barbara Ball
barbaraeball@comcast.net
C. of Charleston
Mathematics Dept.

Atina Brooks
adbrook2@ncsu.edu
N.C. State U.
Statistics Dept.

Amy Langville
langvillea@cofc.edu
C. of Charleston
Mathematics Dept.

NISS NMF Workshop
February 23–24, 2007

Outline

- Two Factorizations:
  - Singular Value Decomposition
  - Nonnegative Matrix Factorization
- Why factor anyway?
- Computing the NMF
  - Early Algorithms
  - Recent Algorithms
- Extensions of NMF

Data Matrix

$\mathbf{A}_{m \times n}$ with rank r

Examples

term-by-document matrix

pixel intensity-by-image matrix

gene-by-DNA microarray matrix

feature-by-item matrix

user-by-purchase matrix

terrorist-by-action matrix

SVD

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\,\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

What is the SVD?

$$\underset{m\times n}{\mathsf{A}} = \underset{m\times m}{\mathsf{U}} \; \underset{m\times n}{\Sigma} \; \underset{n\times n}{\mathsf{V}^T}$$

$$\mathsf{U}^T\mathsf{U} = \mathrm{I}_m$$

$$\mathsf{V}^T\mathsf{V} = \mathrm{I}_n$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

r : rank of A

$$\underset{m \times n}{\mathsf{A}} = \sum_{i=0}^{\min(m,n)} \sigma_i u_i v_i^T$$

$$\underset{m \times n}{\mathsf{A}} = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$$

decreasing importance

The SVD

Data Matrix

$\mathbf{A}_{m \times n}$ with rank r

Examples

term-by-document matrix

pixel intensity-by-image matrix

gene-by-DNA microarray matrix

feature-by-item matrix

user-by-purchase matrix

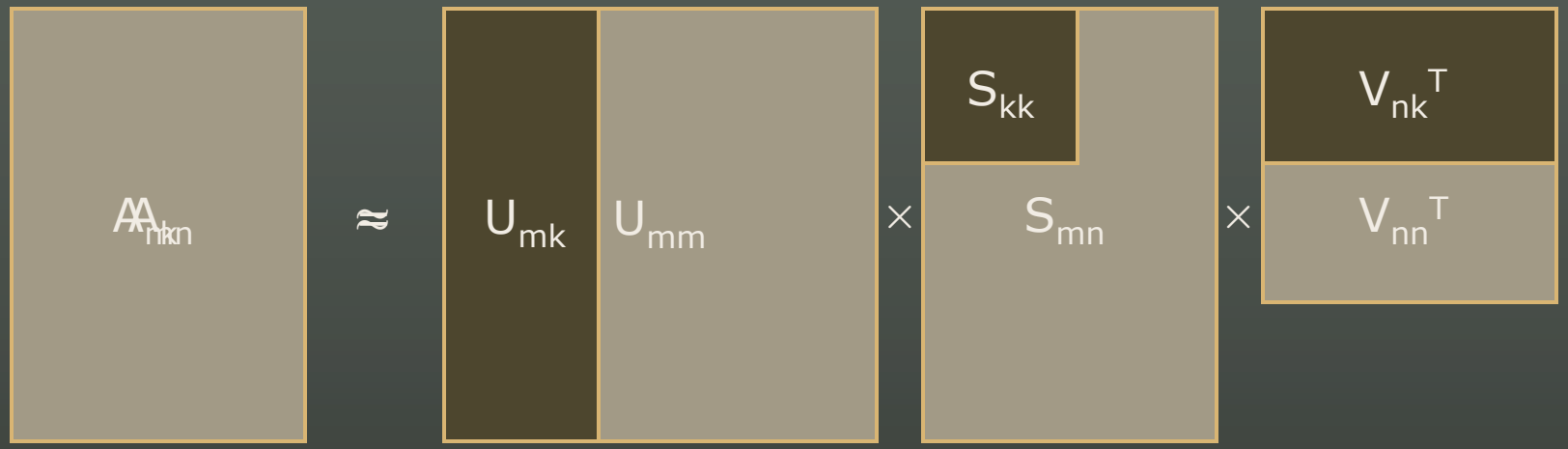
terrorist-by-action matrix

SVD

$$\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

Low Rank Approximation

use $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ in place of $\mathbf{A}$



SVD Rank Reduction

Why use Low Rank Approximation?

- Data Compression and Storage when $k << r$
- Remove noise and uncertainty
  - $\Rightarrow$ improved performance on data mining task of retrieval (e.g., find similar items)
  - $\Rightarrow$ improved performance on data mining task of clustering

Properties of SVD

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

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

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

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

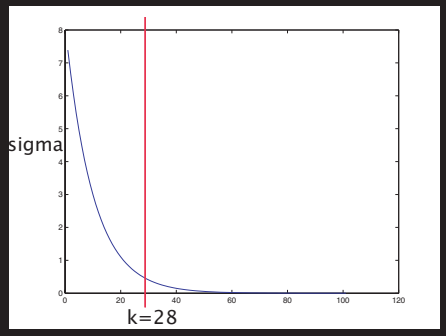
Summary of Truncated SVD

Strengths

- using $\mathbf{A}_k$ in place of $\mathbf{A}$ gives improved performance
- noise reduction isolates essential components of matrix
- best rank-k approximation
- $\mathbf{A}_k$ is unique

Weaknesses

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



Other Low-Rank Approximations

- **QR** decomposition
- any $\mathbf{URV}^T$ factorization
- Semidiscrete decomposition (SDD)

  $\mathbf{A}_k = \mathbf{X}_k \mathbf{D}_k \mathbf{Y}_k^T$, where $\mathbf{D}_k$ is diagonal, and elements of $\mathbf{X}_k, \mathbf{Y}_k \in \{-1, 0, 1\}$.
- **CUR** factorization

Other Low-Rank Approximations

- **QR** decomposition
- any **URV**^{T} factorization
- Semidiscrete decomposition (SDD)

  $\mathbf{A}_k = \mathbf{X}_k \mathbf{D}_k \mathbf{Y}_k^T$, where $\mathbf{D}_k$ is diagonal, and elements of $\mathbf{X}_k, \mathbf{Y}_k \in \{-1, 0, 1\}$.
- **CUR** factorization

BUT

All create basis vectors that are mixed in sign.
Negative elements make interpretation difficult.

Other Low-Rank Approximations

- **QR** decomposition
- any $\mathbf{URV}^T$ factorization
- Semidiscrete decomposition (SDD)

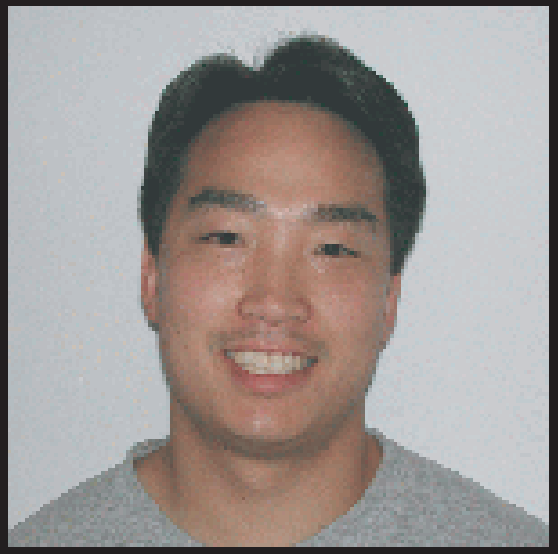
  $\mathbf{A}_k = \mathbf{X}_k \mathbf{D}_k \mathbf{Y}_k^T$, where $\mathbf{D}_k$ is diagonal, and elements of $\mathbf{X}_k, \mathbf{Y}_k \in \{-1, 0, 1\}$.
- **CUR** factorization

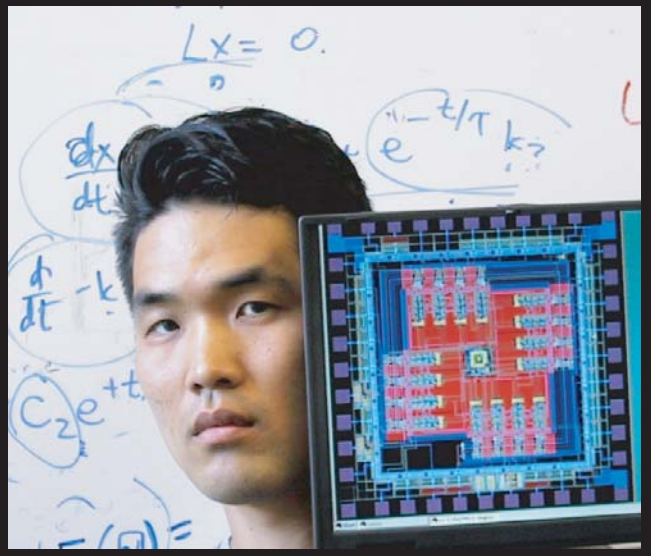
BUT

All create basis vectors that are mixed in sign.
Negative elements make interpretation difficult.

$\Rightarrow$ Nonnegative Matrix Factorization

Nonnegative Matrix Factorization (2000)





Daniel Lee and Sebastian Seung's Nonnegative Matrix Factorization

Idea: use low-rank approximation with nonnegative factors to improve weaknesses of trun-SVD

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

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

Interpretation with NMF

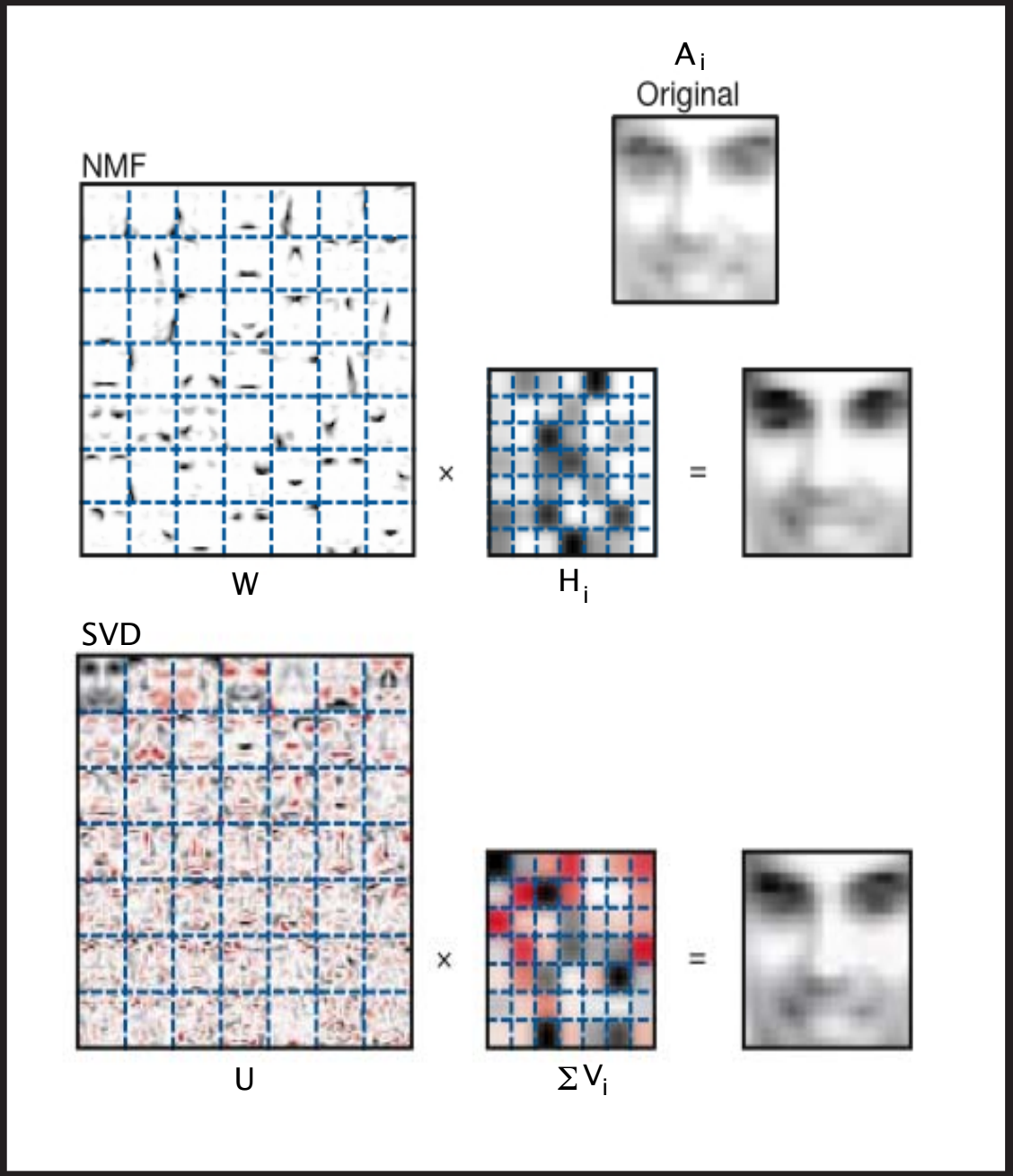
- columns of **W** are the underlying basis vectors, i.e., each of the n columns of **A** can be built from k columns of **W**.

