School of Computer Science
Carnegie Mellon

Topic Models, Latent Space Models, Sparse Coding, and All That

A systematic understanding of probabilistic semantic extraction in large corpora

Eric Xing
Carnegie Mellon University

Acknowledgement: Amr Ahmed, Qirong Ho, and Jun Zhu

We are inundated with data …

(from images.google.cn)

- Humans cannot afford to deal with (e.g., search, browse, or measure similarity) a huge number of text and media documents
- We need computers to help out …

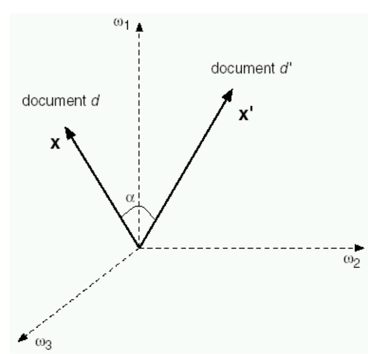
To get started on intelligent systems for automated processing and management of large text or media corpora …

- **Here are some important elements to consider before you start:**
  - Task:
    - Embedding (visualization)? Classification? Clustering? Topic extraction? …
  - Data representation:
    - Input and output (e.g., continuous, binary, counts, …)
  - Model:
    - Latent Semantic Indexing? Bayesian Network? Markov Random Fields? Regression? SVM?
  - Inference:
    - MCMC? Variational? Spectrum Analysis?
  - Learning:
    - MLE? MCLE? Max margin?
  - Computation:
    - Desktop? Hadoop? MPI?
  - Evaluation:
    - Visualization? Human interpretability? Perperlexity? Predictive accuracy?
- **It is better to consider one element at a time!**

Tasks:

- Say, we want to have a mapping …, so that

⇒



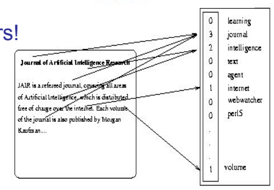
  - Compare similarity
  - Classify contents
  - Cluster/group/categorize docs
  - Distill semantics and perspectives
  - ..

Representation:

- Data: Bag of Words Representation

> As for the Arabian and Palestinean voices that are against the current negotiations and the so-called peace process, they are not against peace per se, but rather for their well-founded predictions that Israel would NOT give an inch of the West bank (and most probably the same for Golan Heights) back to the Arabs. An 18 months of "negotiations" in Madrid, and Washington proved these predictions. Now many will jump on me saying why are you blaming israelis for no-result negotiations. I would say why would the Arabs stall the negotiations, what do they have to loose ?

- Each document is a vector in the word space
- Ignore the order of words in a document. Only count matters!



- A high-dimensional and sparse representation
  - Not efficient text processing tasks, e.g., search, document classification, or similarity measure
  - Not effective for browsing

How to Model Semantic?

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

0.6	0.3	0.1
MT	Syntax	Learning

Mixing Proportion

Source Target SMT Alignment Score BLEU	Parse Tree Noun Phrase Grammar CFG	likelihood EM Hidden Parameters Estimation argMax

