Probabilistic Latent Semantic Analysis (PLSA)

Motivation

- Does orthogonally matter?
- Wouldn't a sound statistical foundation be better?

PLSA

Likelihood of document

$$P(doc) = P(term_1 | doc)P(term_2 | doc)...P(term_L | doc)$$

Introduce term-frequency matrix X

$$\prod_{l=1}^{L} P(term_l | doc) = \prod_{t=1}^{T} P(term_t | doc)^{A(term_t, doc)}$$

PLSA

Introduce hidden topic

$$P(term_t \mid doc) = \sum_{k=1}^{K} P(term_t \mid topic_k) P(topic_k \mid doc)$$

Shorthand t=term$_t$

$$P(t \mid doc) = \sum_{k=1}^{K} P(t \mid k) P(k \mid doc)$$

Relation to LSA?

Likelihood of document

$$P(doc) = \prod_{t=1}^{T} \left\{ \sum_{k=1}^{K} P(t \mid k) P(k \mid doc) \right\}^{A(t,doc)}$$

PLSA: training

Training objective function

$$\sum_{d=1}^{N} \log P(d) = \sum_{d=1}^{N} \sum_{t=1}^{T} A(t,d) \log \sum_{k=1}^{K} P(t \mid k) P(k \mid d)$$

which is to be maximised w.r.t. parameters P$(t \mid k)$ and then also P$(k \mid d)$,

subject to the constraints that $\sum_{t=1}^{T} P(t \mid k) = 1$ and $\sum_{k=1}^{K} P(k \mid d) = 1$.

PLSA: training

Update term-topic matrix

$$P1(t,k) \leftarrow P1(t,k) \sum_{d=1}^{N} \frac{A(t,d)}{\sum_{k=1}^{K} P1(t,k)P2(k,d)} P2(k,d)$$

$$P1(t,k) \leftarrow \frac{P1(t,k)}{\sum_{t=1}^{T} P1(t,k)}$$

Update topic-document matrix

$$P2(k,d) \leftarrow P2(k,d) \sum_{t=1}^{T} \frac{A(t,d)}{\sum_{k=1}^{K} P1(t,k)P2(k,d)} P1(t,k)$$

$$P2(k,d) \leftarrow \frac{P2(k,d)}{\sum_{k=1}^{K} P2(k,d)}$$

PLSA

P(t|k) for some topics

universe	0.0439	drug	0.0672	cells	0.0675	sequence	0.0818	years	0.156
galaxies	0.0375	patients	0.0493	stem	0.0478	sequences	0.0493	million	0.0556
clusters	0.0279	drugs	0.0444	human	0.0421	genome	0.033	ago	0.045
matter	0.0233	clinical	0.0346	cell	0.0309	dna	0.0257	time	0.0317
galaxy	0.0232	treatment	0.028	gene	0.025	sequencing	0.0172	age	0.0243
cluster	0.0214	trials	0.0277	tissue	0.0185	map	0.0123	year	0.024
cosmic	0.0137	therapy	0.0213	cloning	0.0169	genes	0.0122	record	0.0238
dark	0.0131	trial	0.0164	transfer	0.0155	chromosome	0.0119	early	0.0233
light	0.0109	disease	0.0157	blood	0.0113	regions	0.0119	billion	0.0177
density	0.01	medical	0.00997	embryos	0.0111	human	0.0111	history	0.0148
bacteria	0.0983	male	0.0558	theory	0.0811	immune	0.0909	stars	0.0524
bacterial	0.0561	females	0.0541	physics	0.0782	response	0.0375	star	0.0458
resistance	0.0431	female	0.0529	physicists	0.0146	system	0.0358	astrophys	0.0237
coli	0.0381	males	0.0477	einstein	0.0142	responses	0.0322	mass	0.021
strains	0.025	sex	0.0339	university	0.013	antigen	0.0263	disk	0.0173
microbiol	0.0214	reproductive	0.0172	gravity	0.013	antigens	0.0184	black	0.0161
microbial	0.0196	offspring	0.0168	black	0.0127	immunity	0.0176	gas	0.0149
strain	0.0165	sexual	0.0166	theories	0.01	immunology	0.0145	stellar	0.0127
salmonella	0.0163	reproduction	0.0143	aps	0.00987	antibody	0.014	astron	0.0125
resistant	0.0145	eggs	0.0138	matter	0.00954	autoimmune	0.0128	hole	0.00824