- columns of **H** give the weights associated with each basis vector.

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

- $\mathbf{W}_k, \mathbf{H}_k \geq 0 \Rightarrow$ immediate interpretation (additive parts-based rep.)

Image Mining



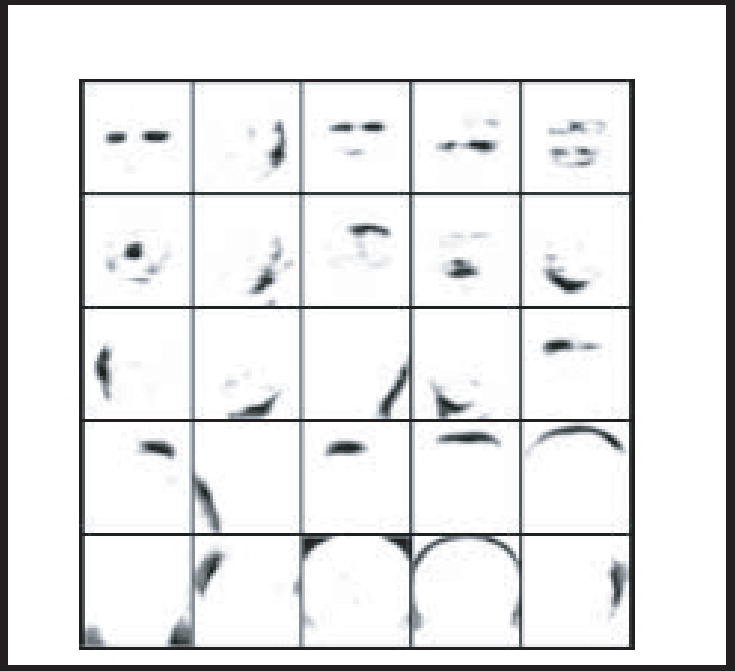
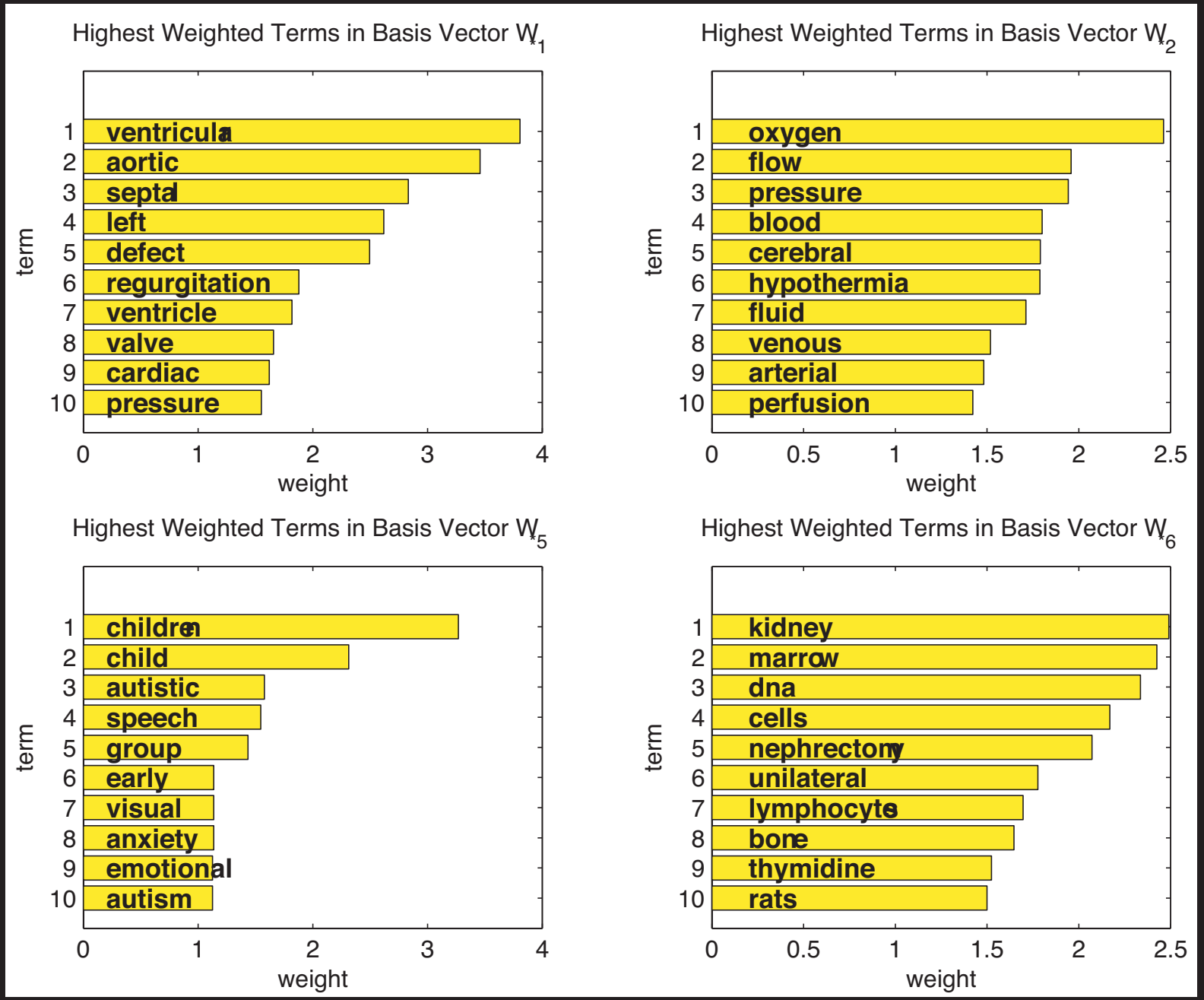


Image Mining Applications

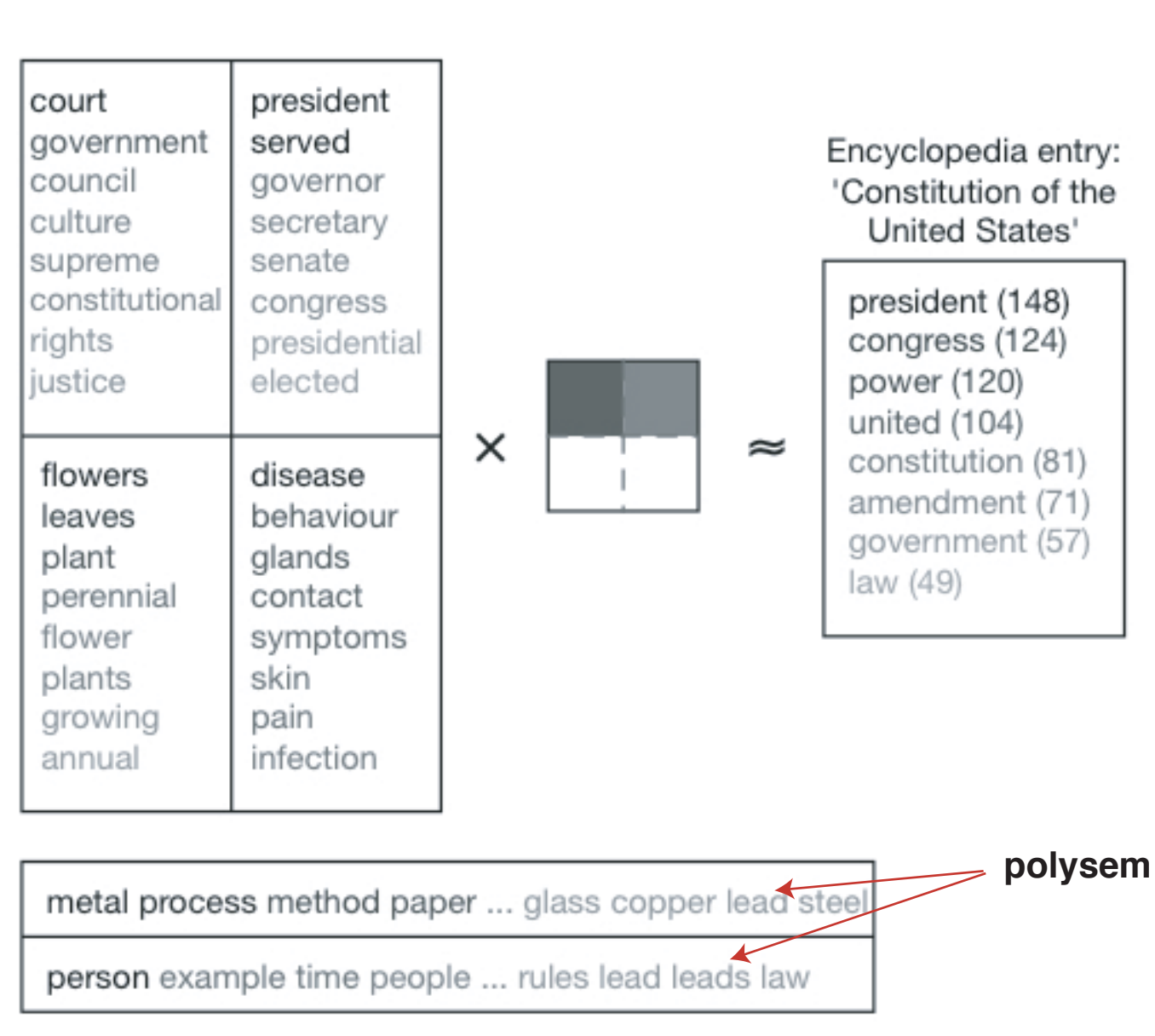
- Data compression
- Find similar images
- Cluster images

Text Mining

MED dataset ($k = 10$)



Text Mining



- polysems broken across several basis vectors $\mathbf{w}_i$

Text Mining Applications

- Data compression $\mathbf{W}_k\mathbf{H}_k$
- Find similar terms $0 \leq cos(\theta) = \mathbf{W}_k\mathbf{H}_k\mathbf{q} \leq 1$
- Find similar documents $0 \leq cos(\theta) = \mathbf{q}^T\mathbf{W}_k\mathbf{H}_k \leq 1$
- Cluster documents

