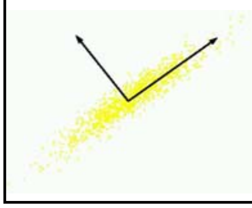
Machine Learning

10-701/15-781, Fall 2011

Alternative Strategies of Learning (1)

PCA versus Topic models:

nonprobabilistic vs. probabilistic approach for subspace learning



Eric Xing

Lecture 16, November 2, 2011

Elements of Learning

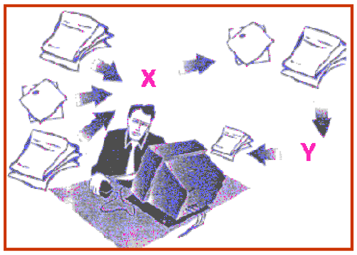
- **Here are some important elements to consider before you start:**
  - Task:
    - Embedding? Classification? Clustering? Topic extraction? …
  - Data and other info:
    - Input and output (e.g., continuous, binary, counts, …)
    - Supervised or unsupervised, of a blend of everything?
    - Prior knowledge? Bias?
  - Models and paradigms:
    - BN? MRF? Regression? SVM?
    - Bayesian/Frequents ? Parametric/Nonparametric?
  - Objective/Loss function:
    - MLE? MCLE? Max margin?
    - Log loss, hinge loss, square loss? …
  - Tractability and exactness trade off:
    - Exact inference? MCMC? Variational? Gradient? Greedy search?
    - Online? Batch? Distributed?
  - Evaluation:
    - Visualization? Human interpretability? Perperlexity? Predictive accuracy?
- **It is better to consider one element at a time!**

Learning Graphical Models

- Scenarios:
  - completely observed GMs
    - directed
    - undirected
  - partially observed GMs
    - directed
    - undirected (an open research topic)
- Estimation principles:
  - Maximal likelihood or conditional likelihood estimation (MLE, MLCE)
  - Bayesian estimation
  - Maximal "Margin"
  - ….
- We use **learning** as a name for the process of estimating the parameters, and in some cases, the topology of the network, from data.

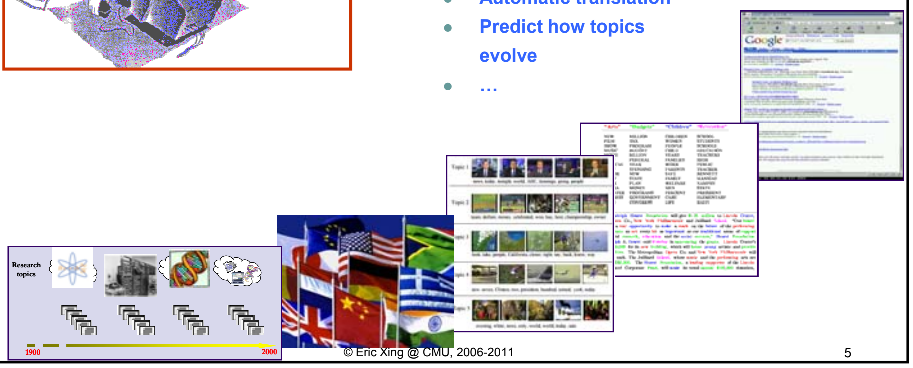
nonprobabilistic vs. probabilistic approach for subspace learning

The Problem: NLP and Data Mining



We want:

- Semantic-based search
- infer topics and categorize documents
- Multimedia inference
- Automatic translation
- Predict how topics evolve
- …

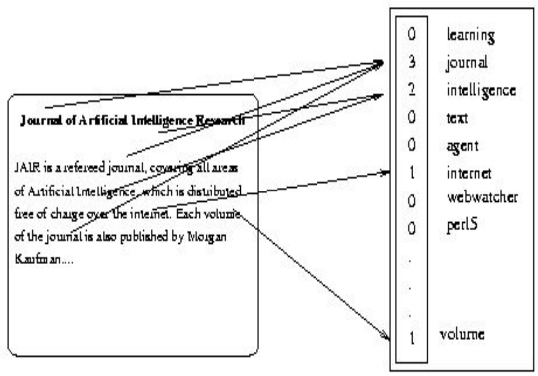


Modeling document collections

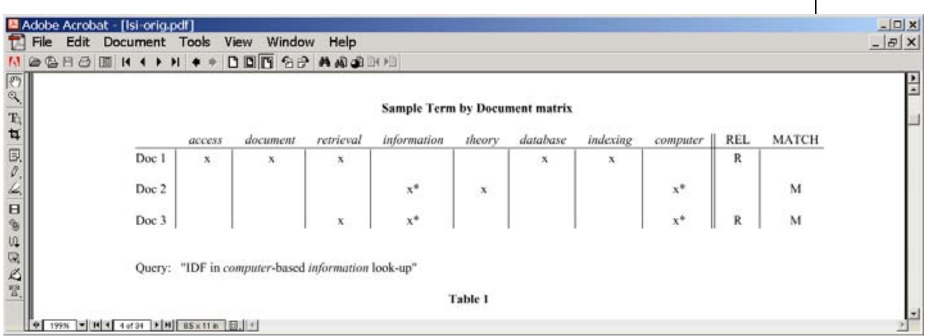
- A document collection is a dataset where each data point is itself a collection of simpler data.
  - Text documents are collections of words.
  - Segmented images are collections of regions.
  - User histories are collections of purchased items.
- Many modern problems ask questions of such data.
  - Is this text document relevant to my query?
  - Which documents are about a particular topic?
  - How have topics changed over time?
  - What does author X write about? Who is likely to write about topic Y? Who wrote this specific document?
  - Which category is this image in? Create a caption for this image.
  - What movies would I probably like?
  - and so on…..