Topics

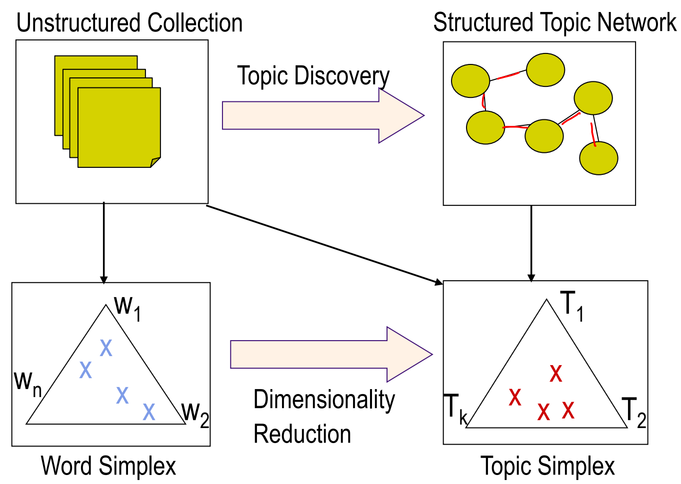
Unigram over vocabulary

Topic Models

A Hierarchical Phrase-Based Model for Statistical Machine Translation

We present a statistical phrase-based Translation model that uses *hierarchical phrases*—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the *formal* machinery of syntax based translation systems without any *linguistic* commitment. In our experiments using BLEU as a metric, the hierarchical Phrase based model achieves a relative Improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.

Topic Model: The Big Picture



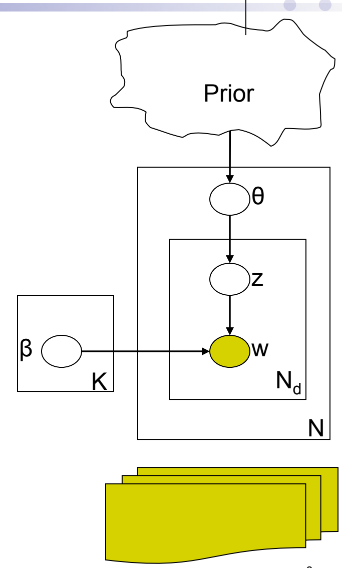
Topic Model as a graphical model

Generating a document

– *Draw* θ from the prior

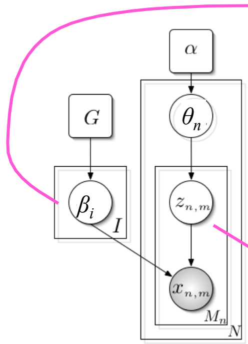
For each word n

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



We can go beyond this by adding more variables and structures to the graph!

Learning and Inference



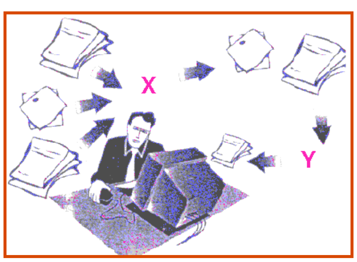
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	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.

Advanced issues:

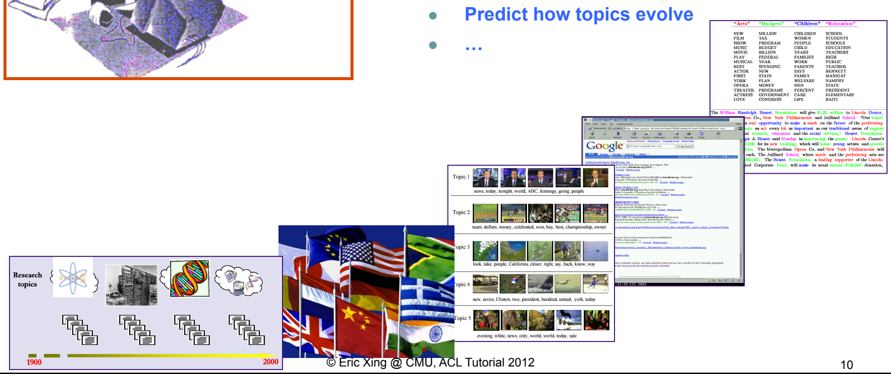
- Objective function: likelihood? Margin? RSS? …
- Data: iid docs, streaming text, multimodal media …
- Algorithm: direct optimization, Monte Carlo, variational methods …
- System: single machine, multi-core machine, distributed system …

Deliverables:



We want:

- Topics and categorization of documents
- Semantic-based ranking of docs
- Multimedia inference
- Automatic translation
- Predict how topics evolve
- …



Questions:

- What is the mathematical and computational basis of all these?
- How to do it right, modular, fast, and real time?
- How to build other related applications on topic models?
- How to scale up?