Clustering with the NMF

Clustering Terms

- use rows of $\mathbf{W}_{m \times k}$

$$
\begin{array}{c} \\ term1 \\ term2 \\ \vdots \end{array}
\begin{array}{c}
\begin{array}{cccc} cl.1 & cl.2 & \dots & cl.k \end{array} \\
\begin{pmatrix} .9 & 0 & \dots & .3 \\ .1 & .8 & \dots & .2 \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}
\end{array}
$$

Clustering Documents

- use cols of $\mathbf{H}_{k \times n}$

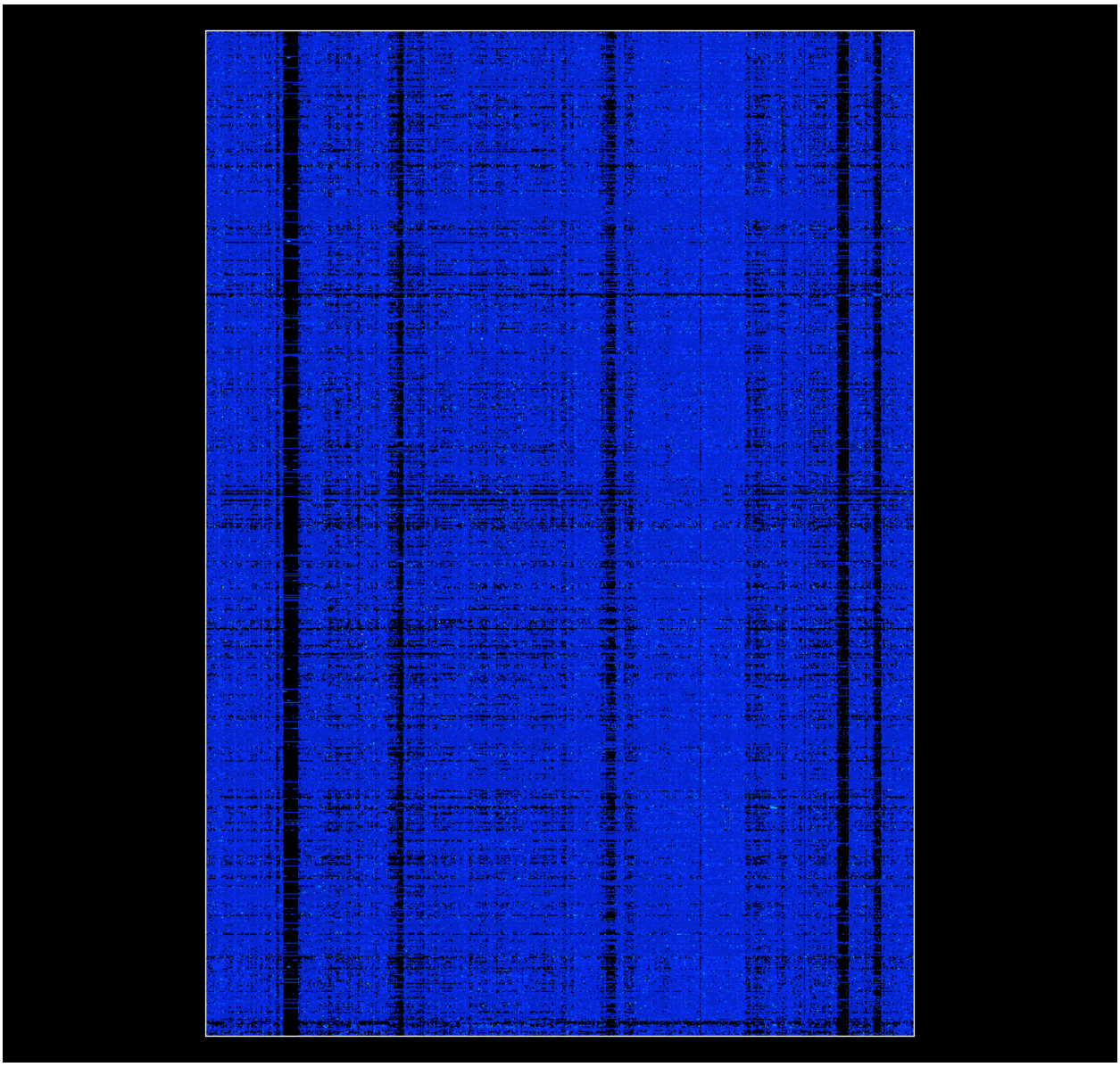
$$
\begin{array}{c} \\ cl.1 \\ \vdots \\ cl.k \end{array}
\begin{array}{c}
\begin{array}{cccc} doc1 & doc2 & \dots & docn \end{array} \\
\begin{pmatrix} .4 & 0 & \dots & .5 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & .8 & \dots & .2 \end{pmatrix}
\end{array}
$$

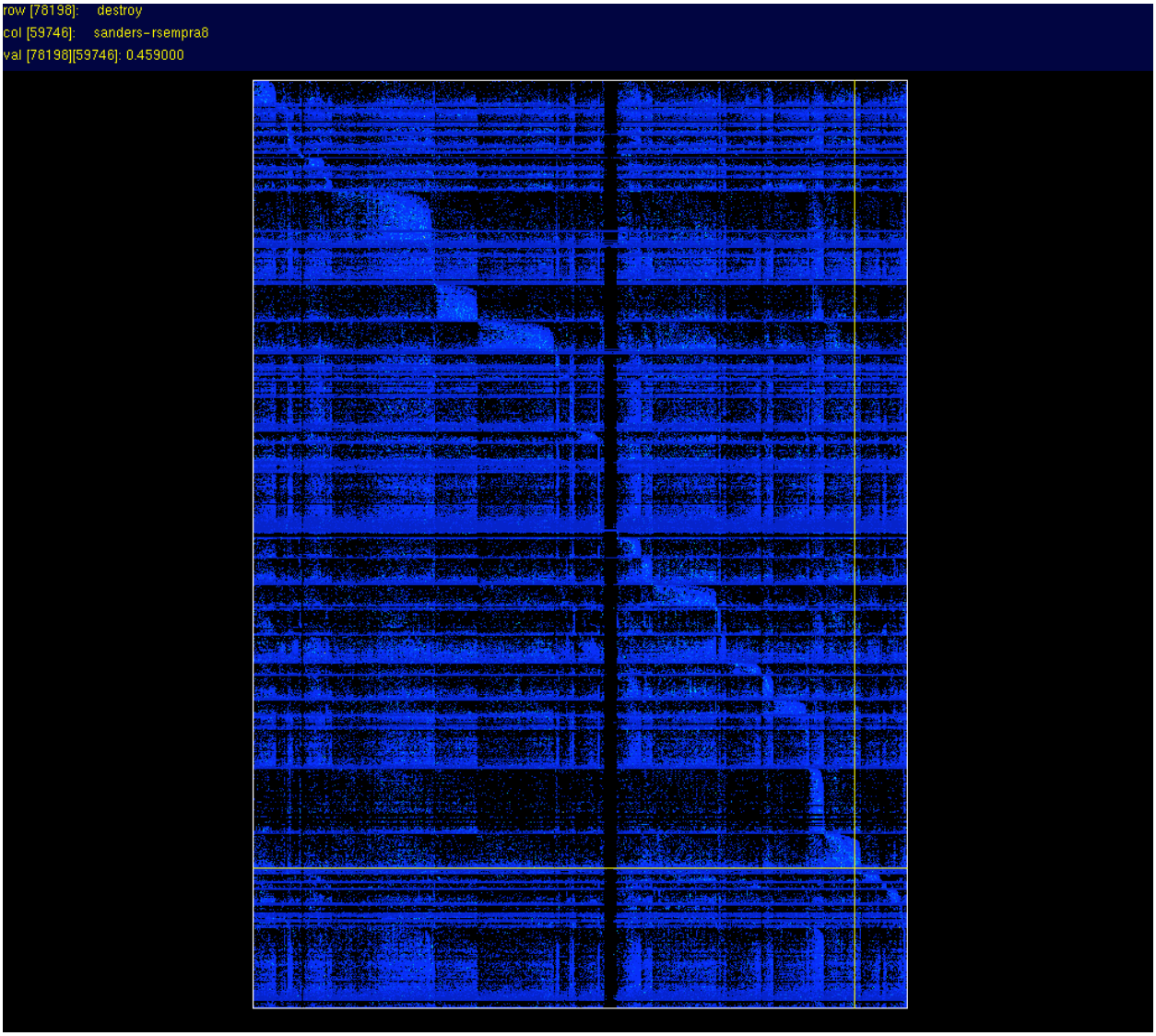
soft clustering is very natural

The Enron Email Dataset (SAS)

- PRIVATE email collection of 150 Enron employees during 2001
- 92,000 terms and 65,000 messages
- Term-by-Message Matrix

$$
\begin{array}{c}
 \\
\vdots \\
subpoena \\
dynegy \\
\vdots
\end{array}
\begin{array}{c}
\begin{array}{cccc}
fastow1 & fastow2 & skilling1 & \dots
\end{array} \\
\left(\begin{array}{cccc}
\vdots & \vdots & \vdots & \dots \\
2 & 0 & 1 & \dots \\
0 & 3 & 0 & \dots \\
\vdots & \vdots & \vdots & \dots
\end{array}\right)
\end{array}
$$