Text document retrieval

- Represent each document by a high-dimensional vector in the space of words



Example



Sample Term by Document matrix

	access	*document*	*retrieval*	*information*	*theory*	*database*	*indexing*	*computer*	REL	MATCH
Doc 1	x	x	x			x	x		R	
Doc 2				x*	x			x*		M
Doc 3			x	x*				x*	R	M

Query: "IDF in *computer*-based *information* look-up"

Table 1

-- Relevant docs may not have the query terms
→ but may have many "related" terms
-- Irrelevant docs may *have* the query terms
→ but may not have any "related" terms

Problems

- Looks for literal term matches
  - Terms in queries (esp short ones) don't always capture user's information need well
- Problems:
  - **Synonymy**: other words with the same meaning
    - Car and automobile
  - No associations between words are made in the vector space representation.

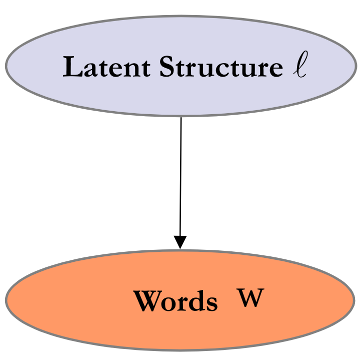
$$\mathrm{sim}_{\mathrm{true}}(d,q) > \cos(\angle(\vec{d},\vec{q}))$$

  - **Polysemy**: the same word having other meanings
    - Apple (fruit and company)
  - The vector space model is unable to discriminate between different meanings of the same word.

$$\mathrm{sim}_{\mathrm{true}}(d,q) < \cos(\angle(\vec{d},\vec{q}))$$

- What if we could match against 'concepts', that represent related words, rather than words themselves

Subspace Learning



Latent Semantic Indexing (LSI)

(Deerwester et al., 1990)

- Uses statistically derived conceptual indices instead of individual words for retrieval
- Assumes that there is some underlying or *latent* structure in word usage that is obscured by variability in word choice
- Key idea: instead of representing documents and queries as vectors in a t-dim space of terms
  - Represent them (and terms themselves) as vectors in a lower-dimensional space whose axes are concepts that effectively group together similar words
  - Uses SVD to reduce document representations,
  - The axes are the Principal Components from SVD (singular value decomposition)
- So what is SVD?

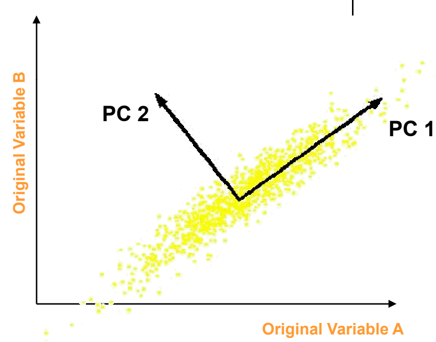
Basic Concept

- Areas of variance in data are where items can be best discriminated and key underlying phenomena observed
- If two items or dimensions are highly correlated or dependent
  - They are likely to represent highly related phenomena
  - If they tell us about the same underlying variance in the data, combining them to form a single measure is reasonable
    - Parsimony
    - Reduction in Error
  - We want to combine related variables, and focus on uncorrelated or independent ones, especially those along which the observations have high variance
- We look for the phenomena underlying the observed covariance/co-dependence in a set of variables
- These phenomena are called “factors” or “principal components” or “independent components,” depending on the methods used
  - Factor analysis: based on variance/covariance/correlation
  - Independent Component Analysis: based on independence

An example:

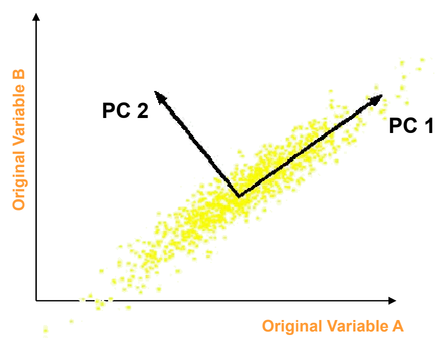
Principal Component Analysis

- Most common form of factor analysis
- The new variables/dimensions
  - Are linear combinations of the original ones
  - Are uncorrelated with one another
    - Orthogonal in original dimension space
  - Capture as much of the original variance in the data as possible
  - Are called Principal Components



- Orthogonal directions of greatest variance in data
- Projections along PC1 discriminate the data most along any one axis

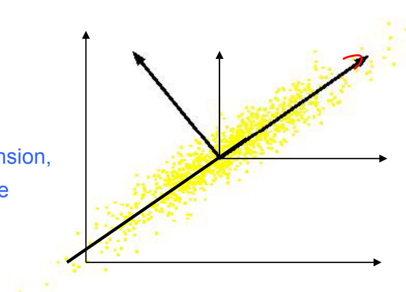
Principal Component Analysis



- First principal component is the direction of greatest variability (covariance) in the data
- Second is the next orthogonal (uncorrelated) direction of greatest variability
  - So first remove all the variability along the first component, and then find the next direction of greatest variability
- And so on …

Computing (learning) the Components

- Data points are vectors in a multidimensional space
- Projection of vector **x** onto an axis (dimension) **u** is $\mathbf{u}^T\mathbf{x}$
- Direction of greatest variability is that in which the average square of the projection is greatest
  - I.e. **u** such that $E((\mathbf{u}^T\mathbf{x})^2)$ over **all x** is maximized
  - Matrix representation:
  - (we subtract the mean along each dimension, and center the original axis system at the centroid of all data points, for simplicity)
  - This direction of **u** is the direction of the first Principal Component