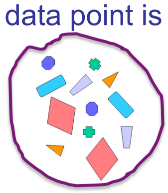
Plan of this tutorial

- 1. Overview of basic topic models
- 2. Computational challenges and two classical algorithmic paths
- 3. Scenario I: Multimodal data
- 4. Scenario II: When supervision is available
- 5. Scenario III: What if I don't know the total number of topics
- 6. Scenario IV: Topic evolution in streaming corpus.
- 7: Advanced subject I: Sparsity in topic modeling (see EMNLP talk)
- 8: Advanced subject II: Scalability, complexity, and fast algorithms (optional)
- 9: Other applications (optional)

1. Overview of topic models

Understanding document corpora

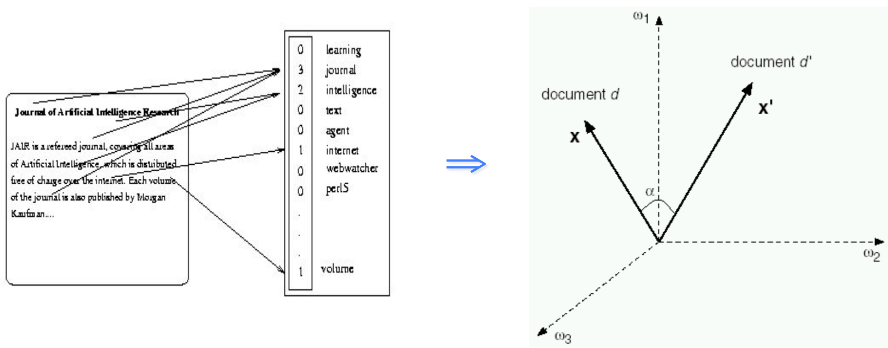
- A document collection is a dataset where each data point is itself a collection of simpler data.



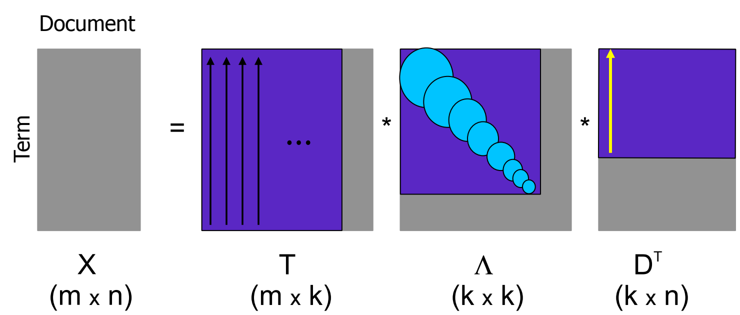
  - Text documents are collections of words.
  - Segmented images are collections of regions.
  - User histories are collections of purchased items.
- Many modern problems ask questions on such data.
  - What topics do these documents “span”?
  - Is this text document relevant to my query?
  - Which category is this text/image in?
  - How have topics changed over time?
  - Who wrote this specific document?
  - What will author X write about?
  - and so on…..

The Vector Space Model

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



Latent Semantic Indexing

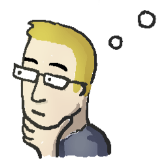


$$\vec{w} = \sum_{k=1}^{K} d_k \lambda_k \vec{T}_k$$

- LSI does not define a properly normalized probability distribution of observed and latent entities
  - Does not support probabilistic reasoning under uncertainty and data fusion

How our brain might work …

Apoptosis + Medicine



How our brain might work …

Apoptosis + Medicine

probabilistic generative model

How our brain might work …

Apoptosis + Medicine

statistical
inference

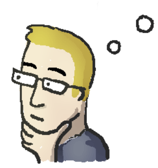
What is Learning

Learning is about seeking a predictive and/or executable understanding of natural/artificial subjects, phenomena, or activities from …