Comparison LSA and PLSA
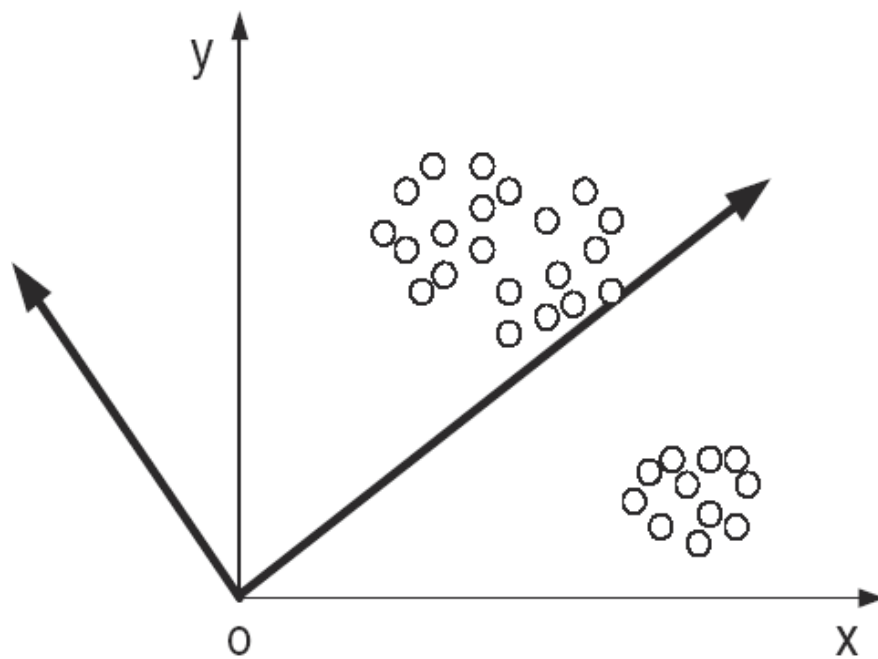

From Th. Hofmann, 2000
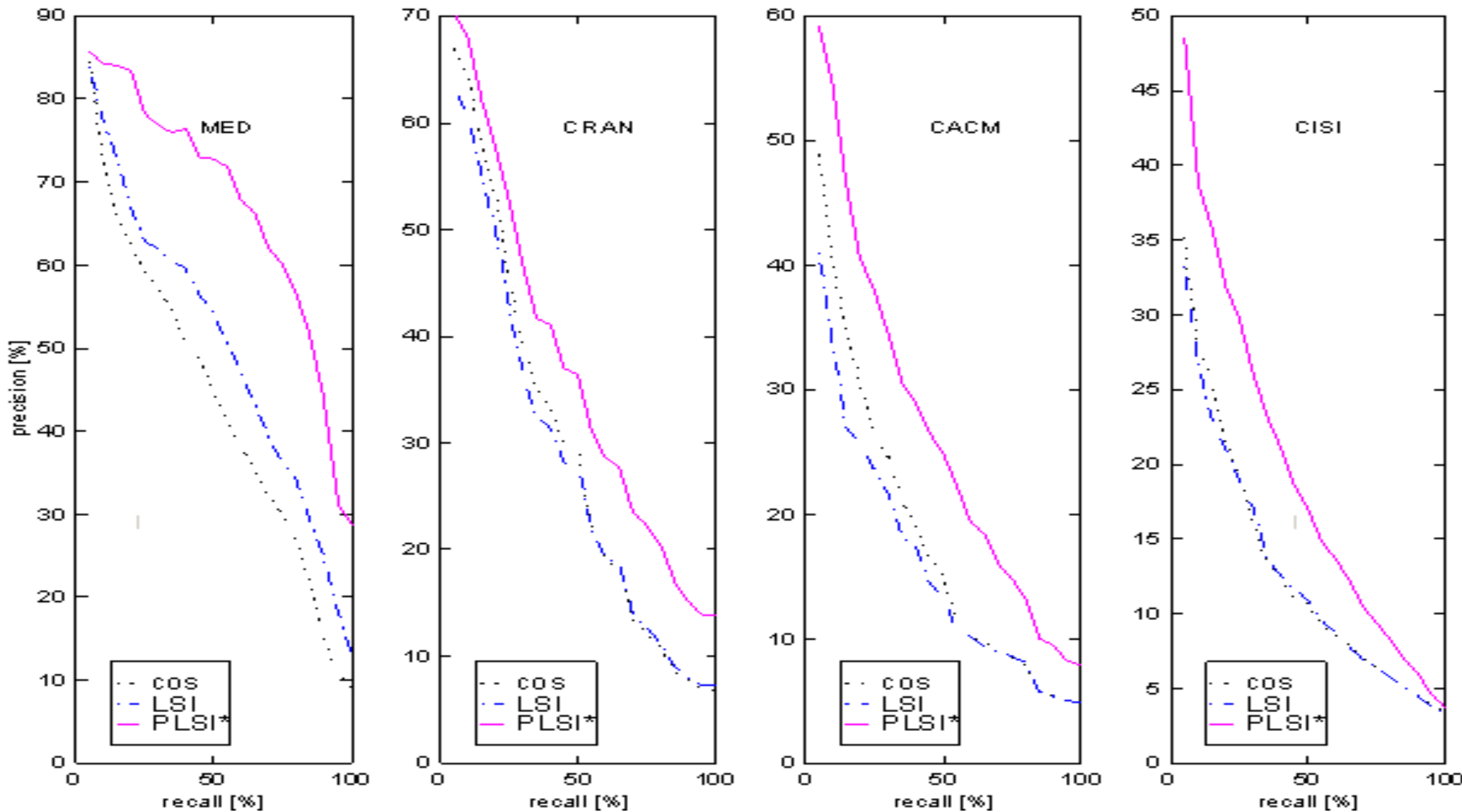
Non-negative Matrix Factorization