Computing the Components

- $E((\mathbf{u}^T\mathbf{x})^2) = \Sigma_i(\mathbf{u}^T\mathbf{x}_i)^2/m = (\mathbf{u}^T\mathbf{X})(\mathbf{u}^T\mathbf{X})^T/m = \mathbf{u}^T(\mathbf{X}\mathbf{X}^T/m)\mathbf{u}$
- The **covariance matrix** $\mathbf{C} = \mathbf{X}\mathbf{X}^T$ contains the correlations (similarities) of the original axes based on how the data values project onto them
- So we are looking for **u** that maximizes $\mathbf{u}^T\mathbf{C}\mathbf{u}$, subject to **u** being unit-length
- It is maximized when **u** is the principal eigenvector of the matrix **C**, in which case
  - $\mathbf{u}^T\mathbf{C}\mathbf{u} = \mathbf{u}^T\lambda\mathbf{u} = \lambda$ if **u** is unit-length, where λ is the principal eigenvalue of the correlation matrix C
  - The eigenvalue denotes the amount of variability captured along that dimension

Why the Eigenvectors?

Maximise $\mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u}$

s.t $\mathbf{u}^T\mathbf{u} = 1$

Construct Langrangian $\mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u} - \lambda\mathbf{u}^T\mathbf{u}$

Vector of partial derivatives set to zero

$$\mathbf{x}\mathbf{x}^T\mathbf{u} - \lambda\mathbf{u} = (\mathbf{x}\mathbf{x}^T - \lambda\mathbf{I})\,\mathbf{u} = 0$$

As $\mathbf{u} \neq \mathbf{0}$ then **u** must be an eigenvector of $\mathbf{X}\mathbf{X}^T$ with eigenvalue λ

Eigenvalues & Eigenvectors

- For symmetric matrices, eigenvectors for distinct eigenvalues are **orthogonal**

$$Sv_{\{1,2\}} = \lambda_{\{1,2\}} v_{\{1,2\}}, \text{ and } \lambda_1 \neq \lambda_2 \Rightarrow v_1 \bullet v_2 = 0$$

- All eigenvalues of a real symmetric matrix are **real**.

$$\text{if } |S - \lambda I| = 0 \text{ and } \mathrm{S} = \mathrm{S}^{\mathrm{T}} \Rightarrow \lambda \in \Re$$

- All eigenvalues of a positive semidefinite matrix are **non-negative**

$$\forall w \in \Re^n, w^T S w \geq 0, \text{ then if } \mathrm{Sv} = \lambda \mathrm{v} \Rightarrow \lambda \geq 0$$

Eigen/diagonal Decomposition

- Let $\mathbf{S} \in \mathbb{R}^{m \times m}$ be a **square** matrix with ***m*** **linearly independent eigenvectors** (a "non-defective" matrix)

- **Theorem**: Exists an **eigen decomposition**

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

diagonal

Unique for distinct eigen-values

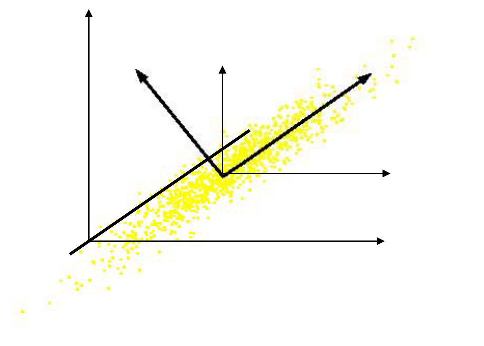
(cf. matrix diagonalization theorem)

- Columns of ***U*** are **eigenvectors** of ***S***
- Diagonal elements of $\mathbf{\Lambda}$ are **eigenvalues** of $\mathbf{S}$

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \quad \lambda_i \geq \lambda_{i+1}$$

Computing the Components

- So, the new axes are the eigenvectors of the matrix of correlations of the original variables, which captures the similarities of the original variables based on how data samples



- Geometri
  - Linear

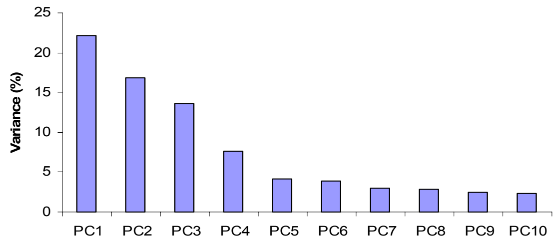
PCs, Variance and Least-Squares

- The first PC retains the greatest amount of variation in the sample
- The k^{th} PC retains the kth greatest fraction of the variation in the sample
- The k^{th} largest eigenvalue of the correlation matrix C is the variance in the sample along the k^{th} PC
- The least-squares view: PCs are a series of linear least squares fits to a sample, each orthogonal to all previous ones

How Many PCs?

- For n original dimensions, sample covariance matrix is nxn, and has up to n eigenvectors. So n PCs.
- Where does dimensionality reduction come from?

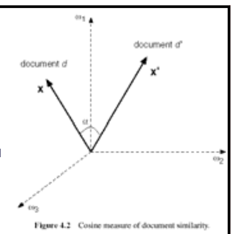
  Can *ignore* the components of lesser significance.