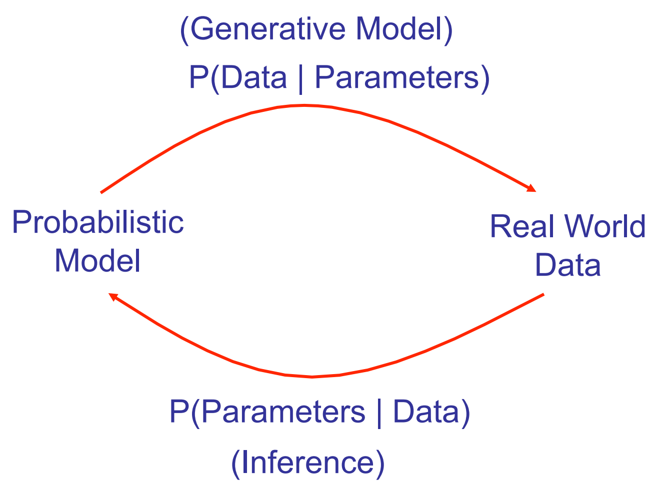
Apoptosis + Medicine

Grammatical rules
Manufacturing procedures
Natural laws
…

Inference



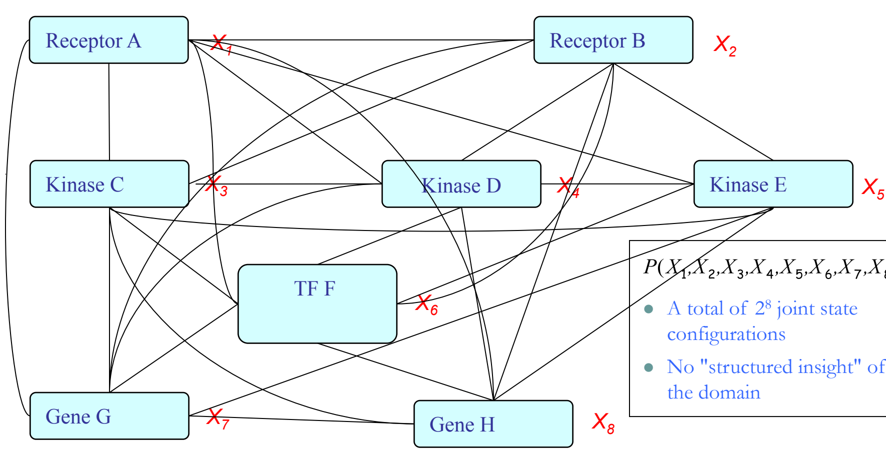
Connecting Probability Models to Data



What is a Graphical Model?

--- example from a signal transduction pathway

- A possible world for cellular signal transduction:



$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$

- A total of 2^8 joint state configurations
- No "structured insight" of the domain

Recap of Basic Prob. Concepts

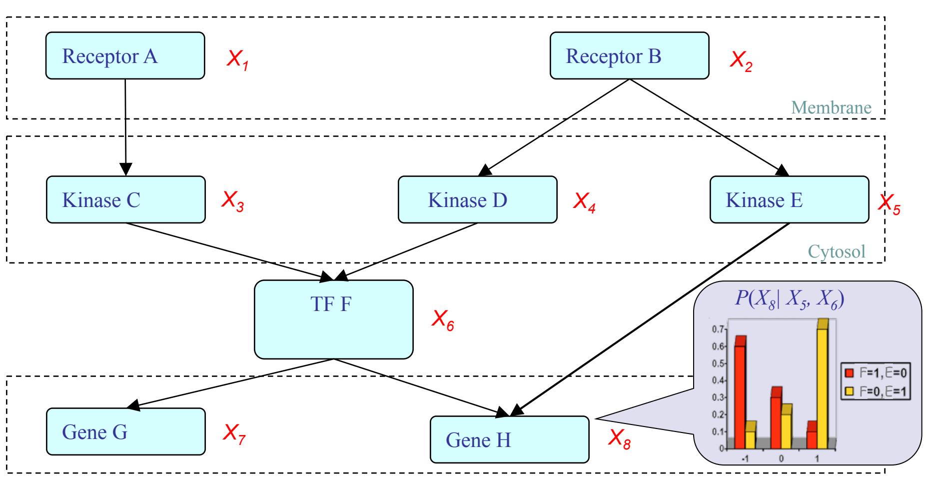
- Representation: what is the joint probability dist. on multiple variables?