row [78196]: destroy
col [59746]: sanders-rsempra8
val [78196][59746]: 0.459000

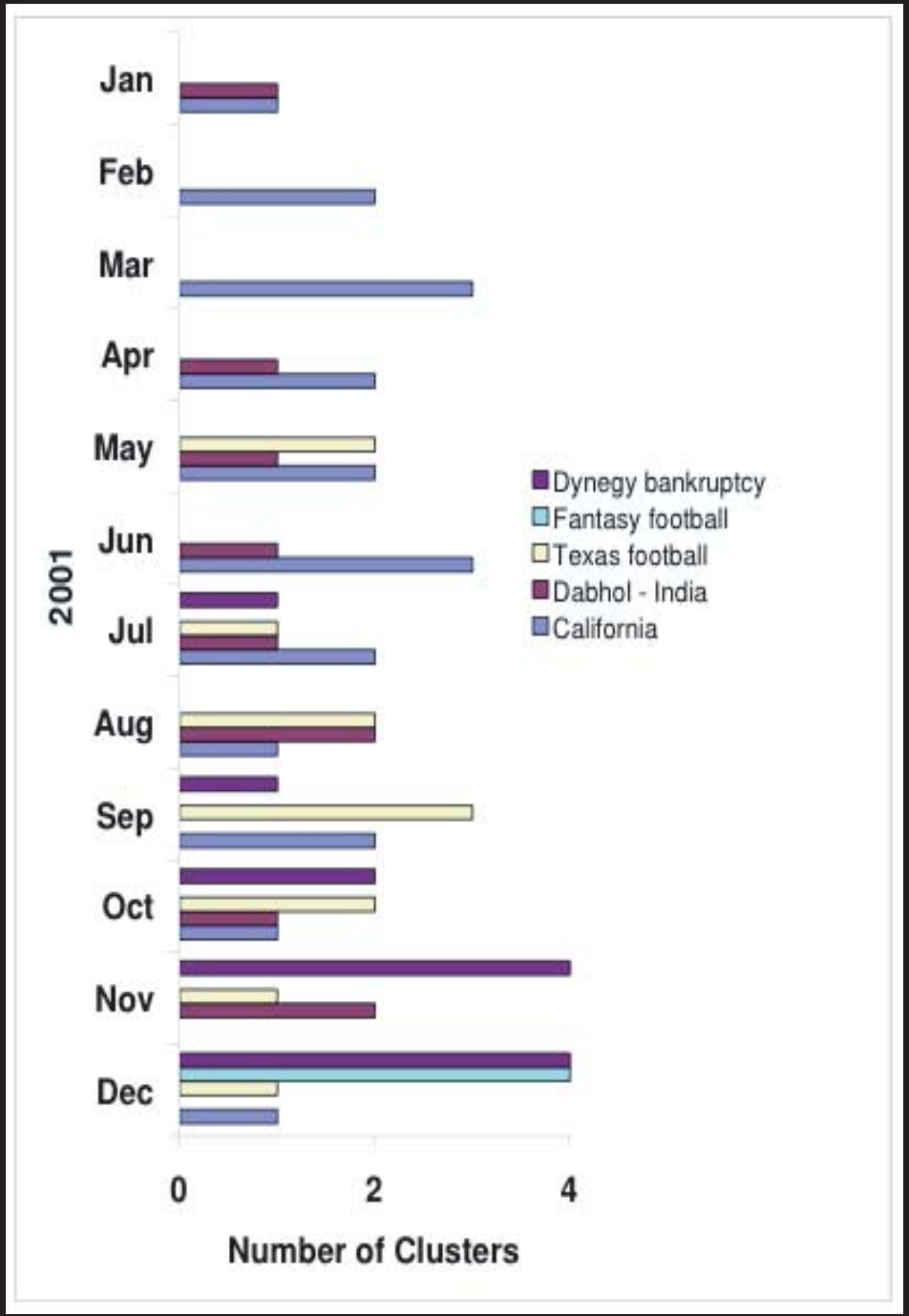
Text Mining Applications

- Data compression $\mathbf{W}_k\mathbf{H}_k$
- Find similar terms $0 \leq cos(\theta) = \mathbf{W}_k\mathbf{H}_k\mathbf{q} \leq 1$
- Find similar documents $0 \leq cos(\theta) = \mathbf{q}^T\mathbf{W}_k\mathbf{H}_k \leq 1$
- Cluster documents
- Topic detection and tracking

Text Mining Applications

Enron email messages 2001

Feature Index (k)	Cluster Size	Topic Description	Dominant Terms
10	497	California	ca, **cpuc,** gov, **socalgas,** sempra, org, sce, gmssr, aelaw, ci
23	43	Louise Kitchen named top woman by Fortune	evp, **fortune,** britain, woman, **ceo,** avon, fiorinai, cfo, hewlett, packard
26	231	Fantasy football	game, wr, qb, play, rb, season, injury, updated, fantasy, image
33	233	Texas longhorn football newsletter	UT, orange, longhorn[s], texas, true, truorange, recruiting, oklahoma defensive
34	65	Enron collapse	**partnership[s], fastow,** shares, **sec,** stock, shareholder, investors, equity, **lay**
39	235	Emails about India	**dahhol, dpc, india, mseb, maharashtra,** indian, lenders, delhi, foreign, minister
46	127	Enron collapse	dow, debt, reserved, wall, copyright jones, cents, analysts, reuters, spokesman



Recommendation Systems

purchase history matrix $\mathbf{A} =$

	User 1	User 2	...	User n
Item 1	1	5	...	0
Item 2	0	0	...	1
⋮	⋮	⋮	⋱	⋮
Item m	0	1	...	2

- Create profiles for classes of users from basis vectors $\mathbf{w}_i$
- Find similar users
- Find similar items

Microarray Study
Kim and Tidor, 2003

- 300 experiments with 5436 *S. cerevisiae* genes
- expression for a gene described by the expression in experiment divided by control experiment of wild type under typical conditions
- basis vector represented by an experiment, containing a relative expression for each gene and its related feature

Functional Relationships

Table 3. The 58 Predictions That Could Be Validated by YPD of the 100 Strongest Functional Relationships Detected by NMF

Coregulated	
dfr1	ecm34
gyp1	yap7
ade16	sir1
hpt1	sir1
rml2	ymr293c
dbp2	mrpl33
mrpl33	rml2
cnb1	yor072w
ade16	ymr041c
gfd1	utr4
cla4 (haploid)	KAR2 (tet promoter)
yel001c	ymr141c
ckb2	gcn4
arg5,6	rpl8a
mrt4	rpl12a
clb6	whi2
erp2	ymr141c
erp2	yel001c
erp2	yor015w
rpl12a	yel033w
ckb2	rtg1
eca39	ras1
Identical Genes	
isw1	isw1, isw2
dig1, dig2	dig1, dig2 (haploid)
fks1 (haploid)	FKS1 (tet promoter)
bub3	bub3 (haploid)

Binding	
cla4 (haploid)	CDC42 (tet promoter)
qcr2 (haploid)	rip1
far1 (haploid)	ste4 (haploid)
bub1 (haploid)	bub3
bub1 (haploid)	bub3 (haploid)
Cell Wall	
fks1 (haploid)	2-deoxy-D-glucose
2-deoxy-D-glucose	Glucosamine
gas1	Tunicamycin
fks1 (haploid)	Glucosamine
yer083c	Tunicamycin
ste12 (haploid)	ste5 (haploid)
Mating	
ste5 (haploid)	ste7 (haploid)
fus3, kss1 (haploid)	ste5 (haploid)
ste18 (haploid)	ste5 (haploid)
ste12 (haploid)	ste18 (haploid)
ste18 (haploid)	ste7 (haploid)
fus3, kss1 (haploid)	ste18 (haploid)
fus3, kss1 (haploid)	ste7 (haploid)
fus3, kss1 (haploid)	ste12 (haploid)
ste12 (haploid)	ste7 (haploid)

Ergosterol Pathway	
erg3 (haploid)	Itraconazole
erg2	Itraconazole
yer044c (haploid)	ERG11 (tet promoter)
ERG11 (tet promoter)	Itraconazole
erg3 (haploid)	ERG11 (tet promoter)
erg3 (haploid)	yer044c (haploid)
erg2	erg3 (haploid)
erg2	yer044c (haploid)
erg2	ERG11 (tet promoter)
Vacuolar ATPase	
cup5	mac1
mac1	vma8
cup5	vma8

More Functional Relationships

Table 4. The 42 Predictions of Functional Relationships That Could Not Be Verified on YPD From the 100 Strongest Relationships Detected

rtg1	vps8
are1, are2 (haploid)	yor015w
pex12	yea4
ckb2	yel008w
yer002w	ymr034c
mrt4	yel033w
ckb2	rts1
mrpl33	ymr293c
imp2	yer050c
cbp2	pet111
cyt1	pet111
yer034w	ynd1
rps24a	ymr014w
yel001c	yor015w
ymr014w	yor006c
aep2	rml2
aep2	mrpl33
ymr014w	yor078w
rml2	yer050c
mrpl33	yer050c
aep2	imp2
sir1	ymr041c
ymr034c	yor015w
pfd2	yor051c
ymr025w	ymr029c
ckb2	vps8
msu1	ymr293c
sbh2	yer084w
mrpl33	msu1
imp2	ymr293c
rtg1	rts1
msu1	yer050c
msu1	rml2
yml003w	ymr034c
aep2	msu1
CDC42 (tet promoter)	KAR2 (tet promoter)
rps24a	yor078w
pfd2	yel044w
gcn4	yel008w
yer050c	ymr293c
aep2	yer050c
aep2	ymr293c

Comparative Study

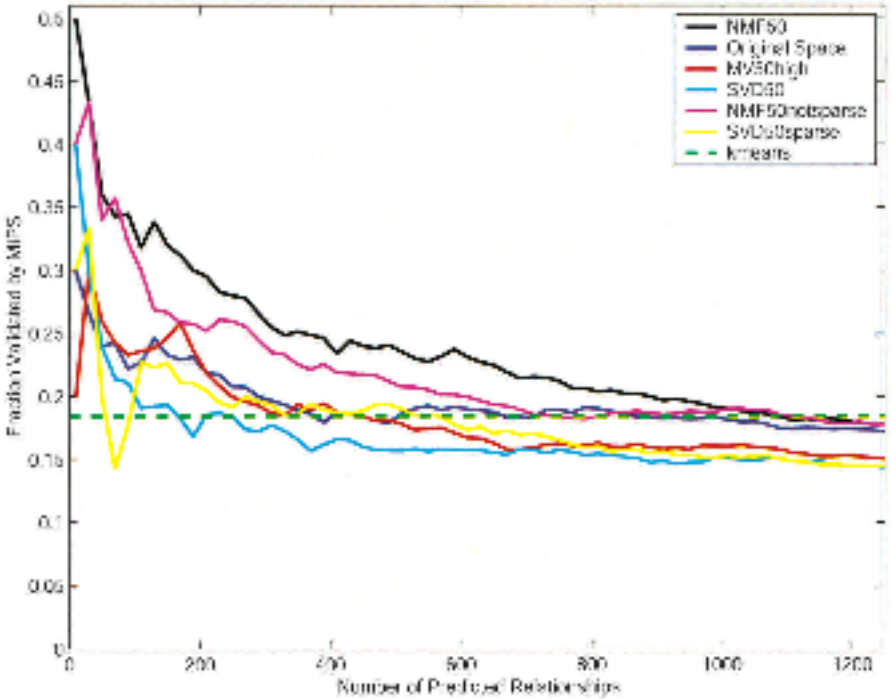
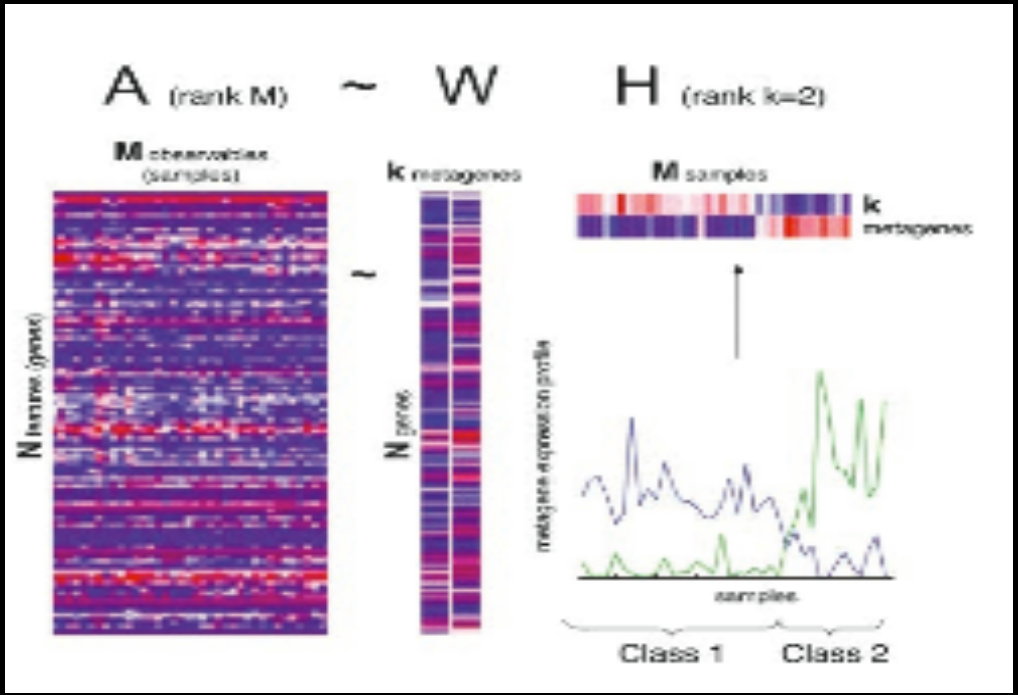


Figure 3 Performance of different spaces at predicting functional relationships between experiments with comparison to the MIPS classification of the deleted genes. (NMF50) NMF space with 50 basis vectors; (Original Space) original gene expression space; (SVD50) SVD space with 50 eigenvectors; (MV50high) space of the 50 most varying genes; (NMF50notsparse) NMF space with 50 basis vector without the sparsification procedure; (SVD50sparse) SVD sparsified; (k-means) predictions taken from k-means clustering with 50 clusters (3176 relationships).