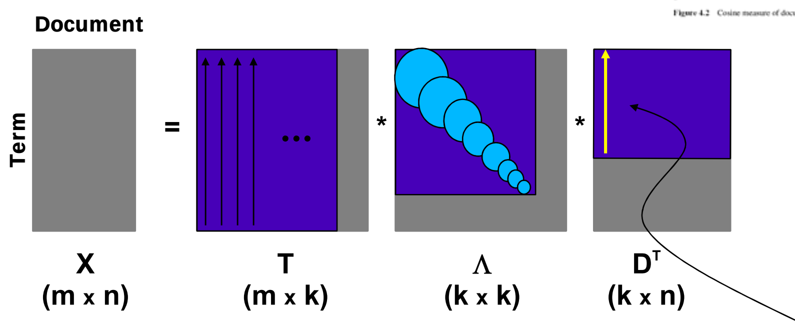


You do lose some information, but if the eigenvalues are small, you don't lose much

- n dimensions in original data
- calculate n eigenvectors and eigenvalues
- choose only the first p eigenvectors, based on their eigenvalues
- final data set has only p dimensions

Latent Semantic Indexing





This is our compressed representation of a document

$$\vec{w} = \sum_{k=1}^{K} d_k \lambda_k \vec{T}_k$$

Recall: Eigen/diagonal decomposition

- Let $\mathbf{S} \in \mathbb{R}^{m \times m}$ be a **square** matrix with ***m*** **linearly independent eigenvectors** (a "non-defective" matrix)

- **Theorem**: Exists an **eigen decomposition**

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

diagonal

Unique for distinct eigen-values

(cf. matrix diagonalization theorem)

- Columns of ***U*** are **eigenvectors** of ***S***
- Diagonal elements of $\mathbf{\Lambda}$ are **eigenvalues** of $\mathbf{S}$

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \quad \lambda_i \geq \lambda_{i+1}$$

Singular Value Decomposition

For an $m \times n$ matrix A of rank r there exists a factorization (Singular Value Decomposition = SVD) as follows:

$$A = U \Sigma V^T$$

$m \times m$ | $m \times m$ | V is $m \times n$

The columns of *U* are orthogonal eigenvectors of AA^T.

The columns of *V* are orthogonal eigenvectors of A^TA.

Eigenvalues $\lambda_1 \ldots \lambda_r$ of AA^T are the eigenvalues of A^TA.

$$\sigma_i = \sqrt{\lambda_i}$$

$$\Sigma = diag(\sigma_1 \ldots \sigma_r)$$

Singular values.

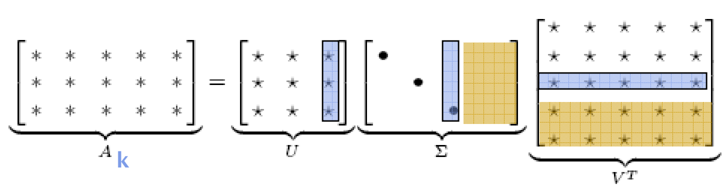
SVD and PCA

- The first root is called the prinicipal eigenvalue which has an associated orthonormal ($\mathbf{u}^T\mathbf{u} = 1$) *eigenvector* $\mathbf{u}$
- Subsequent roots are ordered such that $\lambda_1 > \lambda_2 > \ldots > \lambda_M$ with rank($\mathbf{D}$) non-zero values.
- Eigenvectors form an orthonormal basis i.e. $\mathbf{u}_i^T\mathbf{u}_j = \delta_{ij}$
- The eigenvalue decomposition of $\mathbf{XX}^T = \mathbf{U\Sigma U}^T$
- where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M]$ and $\mathbf{\Sigma} = \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_M]$
- Similarly the eigenvalue decomposition of $\mathbf{X}^T\mathbf{X} = \mathbf{V\Sigma V}^T$
- The SVD is closely related to the above $\mathbf{X}=\mathbf{U}\ \mathbf{\Sigma}^{1/2}\ \mathbf{V}^T$
- The left eigenvectors **U,** right eigenvectors **V,**
- singular values = square root of eigenvalues**.**

Low-rank Approximation

- Solution via SVD

$$A_k = U\ \mathrm{diag}(\sigma_1,\ldots,\sigma_k,\underbrace{0,\ldots,0}_{\substack{\textit{set smallest r-k}\\ \textit{singular values to zero}}})V^T$$



$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T \longleftarrow$$

column notation: sum of rank 1 matrices

Approximation error

- How good (bad) is this approximation?
- It's the best possible, measured by the Frobenius norm of the error:

$$\min_{X:rank(X)=k} \|A - X\|_F = \|A - A_k\|_F = \sigma_{k+1}$$

where the σ_i are ordered such that $\sigma_i \geq \sigma_{i+1}$.

Suggests why Frobenius error drops as *k* increased.

Example

term	ch2	ch3	ch4	ch5	ch6	ch7	ch8	ch9
controllability	1	1	0	0	1	0	0	1
observability	1	0	0	0	1	1	0	1
realization	1	0	1	0	1	0	1	0
feedback	0	1	0	0	0	1	0	0
controller	0	1	0	0	1	1	0	0
observer	0	1	1	0	1	1	0	0
transfer function	0	0	0	0	1	1	0	0
polynomial	0	0	0	0	1	0	1	0
matrices	0	0	0	0	1	0	1	1

U (9x7) =

0.3996	-0.1037	0.5606	-0.3717	-0.3919	-0.3482	0.1029
0.4180	-0.0641	0.4878	0.1566	0.5771	0.1981	-0.1094
0.3464	-0.4422	-0.3997	-0.5142	0.2787	0.0102	-0.2857
0.1888	0.4615	0.0049	-0.0279	-0.2087	0.4193	-0.6629
0.3602	0.3776	-0.0914	0.1596	-0.2045	-0.3701	-0.1023
0.4075	0.3622	-0.3657	-0.2684	-0.0174	0.2711	0.5676
0.2750	0.1667	-0.1303	0.4376	0.3844	-0.3066	0.1230
0.2259	-0.3096	-0.3579	0.3127	-0.2406	-0.3122	-0.2611
0.2958	-0.4232	0.0277	0.4305	-0.3800	0.5114	0.2010

S (7x7) =

3.9901	0	0	0	0	0	0
0	2.2813	0	0	0	0	0
0	0	1.6705	0	0	0	0
0	0	0	1.3522	0	0	0
0	0	0	0	1.1818	0	0
0	0	0	0	0	0.6623	0
0	0	0	0	0	0	0.6487