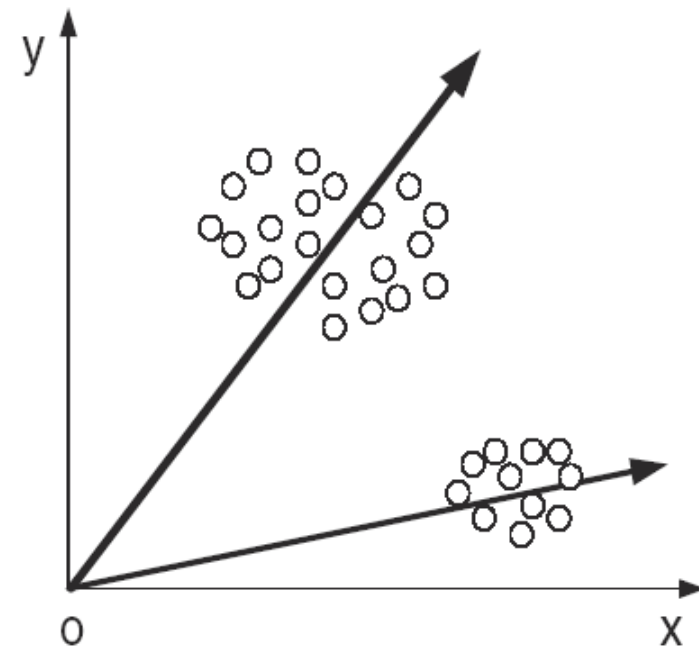See: **Document Clustering Based On Non-negative Matrix Factorization**

Wei Xu, Xin Liu, Yihong Gong
NEC Laboratories America, Inc.
10080 North Wolfe Road, SW3-350
Cupertino, CA 95014, U.S.A.
{xw,xliu,ygong}@ccrl.sj.nec.com

NMF: idea

- Find space that separates clusters better

Directions found by LSI

Directions found by NMF

NMF: the model

- Decompostion of a non-negative matrix X into two matrices W and H, both non-negative

$$A = WH$$

- A: N x M – data matrix
- W: N x R – source matrix
- H: R x M – mixture matrix

NMF: the model

- Determine W and H such that the product WH is as close as possible to A

- W and H are bound to be non-negative values

- Possible metrics
  - Kullback-Leibler-Divergence
  - Frobenius-Norm

$$D(A|WH)$$

$$\frac{1}{2}|A - WH|^2$$

Update

$$H_{ab} = H_{ab} \frac{\left(W^t A\right)_{ab}}{\left(W^t W H\right)_{ab} + \varepsilon}$$

$$W_{ab} = W_{ab} \frac{\left(A H^t\right)_{ab}}{\left(W H H^t\right)_{ab} + \varepsilon}$$

Relation to update
From PLSA?

In case the denominator vanishes, add a small number

Homework

Implement NMF.
Decompose the matrix from exercise 9.
Details in exercise 10.

Summary

Ways to find latent "semantic" spaces:

- LSA
- PLSA
- NMF

Similar factorizations

Different target functions and constraints