Metagenes Study

Brunet et al 2004

Data: gene expressions x samples



Leukemia Samples

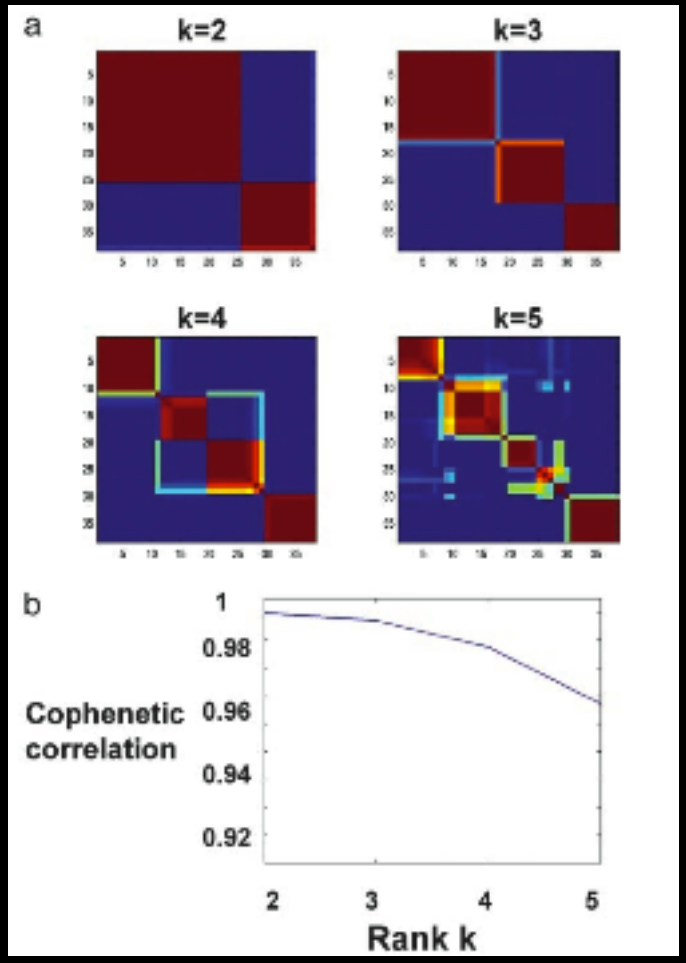


Fig. 4. (*a*) Reordered consensus matrices averaging 50 connectivity matrices computed at $k = 2$–5 for the leukemia data set with the 5,000 most highly varying genes according to their coefficient of variation. Samples are hierarchically clustered by using distances derived from consensus clustering matrix entries, colored from 0 (deep blue, samples are never in the same cluster) to 1 (dark red, samples are always in the same cluster). Compositions of the leukemia clusters determined by HC of consensus matrices are as follows: for $k = 2$: {(25 ALL), (11 AML and 2 ALL)}, $k = 3$: {(17 ALL-B), (8 ALL-T and 1 ALL-B), (11 AML and 1 ALL-B)}, $k = 4$: {(11 ALL-B), (7 ALL-B and 1 AML), (8 ALL-T and 1 ALL-B), (10 AML)}. (*b*) Cophenetic correlation coefficients for hierarchically clustered matrices in *a*.

Samples from Medulloblastoma Tumors

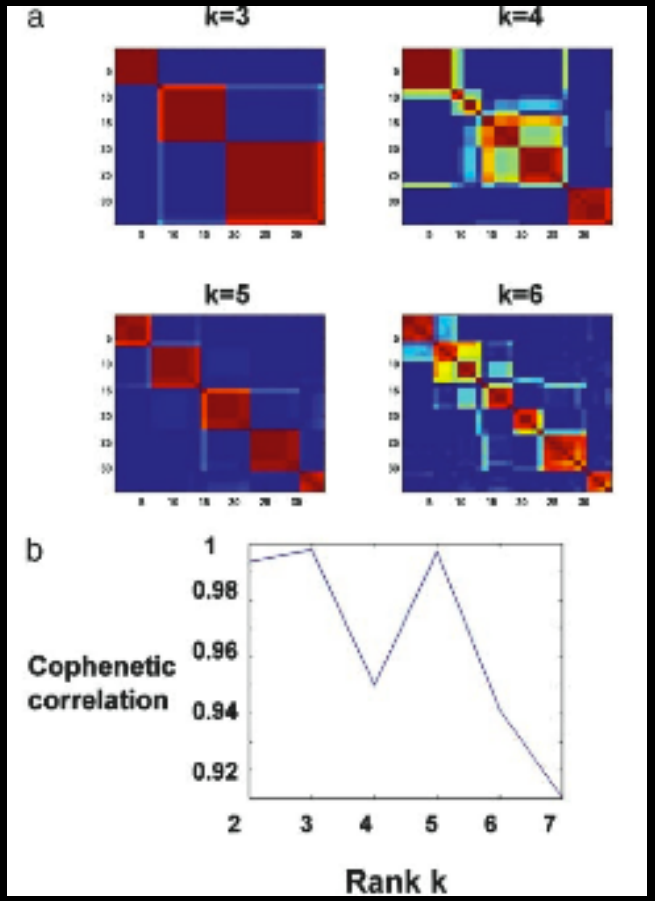


Fig. 6. (*a*) NMF model selection for a data set of 25 classic and 9 desmoplastic medulloblastoma tumors [$n = 5{,}893$; $M = 34$ (14)]. At each rank k, a consensus matrix, averaging 50 connectivity matrices, is reordered by using HC (color map as Fig. 4). In addition to a robust two-class partition (not shown), the consensus is strong for $k = 3, 5$, indicating reproducible partitioning of samples into two, three, and five classes but not four or six. (*b*) Cophenetic correlation coefficients corresponding to the HC of consensus matrices for $k = 2$–7 shows a dip at $k = 4$, where reproducibility is poor, and suggests $k = 5$ as the largest number of classes recognized by NMF for this data set.

CNS Embryonal Tumors

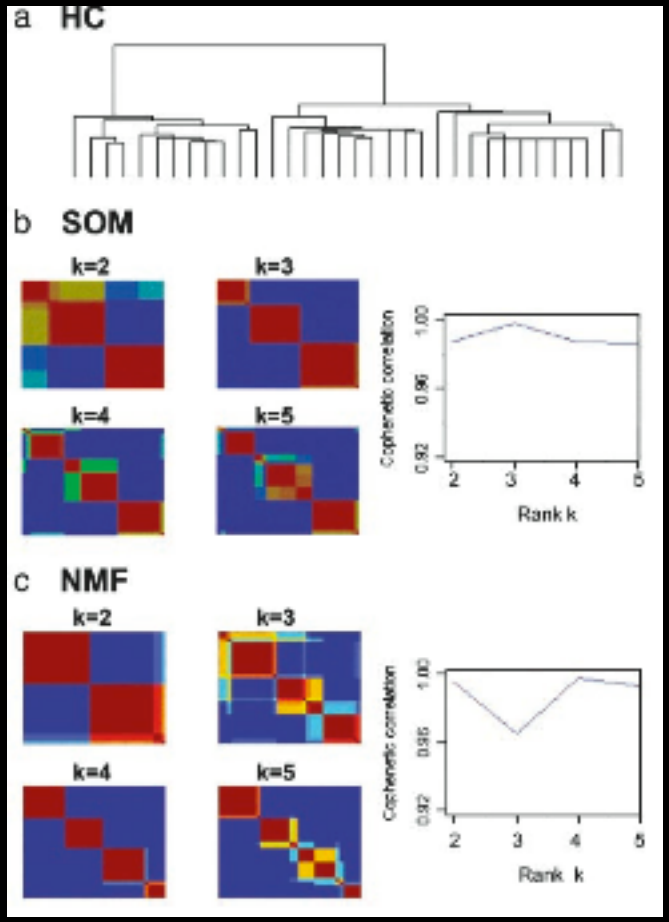


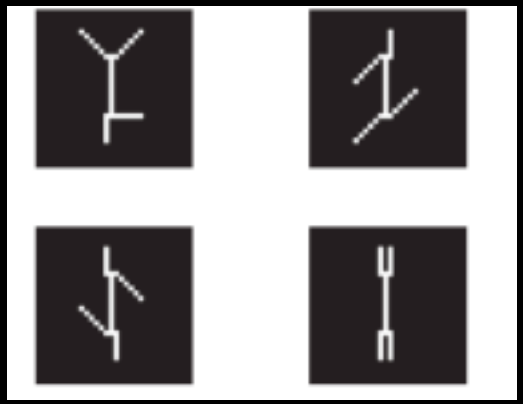
Fig. 7. Analysis of central nervous system embryonal tumors using 5,560 genes. The data set consists of 34 samples, including 10 classic medulloblastomas, 10 malignant gliomas, 10 rhabdoids, and 4 normals. (a) The dendrogram from HC indicates two or three major subclasses but gives no clear indication of a four-class split. (b) Reordered consensus matrices for $k = 2–5$ centroid SOM clusterings from 20 initial conditions. Cophenetic correlation argues for a three-class decomposition. (c) Reordered consensus matrices for 20 NMF initial conditions (50 NMF iterations each), for $k = 2–5$ (color scale same as Fig. 2). Cophenetic correlation coefficient suggests the existence of at most four robust classes.

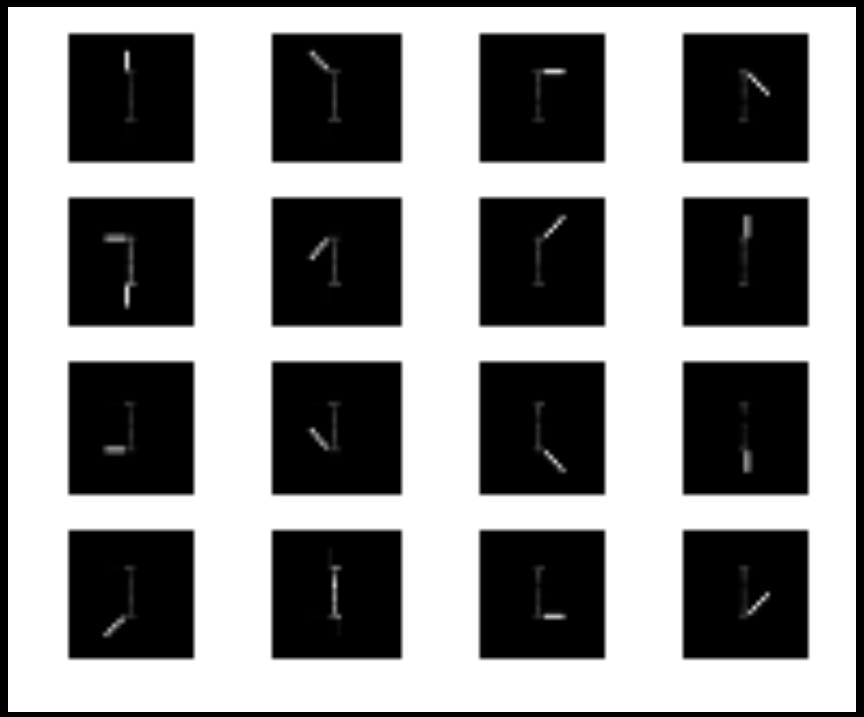
When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?