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8,)$$

  - How many state configurations in total? --- 2^8
  - Are they all needed to be represented?
  - Do we get any scientific/medical insight?
- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how much data do we need?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?
- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?

GM: Structure Simplifies Representation

- Dependencies among variables



Probabilistic Graphical Models

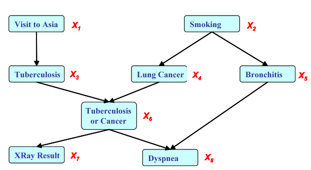
- Represent dependency structure with a graph
  - Node <-> random variable
  - Edges encode dependencies
    - Absence of edge -> conditional independence
  - Directed and undirected versions

- Why is this useful?
  - A language for communication
  - A language for computation
  - A language for development

- Origins:
  - Wright 1920's
  - Independently developed by Spiegelhalter and Lauritzen in statistics and Pearl in computer science in the late 1980's

Probabilistic Graphical Models, con'd

- If X_i's are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

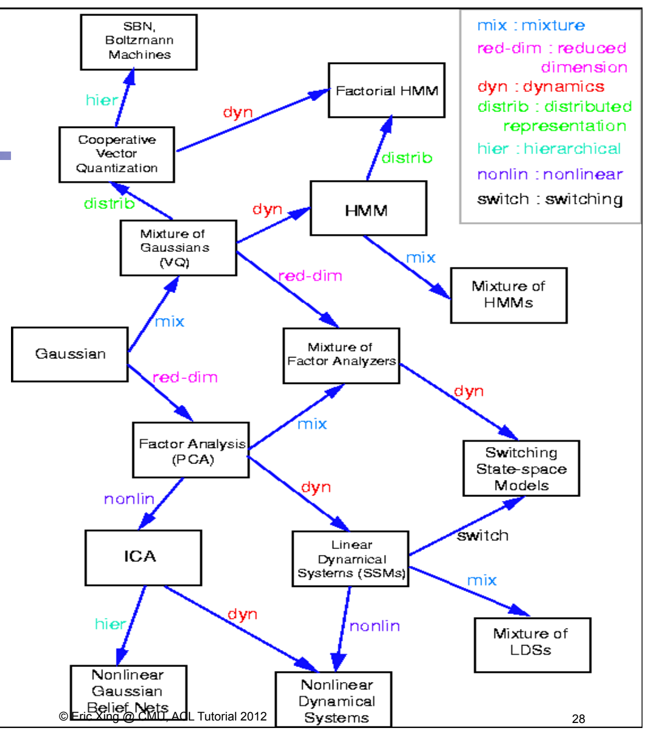
- Why we may favor a PGM?
  - Representation cost: how many probability statements are needed?

    2+2+4+4+4+8+4+8=36, an 8-fold reduction from 2^8!
  - Algorithms for systematic and efficient inference/learning computation
    - Exploring the graph structure and probabilistic (e.g., Bayesian, Markovian) semantics
  - Incorporation of domain knowledge and causal (logical) structures

Probabilistic Inference

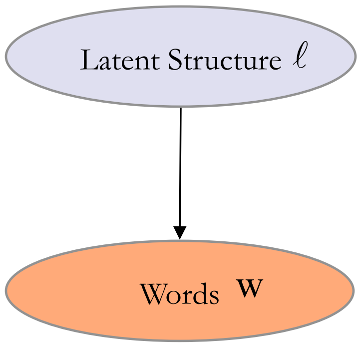
- Computing statistical queries regarding the network, e.g.:
  - Is node X independent on node Y given nodes Z,W ?
  - What is the probability of X=true if (Y=false and Z=true)?
  - What is the joint distribution of (X,Y) if Z=false?
  - What is the likelihood of some full assignment?
  - What is the most likely assignment of values to all or a subset the nodes of the network?
- General purpose algorithms exist to fully automate such computation
  - Computational cost depends on the topology of the network
  - Exact inference:
    - The junction tree algorithm
  - Approximate inference;
    - Loopy belief propagation, variational inference, Monte Carlo sampling