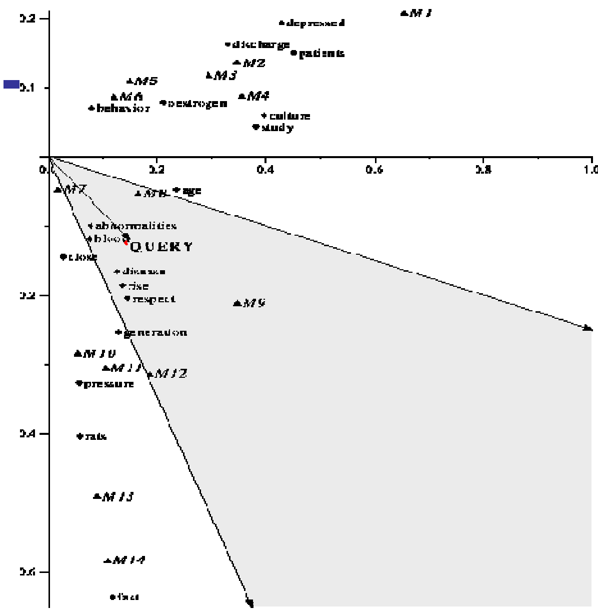
T

V (7x8) =

0.2917	-0.2674	0.3883	-0.5393	0.3926	-0.2112	-0.4505
0.3399	0.4811	0.0649	-0.3760	-0.6959	-0.0421	-0.1462
0.1889	-0.0351	-0.4582	-0.5788	0.2211	0.4247	0.4346
-0.0000	-0.0000	-0.0000	-0.0000	0.0000	-0.0000	0.0000
0.6838	-0.1913	-0.1609	0.2535	0.0050	-0.5229	0.3636
0.4134	0.5716	-0.0566	0.3383	0.4493	0.3198	-0.2839
0.2176	-0.5151	-0.4369	0.1694	-0.2893	0.3161	-0.5330
0.2791	-0.2591	0.6442	0.1593	-0.1648	0.5455	0.2998

This happens to be a rank-7 matrix
-so only 7 dimensions required

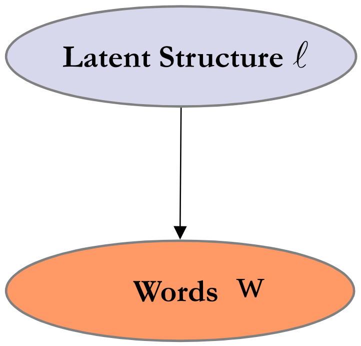
Singular values = Sqrt of Eigen values of AA^T



What LSI can do

- LSI analysis effectively does
  - Dimensionality reduction
  - Noise reduction
  - Exploitation of redundant data
  - Correlation analysis and Query expansion (with related words)
- Some of the individual effects can be achieved with simpler techniques (e.g. thesaurus construction). LSI does them together.
- LSI handles synonymy well, not so much polysemy
- Challenge: SVD is complex to compute ($O(n^3)$)
  - Needs to be updated as a whole as new documents are found/updated, not an online algorithm

Probabilistic Latent Semantic Indexing



Distribution over words

$$P(\mathrm{w}) = \sum_{\ell} P(\mathrm{w}, \ell)$$

Inferring latent structure

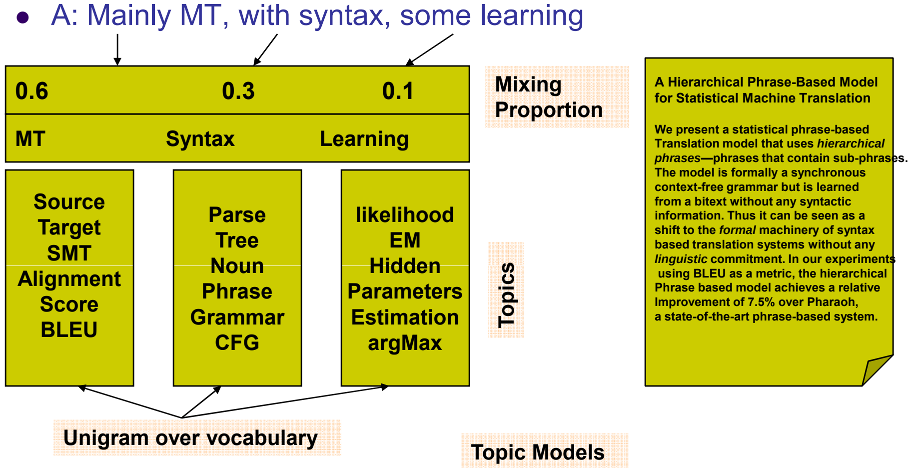
$$P(\ell \mid \mathrm{w}) = \frac{P(\mathrm{w} \mid \ell) P(\ell)}{P(\mathrm{w})}$$

Prediction

$$P(w_{n+1} \mid \mathrm{w}) = \ldots$$

How to Model Semantic?