Donoho and Stodden, 2003

- Set of weighted generators, non-negative
- Each combination separable from other combinations
- All combinations represented in dataset

Example of a Separable Factorial Articulation Family





Properties of NMF

- basis vectors $\mathbf{w}_i$ are not $\perp \Rightarrow$ can have overlap of topics
- can restrict **W**, **H** to be sparse
- $\mathbf{W}_k$, $\mathbf{H}_k \geq 0 \Rightarrow$ 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 $\mathbf{w}_i$

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

- NMF is algorithm-dependent: **W**, **H** not unique

Report Card for SVD and NMF

Subject	SVD	NMF
low rank approximation improves performance on data mining tasks	A	A
noise reduction isolates essential components of matrix	A	A
quality of low rank approximation, $\|\mathbf{A} - \mathbf{A}_k\|$	A+	B+
uniqueness of low rank approximation	A+	B
storage of low rank factors	D	A
interpretation of vectors in low rank factors	C	A
choosing truncation point k	C	C
orthogonality restriction on vectors in low rank factors	B	A
updating the factors in the factorization	B	C
downdating the factors in the factorization	A	C

Computation of NMF

(Lee and Seung 2000)

Mean squared error objective function

$$\min \|\mathbf{A} - \mathbf{WH}\|_F^2$$

$$s.t. \quad \mathbf{W}, \mathbf{H} \geq 0$$

Nonlinear Optimization Problem

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

  (EX: $\mathbf{A}_{70K \times 10K}$ and k=10 topics $\Rightarrow$ 800K unknowns)
- above objective is one of many possible

Other Objective Functions

Divergence objective function

$$\min \sum_{i,j} (\mathbf{A}_{ij} \log \frac{\mathbf{A}_{ij}}{[\mathbf{WH}]_{ij}} - \mathbf{A}_{ij} + [\mathbf{WH}]_{ij})$$

Weighted Mean Squared Error objective function

$$\min \|\mathbf{B}. * (\mathbf{A} - \mathbf{WH})\|_F^2$$

Weighted Divergence objective function

$$\min \sum_{i,j} \mathbf{B}_{ij}. * (\mathbf{A}_{ij} \log \frac{\mathbf{A}_{ij}}{[\mathbf{WH}]_{ij}} - \mathbf{A}_{ij} + [\mathbf{WH}]_{ij})$$

Bregman Divergence Class of objective functions

(coming tomorrow–Inderjit Dhillon)

Suite of Other Divergence objective functions

(NMFLAB–Cichocki)

Table 1. Amari Alpha-NMF algorithms

Amari alpha divergence: $D_A^{(\alpha)}(y_{ik}\|\|z_{ik}) = \sum_{ik} \frac{y_{ik}^{\alpha} z_{ik}^{1-\alpha} - \alpha y_{ik} + (\alpha-1) z_{ik}}{\alpha(\alpha-1)}$	
Algorithm:	$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \, (\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}})^{\alpha} \right)^{\frac{\omega_X}{\alpha}} \right)^{1+\alpha_{sX}}$
	$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \, (\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}})^{\alpha} \right)^{\frac{\omega_A}{\alpha}} \right)^{1+\alpha_{sA}}$
	$a_{ij} \leftarrow a_{ij} / \sum_p a_{pj},$
	$0 < \omega_X < 2, \quad 0 < \omega_A < 2$
Pearson distance: $(\alpha = 2)$:	$D_A^{(\alpha=2)}(y_{ik}\|\|z_{ik}) = \sum_{ik} \frac{(y_{ik} - [\boldsymbol{A}\boldsymbol{X}]_{ik})^2}{[\boldsymbol{A}\boldsymbol{X}]_{ik}},$
Algorithm:	$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \, (\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}})^{2} \right)^{\frac{\omega_X}{2}} \right)^{1+\alpha_{sX}}$
	$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \, (\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}})^{2} \right)^{\frac{\omega_A}{2}} \right)^{1+\alpha_{sA}}$
	$a_{ij} \leftarrow a_{ij} / \sum_p a_{pj},$
	$0 < \omega_X < 2, \quad 0 < \omega_A < 2$
Hellinger distance: $(\alpha = \frac{1}{2})$:	$D_A^{(\alpha=0.5)}(y_{ik}\|\|z_{ik}) = \sum_{ik} \frac{(y_{ik} - [\boldsymbol{A}\boldsymbol{X}]_{ik})^2}{[\boldsymbol{A}\boldsymbol{X}]_{ik}},$
Algorithm:	$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \sqrt{\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}}} \right)^{2\omega_X} \right)^{1+\alpha_{sX}}$
	$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \sqrt{\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}}} \right)^{2\omega_A} \right)^{1+\alpha_{sA}}$
	$a_{ij} \leftarrow a_{ij} / \sum_p a_{pj},$
	$0 < \omega_X < 2, \quad 0 < \omega_A < 2$

Table 2. Amari Alpha-NMF algorithms (continued)

Kullback-Leibler divergence: ($\alpha \to 1$):

$$\lim_{\alpha\to 1} D_A^{(\alpha)}(y_{ik}||z_{ik}) = \sum_{ik} y_{ik} \log \frac{y_{ik}}{[\boldsymbol{AX}]_{ik}} - y_{ik} + [\boldsymbol{AX}]_{ik},$$

Algorithm:

$$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}} \right)^{\omega_X} \right)^{1+\alpha_{sX}}$$

$$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}} \right)^{\omega_A} \right)^{1+\alpha_{sA}}$$

$$a_{ij} \leftarrow a_{ij} / \textstyle\sum_p a_{pj}$$

$$0 < \omega_X < 2, \quad 0 < \omega_A < 2$$

Dual Kullback-Leibler divergence: ($\alpha \to 0$):

$$\lim_{\alpha\to 0} D_A^{(\alpha)}(y_{ik}||z_{ik}) = \sum_{ik} [\boldsymbol{AX}]_{ik} \log \frac{[\boldsymbol{AX}]_{ik}}{y_{ik}} + y_{ik} - [\boldsymbol{AX}]_{ik}$$

Algorithm:

$$x_{jk} \leftarrow \left(x_{jk} \prod_{i=1}^{m} \left(\frac{y_{ik}}{[\boldsymbol{AX}]_{ik}} \right)^{\omega_X\, a_{ij}} \right)^{1+\alpha_{sX}}$$

$$a_{ij} \leftarrow \left(a_{ij} \prod_{k=1}^{N} \left(\frac{y_{ik}}{[\boldsymbol{AX}]_{ik}} \right)^{\tilde{\eta}_j x_{jk}} \right)^{1+\alpha_{sA}}$$

$$a_{ij} \leftarrow a_{ij} / \textstyle\sum_p a_{pj}$$

$$0 < \omega_X < 2, \quad 0 < \omega_A < 2$$

Table 3. Other generalized NMF algorithms

Beta generalized divergence:

$$D_K^{(\beta)}(y_{ik}||z_{ik}) = \sum_{ik} y_{ik} \frac{y_{ik}^{\beta-1} - [\boldsymbol{AX}]_{ik}^{\beta-1}}{\beta(\beta-1)} + [\boldsymbol{AX}]_{ik}^{\beta-1} \frac{[\boldsymbol{AX}]_{ik} - y_{ik}}{\beta}$$

Kompass algorithm:

$$x_{jk} \leftarrow x_{jk} \frac{\sum_{i=1}^{m} a_{ij}\ (y_{ik}/[\boldsymbol{AX}]_{ik}^{2-\beta})}{\sum_{i=1}^{m} a_{ij}\ [\boldsymbol{AX}]_{ik}^{\beta-1} + \varepsilon}$$

$$a_{ij} \leftarrow \left(a_{ij} \frac{\sum_{k=1}^{N} x_{jk}\ (y_{ik}/[\boldsymbol{AX}]_{ik}^{2-\beta})}{\sum_{k=1}^{N} x_{jk}\ [\boldsymbol{AX}]_{ik}^{\beta-1} + \varepsilon} \right)^{1+\alpha_{sA}}$$

$$a_{ij} \leftarrow a_{ij} / \textstyle\sum_p a_{pj},$$

Triangular discrimination:

$$D_T^{(\beta)}(y_{ik}||z_{ik}) = \sum_{ik} \frac{y_{ik}^{\beta} z_{ik}^{1-\beta} - \beta y_{ik} + (\beta-1) z_{ik}}{\beta(\beta-1)}$$

Algorithm:

$$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \left(\frac{2y_{ik}}{y_{ik} + [\boldsymbol{A}\,\boldsymbol{X}]_{ik}} \right)^2 \right)^{\omega_X} \right)^{1+\alpha_{sX}}$$

$$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \left(\frac{2y_{ik}}{y_{ik} + [\boldsymbol{A}\,\boldsymbol{X}]_{ik}} \right)^2 \right)^{\omega_A} \right)^{1+\alpha_{sA}}$$

$$a_{ij} \leftarrow a_{ij} / \textstyle\sum_p a_{pj}, \quad 0 < \omega_X < 2, \quad 0 < \omega_A < 2$$

Itakura-Saito distance:

$$D_{IS}(y_{ik}||z_{ik}) = \sum_{ik} \frac{y_{ik}}{z_{ik}} - \log\left(\frac{y_{ik}}{z_{ik}}\right) - 1$$

Algorithm:

$$\boldsymbol{X} \leftarrow \boldsymbol{X} \odot [(\boldsymbol{A}^T \boldsymbol{P}) \oslash (\boldsymbol{A}^T \boldsymbol{Q} + \varepsilon)].^{\beta}$$

$$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot [(\boldsymbol{P}\boldsymbol{X}^T) \oslash (\boldsymbol{Q}\boldsymbol{X}^T + \varepsilon)].^{\beta}$$

$$a_{ij} \leftarrow a_{ij} / \textstyle\sum_p a_{pj}, \ \ \beta = [0.5, 1]$$

$$\boldsymbol{P} = \boldsymbol{Y} \oslash (\boldsymbol{AX} + \varepsilon).^2, \ \ \boldsymbol{Q} = 1 \oslash (\boldsymbol{AX} + \varepsilon)$$

Table 4. Generalized SMART NMF adaptive algorithms and corresponding loss functions - part I.