An (incomplete) genealogy of graphical models



(Picture by Zoubin Ghahramani and Sam Roweis)

Latent Semantic Structure in GM



Distribution over words

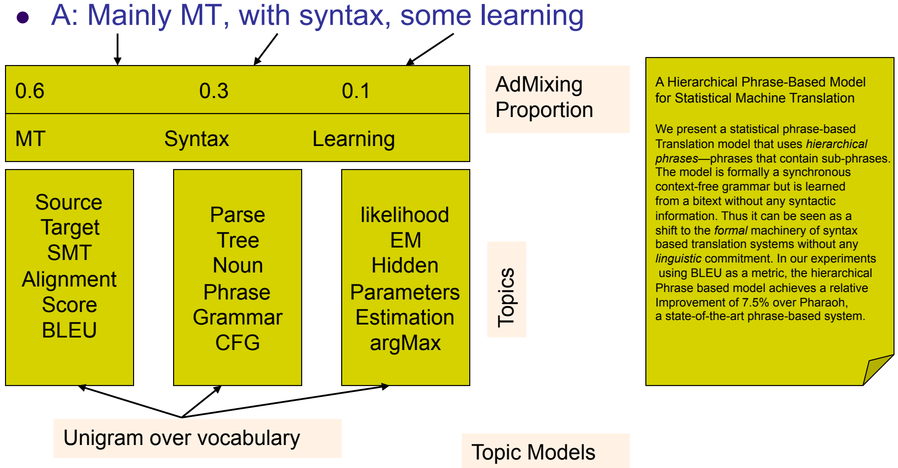
$$P(\mathbf{w}) = \sum_{\ell} P(\mathbf{w}, \ell)$$

Inferring latent structure

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

How to Model Semantics?

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



Why this is Useful?

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

0.6	0.3	0.1
MT	Syntax	Learning

AdMixing Proportion

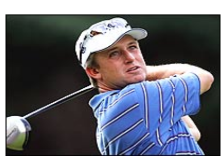
A Hierarchical Phrase-Based Model for Statistical Machine Translation

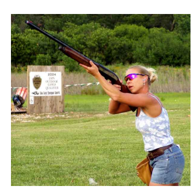
We present a statistical phrase-based Translation model that uses *hierarchical phrases*—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the *formal* machinery of syntax based translation systems without any *linguistic* commitment. In our experiments using BLEU as a metric, the hierarchical Phrase based model achieves a relative Improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.

- Q: give me similar document?
  - Structured way of browsing the collection
- Other tasks
  - Dimensionality reduction
    - TF-IDF vs. topic mixing proportion
    - Classification, clustering, and more …

Words in Contexts

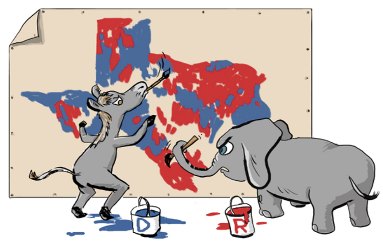
- "It was a nice **shot**."