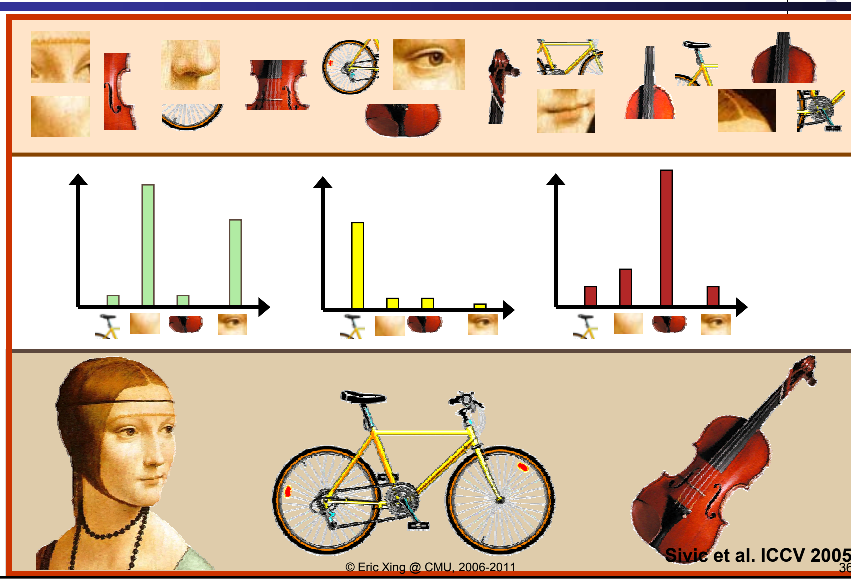
- Q: What is it about?
- A: Mainly MT, with syntax, some learning



Words in Contexts

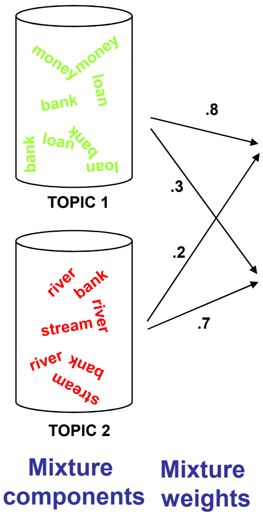
- the opposition Labor **Party** fared even worse, with a predicted 35 **seats**, seven less than last **election**.

"Words" in Contexts (con'd)



Sivic et al. ICCV 2005

GENERATIVE PROCESS



DOCUMENT 1: money^1 bank^1 bank^1 loan^1 river^2 stream^2 bank^1 money^1 river^2 bank^1 money^1 bank^1 loan^1 money^1 stream^2 bank^1 money^1 bank^1 bank^1 loan^1 river^2 stream^2 bank^1 money^1 river^2 bank^1 money^1 bank^1 loan^1 bank^1 money^1 stream^2

DOCUMENT 2: river^2 stream^2 bank^2 stream^2 bank^2 money^1 loan^1 river^2 stream^2 loan^1 bank^2 river^2 bank^2 bank^1 stream^2 river^2 loan^1 bank^2 stream^2 bank^2 money^1 loan^1 river^2 stream^2 bank^2 stream^2 bank^2 money^1 river^2 stream^2 loan^1 bank^2 river^2 bank^2 money^1 bank^1 stream^2 river^2 bank^2 stream^2 bank^2 money^1

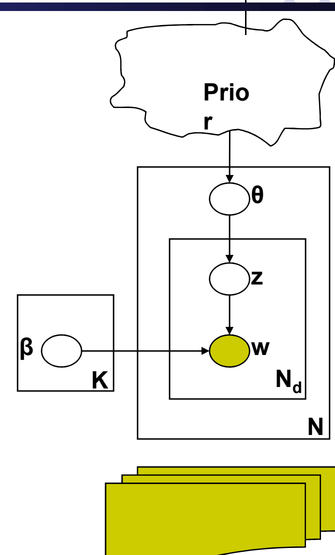
Topic Models = Mixture Membership Models

Generating a document

– *Draw* θ from the prior

For each word n

- Draw z_n from $multinomial(\theta)$
- Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$



Which prior to use?

Latent Dirichlet Allocation

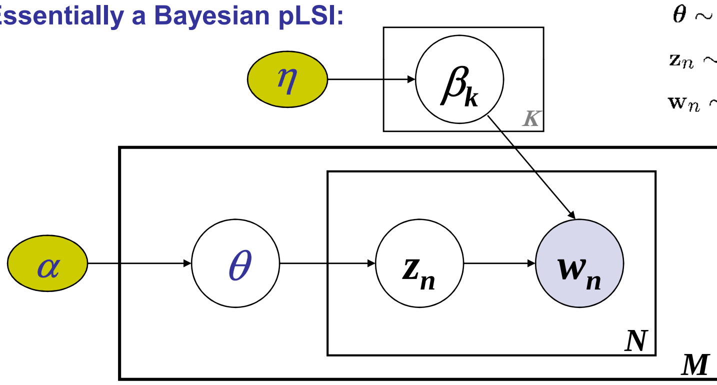
Blei, Ng and Jordan (2003)

Essentially a Bayesian pLSI:

$\theta \sim \mathrm{Dir}(\boldsymbol{\alpha})$

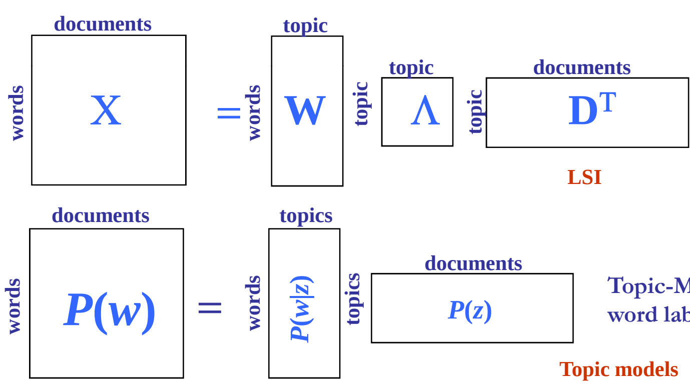
$\mathbf{z}_n \sim \mathrm{Mult}(\boldsymbol{\theta})$

$\mathbf{w}_n \sim p(\mathbf{w}_n | \mathbf{z}_n, \boldsymbol{\beta})$



$$p(\mathbf{w}) = \sum_{\mathbf{z}} \int p(\theta) p(\beta) \left(\prod_{n=1}^{N} p(z_n | \theta) p(w_n | \beta_{z_n}) \right) d\theta \, d\beta$$