Generalized SMART algorithms	
$a_{ij} \leftarrow a_{ij} \exp\left(\sum_{k=1}^{N} \tilde{\eta}_j \, x_{jk} \, \rho(y_{ik}, z_{ik})\right), \quad x_{jk} \leftarrow x_{jk} \exp\left(\sum_{i=1}^{m} \eta_j \, a_{ij} \, \rho(y_{ik}, z_{ik})\right),$ $a_j = \sum_{i=1}^{m} a_{ij} = 1, \ \forall j, \quad a_{ij} \geq 0, \quad y_{ik} > 0, \ z_{ik} = [\boldsymbol{A}\boldsymbol{X}]_{ik} > 0, \ x_{jk} \geq 0$	
Divergence: $D(\boldsymbol{Y}\|\|\boldsymbol{A}\,\boldsymbol{X})$	Error function: $\rho(y_{ik}, z_{ik})$
Dual Kullback-Leibler I-divergence: $D_{KL2}(\boldsymbol{A}\boldsymbol{X}\|\|\boldsymbol{Y})$	
$\sum_{ik} \left(z_{ik} \ln \frac{z_{ik}}{y_{ik}} + y_{ik} - z_{ik} \right),$	$\rho(y_{ik}, z_{ik}) = \ln\left(\frac{y_{ik}}{z_{ik}}\right),$
Relative Arithmetic-Geometric divergence: $D_{RAG}(\boldsymbol{Y}\|\|\boldsymbol{A}\boldsymbol{X})$	
$\sum_{ik} \left((y_{ik} + z_{ik}) \ln\left(\frac{y_{ik} + z_{ik}}{2y_{ik}}\right) + y_{ik} - z_{ik} \right),$	$\rho(y_{ik}, z_{ik}) = \ln\left(\frac{2y_{ik}}{y_{ik} + z_{ik}}\right),$
Symmetric Arithmetic-Geometric divergence: $D_{SAG}(\boldsymbol{Y}\|\|\boldsymbol{A}\boldsymbol{X})$	
$2\sum_{ik} \left(\frac{y_{ik} + z_{ik}}{2} \ln\left(\frac{y_{ik} + z_{ik}}{2\sqrt{y_{ik} z_{ik}}}\right) \right),$	$\rho(y_{ik}, z_{ik}) = \frac{y_{ik} - z_{ik}}{2z_{ik}} + \ln\left(\frac{2\sqrt{y_{ik} z_{ik}}}{y_{ik} + z_{ik}}\right),$
J-divergence: $D_J(\boldsymbol{Y}\|\|\boldsymbol{A}\boldsymbol{X})$	
$\sum_{ik} \left(\frac{y_{ik} - z_{ik}}{2} \ln\left(\frac{y_{ik}}{z_{ik}}\right) \right),$	$\rho(y_{ik}, z_{ik}) = \frac{1}{2} \ln\left(\frac{y_{ik}}{z_{ik}}\right) + \frac{y_{ik} - z_{ik}}{2z_{ik}},$

Table 5. Generalized SMART NMF adaptive algorithms and corresponding loss functions - part II.

Relative Jensen-Shannon divergence: $D_{RJS}(\boldsymbol{Y}\|\|\boldsymbol{AX})$	
$\sum_{ik}\left(2y_{ik}\ln\left(\frac{2y_{ik}}{y_{ik}+z_{ik}}\right)+z_{ik}-y_{ik}\right),$	$\rho(y_{ik},z_{ik})=\frac{y_{ik}-z_{ik}}{2z_{ik}}+\ln\left(\frac{2\sqrt{y_{ik}z_{ik}}}{y_{ik}+z_{ik}}\right),$
Dual Jensen-Shannon divergence: $D_{DJS}(\boldsymbol{Y}\|\|\boldsymbol{AX})$	
$\sum_{ik}y_{ik}\ln\left(\frac{2z_{ik}}{z_{ik}+y_{ik}}\right)+y_{ik}\ln\left(\frac{2y_{ik}}{z_{ik}+y_{ik}}\right),$	$\rho(y_{ik},z_{ik})=\ln\left(\frac{z_{ik}+y_{ik}}{2y_{ik}}\right),$
Symmetric Jensen-Shannon divergence: $D_{SJS}(\boldsymbol{Y}\|\|\boldsymbol{AX})$	
$\sum_{ik}y_{ik}\ln\left(\frac{2y_{ik}}{y_{ik}+z_{ik}}\right)+z_{ik}\ln\left(\frac{2z_{ik}}{y_{ik}+z_{ik}}\right),$	$\rho(y_{ik},z_{ik})=\ln\left(\frac{y_{ik}+z_{ik}}{2z_{ik}}\right),$
Triangular discrimination: $D_T(\boldsymbol{Y}\|\|\boldsymbol{AX})$	
$\sum_{ik}\left\{\frac{(y_{ik}-z_{ik})^2}{y_{ik}+z_{ik}}\right\},$	$\rho(y_{ik},z_{ik})=\left(\frac{2y_{ik}}{y_{ik}+z_{ik}}\right)^2-1,$
Bose-Einstein divergence: $D_{BE}(\boldsymbol{Y}\|\|\boldsymbol{AX})$	
$\sum_{ik}y_{ik}\ln\left(\frac{(1+\alpha)y_{ik}}{y_{ik}+\alpha z_{ik}}\right)+\alpha z_{ik}\ln\left(\frac{(1+\alpha)z_{ik}}{y_{ik}+\alpha z_{ik}}\right),$	$\rho(y_{ik},z_{ik})=\alpha\ln\left(\frac{y_{ik}+\alpha z_{ik}}{(1+\alpha)z_{ik}}\right),$

Early NMF Algorithms

- Alternating Least Squares
  - Paatero 1994
  - ALS algorithms that incorporate sparsity
- Multiplicative update rules
  - Lee-Seung 2000
  - Hoyer 2002
- Gradient Descent
  - Hoyer 2004
  - Berry-Plemmons 2004

PMF Algorithm: Paatero & Tapper 1994

Mean Squared Error—Alternating Least Squares

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

W = abs(randn(m,k));

for i = 1 : maxiter

LS for j = 1 : $n = \#docs$, solve

$$\min_{\mathbf{H}_{*j}} \|\mathbf{A}_{*j} - \mathbf{WH}_{*j}\|_2^2$$
$$\text{s.t. } \mathbf{H}_{*j} \geq 0$$

LS for j = 1 : $m = \#terms$, solve

$$\min_{\mathbf{W}_{j*}} \|\mathbf{A}_{j*} - \mathbf{W}_{j*}\mathbf{H}\|_2^2$$
$$\text{s.t. } \mathbf{W}_{j*} \geq 0$$

end

ALS Algorithm

W = abs(randn(m,k));

for i = 1 : maxiter

LS solve matrix equation $\mathbf{W}^T\mathbf{W}\mathbf{H} = \mathbf{W}^T\mathbf{A}$ for **H**

NONNEG $\mathbf{H} = \mathbf{H}. * (\mathbf{H} >= 0)$

LS solve matrix equation $\mathbf{H}\mathbf{H}^T\mathbf{W}^T = \mathbf{H}\mathbf{A}^T$ for **W**

NONNEG $\mathbf{W} = \mathbf{W}. * (\mathbf{W} >= 0)$

end

ALS Summary

Pros

- \+ fast
- \+ works well in practice
- \+ speedy convergence
- \+ only need to initialize $\mathbf{W}^{(0)}$
- \+ 0 elements not $locked$

Cons

- – no sparsity of **W** and **H** incorporated into mathematical setup
- – ad hoc nonnegativity: negative elements are set to 0
- – ad hoc sparsity: negative elements are set to 0
- – no convergence theory

Early NMF Algorithms

- Alternating Least Squares
  - Paatero 1994
  - ALS algorithms that incorporate sparsity
- Multiplicative update rules
  - Lee-Seung 2000
  - Hoyer 2002
- Gradient Descent
  - Hoyer 2004
  - Berry-Plemmons 2004

NMF Algorithm: Lee and Seung 2000

Mean Squared Error objective function

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

```
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!

(proof of convergence to fixed point based on E-M convergence proof)

NMF Algorithm: Lee and Seung 2000

Divergence objective function

$$\min \sum_{i,j} (\mathbf{A}_{ij} \log \frac{\mathbf{A}_{ij}}{[\mathbf{WH}]_{ij}} - \mathbf{A}_{ij} + [\mathbf{WH}]_{ij})$$

$$s.t. \quad \mathbf{W}, \mathbf{H} \geq 0$$

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

(proof of convergence to fixed point based on E-M convergence proof)

(objective function tails off after 50-100 iterations)

Multiplicative Update Summary

Pros

- \+ convergence theory: guaranteed to converge to fixed point
- \+ good initialization $\mathbf{W}^{(0)}, \mathbf{H}^{(0)}$ speeds convergence and gets to better fixed point

Cons

- – fixed point may be local min or saddle point
- – good initialization $\mathbf{W}^{(0)}, \mathbf{H}^{(0)}$ speeds convergence and gets to better fixed point
- – slow: many M-M multiplications at each iteration
- – hundreds/thousands of iterations until convergence
- – no sparsity of $\mathbf{W}$ and $\mathbf{H}$ incorporated into mathematical setup
- – 0 elements $locked$

Table 1. Amari Alpha-NMF algorithms

Amari alpha divergence: $D_A^{(\alpha)}(y_{ik}\|\|z_{ik}) = \sum_{ik} \frac{y_{ik}^{\alpha} z_{ik}^{1-\alpha} - \alpha y_{ik} + (\alpha-1) z_{ik}}{\alpha(\alpha-1)}$	
Algorithm:	$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \, (\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}})^{\alpha} \right)^{\frac{\omega_X}{\alpha}} \right)^{1+\alpha_{sX}}$
	$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \, (\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}})^{\alpha} \right)^{\frac{\omega_A}{\alpha}} \right)^{1+\alpha_{sA}}$
	$a_{ij} \leftarrow a_{ij} / \sum_p a_{pj},$
	$0 < \omega_X < 2, \quad 0 < \omega_A < 2$
Pearson distance: $(\alpha = 2)$:	$D_A^{(\alpha=2)}(y_{ik}\|\|z_{ik}) = \sum_{ik} \frac{(y_{ik} - [\boldsymbol{A}\boldsymbol{X}]_{ik})^2}{[\boldsymbol{A}\boldsymbol{X}]_{ik}},$
Algorithm:	$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \, (\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}})^{2} \right)^{\frac{\omega_X}{2}} \right)^{1+\alpha_{sX}}$
	$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \, (\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}})^{2} \right)^{\frac{\omega_A}{2}} \right)^{1+\alpha_{sA}}$
	$a_{ij} \leftarrow a_{ij} / \sum_p a_{pj},$
	$0 < \omega_X < 2, \quad 0 < \omega_A < 2$
Hellinger distance: $(\alpha = \frac{1}{2})$:	$D_A^{(\alpha=0.5)}(y_{ik}\|\|z_{ik}) = \sum_{ik} \frac{(y_{ik} - [\boldsymbol{A}\boldsymbol{X}]_{ik})^2}{[\boldsymbol{A}\boldsymbol{X}]_{ik}},$
Algorithm:	$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \sqrt{\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}}} \right)^{2\omega_X} \right)^{1+\alpha_{sX}}$
	$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \sqrt{\frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}}} \right)^{2\omega_A} \right)^{1+\alpha_{sA}}$
	$a_{ij} \leftarrow a_{ij} / \sum_p a_{pj},$
	$0 < \omega_X < 2, \quad 0 < \omega_A < 2$

Table 2. Amari Alpha-NMF algorithms (continued)

Kullback-Leibler divergence: ($\alpha \to 1$):

$$\lim_{\alpha\to 1} D_A^{(\alpha)}(y_{ik}||z_{ik}) = \sum_{ik} y_{ik} \log \frac{y_{ik}}{[\boldsymbol{AX}]_{ik}} - y_{ik} + [\boldsymbol{AX}]_{ik},$$

Algorithm:

$$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}} \right)^{\omega_X} \right)^{1+\alpha_{sX}}$$

$$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \frac{y_{ik}}{[\boldsymbol{A}\,\boldsymbol{X}]_{ik}} \right)^{\omega_A} \right)^{1+\alpha_{sA}}$$

$$a_{ij} \leftarrow a_{ij} / \textstyle\sum_p a_{pj}$$

$$0 < \omega_X < 2, \quad 0 < \omega_A < 2$$

Dual Kullback-Leibler divergence: ($\alpha \to 0$):

$$\lim_{\alpha\to 0} D_A^{(\alpha)}(y_{ik}||z_{ik}) = \sum_{ik} [\boldsymbol{AX}]_{ik} \log \frac{[\boldsymbol{AX}]_{ik}}{y_{ik}} + y_{ik} - [\boldsymbol{AX}]_{ik}$$

Algorithm:

$$x_{jk} \leftarrow \left(x_{jk} \prod_{i=1}^{m} \left(\frac{y_{ik}}{[\boldsymbol{AX}]_{ik}} \right)^{\omega_X\, a_{ij}} \right)^{1+\alpha_{sX}}$$

$$a_{ij} \leftarrow \left(a_{ij} \prod_{k=1}^{N} \left(\frac{y_{ik}}{[\boldsymbol{AX}]_{ik}} \right)^{\tilde{\eta}_j x_{jk}} \right)^{1+\alpha_{sA}}$$

$$a_{ij} \leftarrow a_{ij} / \textstyle\sum_p a_{pj}$$

$$0 < \omega_X < 2, \quad 0 < \omega_A < 2$$

Table 3. Other generalized NMF algorithms

Beta generalized divergence:

$$D_K^{(\beta)}(y_{ik}||z_{ik}) = \sum_{ik} y_{ik} \frac{y_{ik}^{\beta-1} - [\boldsymbol{AX}]_{ik}^{\beta-1}}{\beta(\beta-1)} + [\boldsymbol{AX}]_{ik}^{\beta-1} \frac{[\boldsymbol{AX}]_{ik} - y_{ik}}{\beta}$$

Kompass algorithm:

$$x_{jk} \leftarrow x_{jk} \frac{\sum_{i=1}^{m} a_{ij}\ (y_{ik}/[\boldsymbol{AX}]_{ik}^{2-\beta})}{\sum_{i=1}^{m} a_{ij}\ [\boldsymbol{AX}]_{ik}^{\beta-1} + \varepsilon}$$

$$a_{ij} \leftarrow \left(a_{ij} \frac{\sum_{k=1}^{N} x_{jk}\ (y_{ik}/[\boldsymbol{AX}]_{ik}^{2-\beta})}{\sum_{k=1}^{N} x_{jk}\ [\boldsymbol{AX}]_{ik}^{\beta-1} + \varepsilon} \right)^{1+\alpha_{sA}}$$

$$a_{ij} \leftarrow a_{ij} / \textstyle\sum_p a_{pj},$$

Triangular discrimination:

$$D_T^{(\beta)}(y_{ik}||z_{ik}) = \sum_{ik} \frac{y_{ik}^{\beta} z_{ik}^{1-\beta} - \beta y_{ik} + (\beta-1) z_{ik}}{\beta(\beta-1)}$$

Algorithm:

$$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \left(\frac{2y_{ik}}{y_{ik} + [\boldsymbol{A}\,\boldsymbol{X}]_{ik}} \right)^2 \right)^{\omega_X} \right)^{1+\alpha_{sX}}$$

$$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \left(\frac{2y_{ik}}{y_{ik} + [\boldsymbol{A}\,\boldsymbol{X}]_{ik}} \right)^2 \right)^{\omega_A} \right)^{1+\alpha_{sA}}$$

$$a_{ij} \leftarrow a_{ij} / \textstyle\sum_p a_{pj}, \quad 0 < \omega_X < 2, \quad 0 < \omega_A < 2$$

Itakura-Saito distance:

$$D_{IS}(y_{ik}||z_{ik}) = \sum_{ik} \frac{y_{ik}}{z_{ik}} - \log\left(\frac{y_{ik}}{z_{ik}}\right) - 1$$

Algorithm:

$$\boldsymbol{X} \leftarrow \boldsymbol{X} \odot [(\boldsymbol{A}^T \boldsymbol{P}) \oslash (\boldsymbol{A}^T \boldsymbol{Q} + \varepsilon)].^{\beta}$$

$$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot [(\boldsymbol{P}\boldsymbol{X}^T) \oslash (\boldsymbol{Q}\boldsymbol{X}^T + \varepsilon)].^{\beta}$$

$$a_{ij} \leftarrow a_{ij} / \textstyle\sum_p a_{pj}, \ \ \beta = [0.5, 1]$$

$$\boldsymbol{P} = \boldsymbol{Y} \oslash (\boldsymbol{AX} + \varepsilon).^2, \ \ \boldsymbol{Q} = 1 \oslash (\boldsymbol{AX} + \varepsilon)$$

Table 4. Generalized SMART NMF adaptive algorithms and corresponding loss functions - part I.

Generalized SMART algorithms	
$a_{ij} \leftarrow a_{ij} \exp\left(\sum_{k=1}^{N} \tilde{\eta}_j \, x_{jk} \, \rho(y_{ik}, z_{ik})\right), \quad x_{jk} \leftarrow x_{jk} \exp\left(\sum_{i=1}^{m} \eta_j \, a_{ij} \, \rho(y_{ik}, z_{ik})\right),$ $a_j = \sum_{i=1}^{m} a_{ij} = 1, \ \forall j, \quad a_{ij} \geq 0, \quad y_{ik} > 0, \ z_{ik} = [\boldsymbol{A}\boldsymbol{X}]_{ik} > 0, \ x_{jk} \geq 0$	
Divergence: $D(\boldsymbol{Y}\|\|\boldsymbol{A}\,\boldsymbol{X})$	Error function: $\rho(y_{ik}, z_{ik})$
Dual Kullback-Leibler I-divergence: $D_{KL2}(\boldsymbol{A}\boldsymbol{X}\|\|\boldsymbol{Y})$	
$\sum_{ik} \left(z_{ik} \ln \frac{z_{ik}}{y_{ik}} + y_{ik} - z_{ik} \right),$	$\rho(y_{ik}, z_{ik}) = \ln\left(\frac{y_{ik}}{z_{ik}}\right),$
Relative Arithmetic-Geometric divergence: $D_{RAG}(\boldsymbol{Y}\|\|\boldsymbol{A}\boldsymbol{X})$	
$\sum_{ik} \left((y_{ik} + z_{ik}) \ln\left(\frac{y_{ik} + z_{ik}}{2y_{ik}}\right) + y_{ik} - z_{ik} \right),$	$\rho(y_{ik}, z_{ik}) = \ln\left(\frac{2y_{ik}}{y_{ik} + z_{ik}}\right),$
Symmetric Arithmetic-Geometric divergence: $D_{SAG}(\boldsymbol{Y}\|\|\boldsymbol{A}\boldsymbol{X})$	
$2\sum_{ik} \left(\frac{y_{ik} + z_{ik}}{2} \ln\left(\frac{y_{ik} + z_{ik}}{2\sqrt{y_{ik} z_{ik}}}\right) \right),$	$\rho(y_{ik}, z_{ik}) = \frac{y_{ik} - z_{ik}}{2z_{ik}} + \ln\left(\frac{2\sqrt{y_{ik} z_{ik}}}{y_{ik} + z_{ik}}\right),$
J-divergence: $D_J(\boldsymbol{Y}\|\|\boldsymbol{A}\boldsymbol{X})$	
$\sum_{ik} \left(\frac{y_{ik} - z_{ik}}{2} \ln\left(\frac{y_{ik}}{z_{ik}}\right) \right),$	$\rho(y_{ik}, z_{ik}) = \frac{1}{2} \ln\left(\frac{y_{ik}}{z_{ik}}\right) + \frac{y_{ik} - z_{ik}}{2z_{ik}},$

Table 5. Generalized SMART NMF adaptive algorithms and corresponding loss functions - part II.

Relative Jensen-Shannon divergence: $D_{RJS}(\boldsymbol{Y}\|\|\boldsymbol{AX})$	
$\sum_{ik}\left(2y_{ik}\ln\left(\frac{2y_{ik}}{y_{ik}+z_{ik}}\right)+z_{ik}-y_{ik}\right),$	$\rho(y_{ik},z_{ik})=\frac{y_{ik}-z_{ik}}{2z_{ik}}+\ln\left(\frac{2\sqrt{y_{ik}z_{ik}}}{y_{ik}+z_{ik}}\right),$
Dual Jensen-Shannon divergence: $D_{DJS}(\boldsymbol{Y}\|\|\boldsymbol{AX})$	
$\sum_{ik}y_{ik}\ln\left(\frac{2z_{ik}}{z_{ik}+y_{ik}}\right)+y_{ik}\ln\left(\frac{2y_{ik}}{z_{ik}+y_{ik}}\right),$	$\rho(y_{ik},z_{ik})=\ln\left(\frac{z_{ik}+y_{ik}}{2y_{ik}}\right),$
Symmetric Jensen-Shannon divergence: $D_{SJS}(\boldsymbol{Y}\|\|\boldsymbol{AX})$	
$\sum_{ik}y_{ik}\ln\left(\frac{2y_{ik}}{y_{ik}+z_{ik}}\right)+z_{ik}\ln\left(\frac{2z_{ik}}{y_{ik}+z_{ik}}\right),$	$\rho(y_{ik},z_{ik})=\ln\left(\frac{y_{ik}+z_{ik}}{2z_{ik}}\right),$
Triangular discrimination: $D_T(\boldsymbol{Y}\|\|\boldsymbol{AX})$	
$\sum_{ik}\left\{\frac{(y_{ik}-z_{ik})^2}{y_{ik}+z_{ik}}\right\},$	$\rho(y_{ik},z_{ik})=\left(\frac{2y_{ik}}{y_{ik}+z_{ik}}\right)^2-1,$
Bose-Einstein divergence: $D_{BE}(\boldsymbol{Y}\|\|\boldsymbol{AX})$	
$\sum_{ik}y_{ik}\ln\left(\frac{(1+\alpha)y_{ik}}{y_{ik}+\alpha z_{ik}}\right)+\alpha z_{ik}\ln\left(\frac{(1+\alpha)z_{ik}}{y_{ik}+\alpha z_{ik}}\right),$	$\rho(y_{ik},z_{ik})=\alpha\ln\left(\frac{y_{ik}+\alpha z_{ik}}{(1+\alpha)z_{ik}}\right),$

Early NMF Algorithms

- Alternating Least Squares
  - Paatero 1994
  - ALS algorithms that incorporate sparsity
- Multiplicative update rules
  - Lee-Seung 2000
  - Hoyer 2002
- Gradient Descent
  - Hoyer 2004
  - Berry-Plemmons 2004

Sparsity Measures

- Berry et al. $\|\mathbf{x}\|_2^2$
- Hoyer $spar(\mathbf{x}_{n\times 1}) = \frac{\sqrt{n} - \|\mathbf{x}\|_1/\|\mathbf{x}\|_2}{\sqrt{n}-1}$
- Diversity measure $E^{(p)}(\mathbf{x}) = \sum_{i=1}^{n} |x_i|^p,\ 0 \leq p \leq 1$

  $E^{(p)}(\mathbf{x}) = -\sum_{i=1}^{n} |x_i|^p,\ p < 0$

  Rao and Kreutz-Delgado: algorithms for minimizing $E^{(p)}(\mathbf{x})$ s.t. $\mathbf{Ax} = \mathbf{b}$, but expensive iterative procedure
- Ideal $nnz(\mathbf{x})$ not continuous, NP-hard to use this in optim.

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

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

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

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 for **H**: $(\mathbf{W}^T\mathbf{W} + \lambda\,\mathbf{I})\,\mathbf{H} = \mathbf{W}^T\mathbf{A}$; (small matrix 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

(objective function tails off after 15-30 iterations)

Berry et al. 2004 Summary

Pros

+ fast: less work per iteration than most other NMF algorithms
+ fast: small # of iterations until convergence
+ sparsity parameter for **H**

Cons

– 0 elements in **W** are $locked$
– no sparsity parameter for **W**
– ad hoc nonnegativity: negative elements in **H** are set to 0, could run lsqnonneg or snnls instead
– no convergence theory