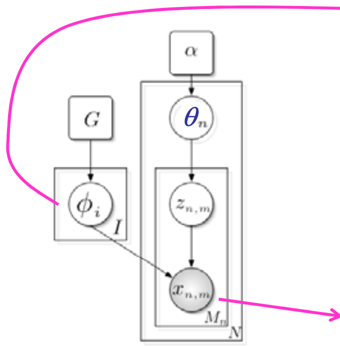
LSI versus Topic Model (probabilistic LSI)



$\vec{x} = W'\vec{d}$

Topic-Mixing is via repeated word labeling

Inference Tasks



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Bayesian inference

- A possible query:

$$p(\theta_n \mid D) = ?$$
$$p(z_{n,m} \mid D) = ?$$

- Close form solution?

$$p(\theta_n \mid D) = \frac{p(\theta_n, D)}{p(D)}$$
$$= \frac{\sum_{\{z_{n,m}\}} \int \left(\prod_n \left(\prod_m p(x_{n,m} \mid \phi_{z_n}) p(z_{n,m} \mid \theta_n) \right) p(\theta_n \mid \alpha) \right) p(\phi \mid G) d\theta_{-i} d\phi}{p(D)}$$

$$p(D) = \sum_{\{z_{n,m}\}} \int \cdots \int \left(\prod_n \left(\prod_m p(x_{n,m} \mid \phi_{z_n}) p(z_{n,m} \mid \theta_n) \right) p(\theta_n \mid \alpha) \right) p(\phi \mid G) d\theta_1 \cdots d\theta_N d\phi$$

- Sum in the denominator over T^n terms, and integrate over n k-dimensional topic vectors

Approximate Inference

- Variational Inference
  - Mean field approximation (Blei et al)
  - Expectation propagation (Minka et al)
  - Variational 2nd-order Taylor approximation (Xing)
- Markov Chain Monte Carlo
  - Gibbs sampling (Griffiths et al)

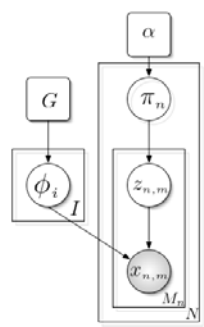
Collapsed Gibbs sampling
(Tom Griffiths & Mark Steyvers)

- Collapsed Gibbs sampling
  - Integrate out π

For variables $\mathbf{z} = z_1, z_2, \ldots, z_n$

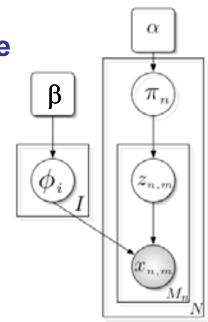
Draw $z_i^{(t+1)}$ from $P(z_i|\mathbf{z}_{-i}, \boldsymbol{w})$

$\mathbf{z}_{-i} = z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_n^{(t)}$



Gibbs sampling

- **Need full conditional distributions for variable**
- **Since we only sample *z* we need**



$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto P(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i})P(z_i = j|\mathbf{z}_{-i})$$

$$= \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

$n_j^{(w)}$ **number of times word *w* assigned to topic *j***

$n_j^{(d)}$ **number of times topic *j* used in document *d***

Gibbs sampling

			iteration 1
i	w_i	d_i	z_i
1	MATHEMATICS	1	2
2	KNOWLEDGE	1	2
3	RESEARCH	1	1
4	WORK	1	2
5	MATHEMATICS	1	1
6	RESEARCH	1	2
7	WORK	1	2
8	SCIENTIFIC	1	1
9	MATHEMATICS	1	2
10	WORK	1	1
11	SCIENTIFIC	2	1
12	KNOWLEDGE	2	1
.	.	.	.
.	.	.	.
.	.	.	.
50	JOY	5	2

Gibbs sampling

			iteration	
			1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

Gibbs sampling

			iteration	
			1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	?
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration	
			1	**2**
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	?
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration	
			1	**2**
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	?
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration	
			1	**2**
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	2
5	MATHEMATICS	1	1	?
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

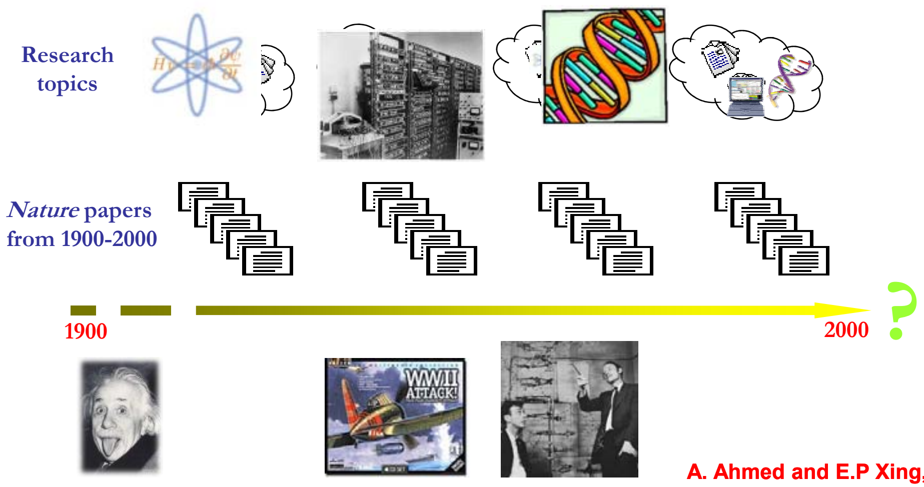
$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling

			iteration			
			1	**2**	**…**	**1000**
i	w_i	d_i	z_i	z_i		z_i
1	MATHEMATICS	1	2	2		2
2	KNOWLEDGE	1	2	1		2
3	RESEARCH	1	1	1		2
4	WORK	1	2	2		1
5	MATHEMATICS	1	1	2		2
6	RESEARCH	1	2	2		2
7	WORK	1	2	2		2
8	SCIENTIFIC	1	1	1	…	1
9	MATHEMATICS	1	2	2		2
10	WORK	1	1	2		2
11	SCIENTIFIC	2	1	1		2
12	KNOWLEDGE	2	1	2		2
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
50	JOY	5	2	1		1

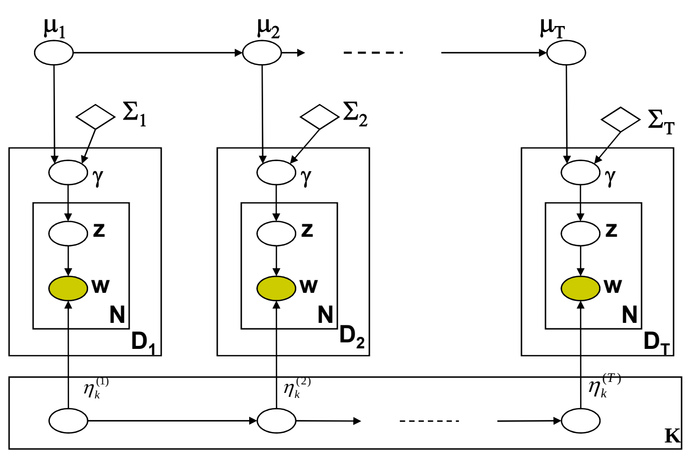
$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Extension 1: topic evolution?

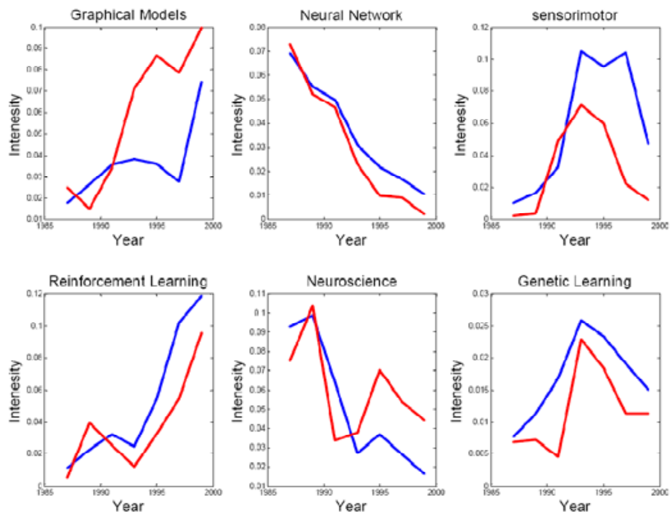


A. Ahmed and E.P Xing, Submitted 2007

The Dynamic CTM



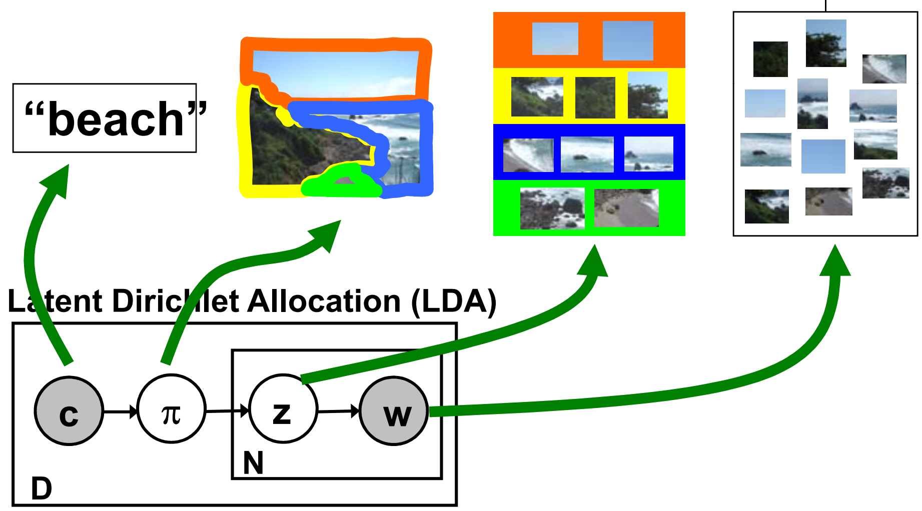
Topic Trends



Topic Words over Time

Leg Insect Controller Sensory sweeping	learning Sweeping state prioritized time control	learning State neural control reinforcement policy	Policy learning State reinforcement Optimal action Reward markov nonlinear	RL
brain information neurons fibers potential model synaptic	model activity network neurons cell phase firing	Model time neurons Spike input Cells head information Rate synaptic temporal	Synergy synaptic Pruning neurons adaptation Spike neuron degradation Irregular hebbian	Neuroscience

Extension 2: Topic Models for Images



Fei-Fei et al. ICCV 2005

Summary:

- Principle of sub-space learning
  - Projection method to reduce the number of dimensions
  - Transfer a set of correlated variables into a new set of possibly uncorrelated variables
  - Map the data into a space of lower dimensionality
  - Form of unsupervised learning
- Properties
  - PCA: It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables; new axes are orthogonal and represent the directions with maximum variability
  - LDA: it can be viewed as a probabilistic generative model where each word in a doc is generated from a doc-specific topic vector defining proportion of memberships from a collection of topic-specific word distributions
- Application: In many settings in pattern recognition and retrieval, we have a feature-object matrix.
  - Dimensionality reduction for each doc/image/user
  - Topic extraction and summarization of a corpus
  - Trend analysis and discovery