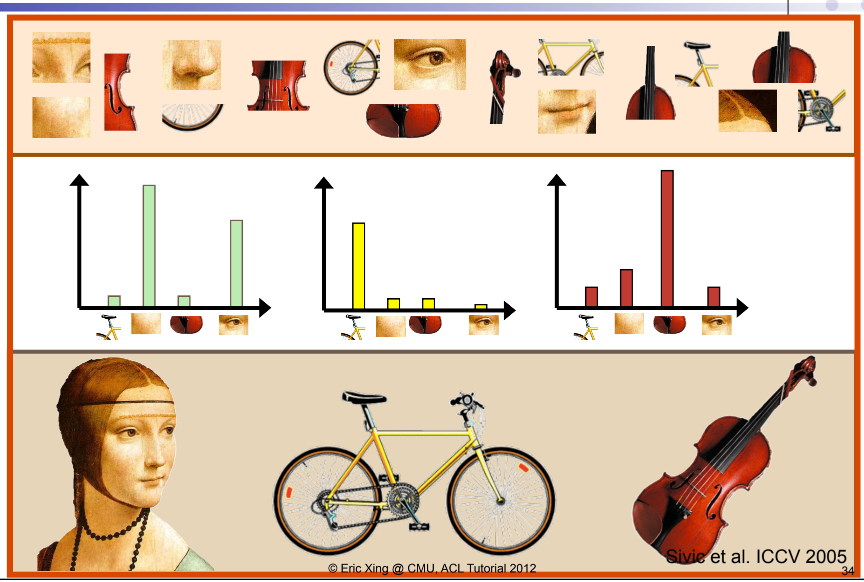


Words in Contexts (con'd)

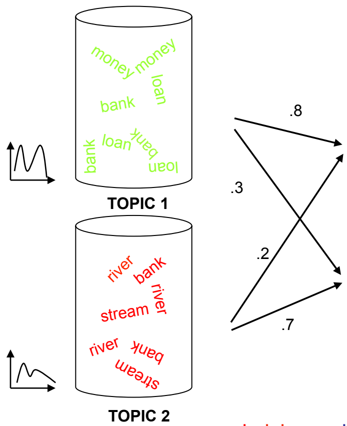
- the opposition Labor **Party** fared even worse, with a predicted 35 **seats**, seven less than last **election**.



"Words" in Contexts (con'd)



A possible generative process of a document



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

Mixture Components (distributions over elements)

admixing weight vector θ (represents all components' contributions)

Bayesian approach: use priors
Admixture weights $\sim$ Dirichlet(α)
Mixture components $\sim$ Dirichlet(Γ)

Method One:

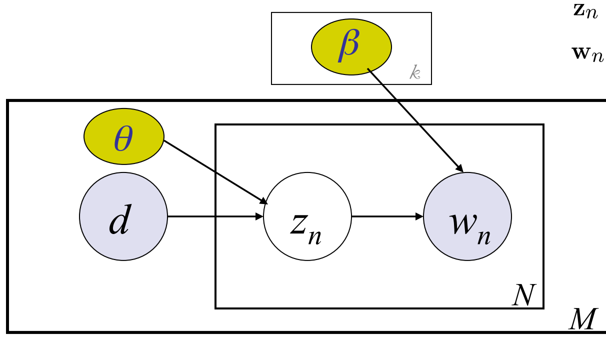
- **Hierarchical Bayesian Admixture (a.k.a. probabilistic Topic Models)**

Probabilistic LSI

Hoffman (1999)

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

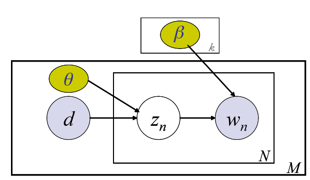
$\mathbf{w}_n \sim p(\mathbf{w}_n|\mathbf{z}_n, \boldsymbol{\beta})$



$$p(d, w_n) = p(d)\sum_{\mathbf{z}}\left(\prod_{n=1}^{N} p(w_n \mid z_n)\, p(z_n \mid d)\right)$$

Probabilistic LSI

- A "generative" model
- Models each word in a document as a sample from a mixture model.
- Each word is generated from a single topic, different words in the document may be generated from different topics.
- A topic is characterized by a distribution over words.
- Each document is represented as a list of admixing proportions for the components (i.e. topic vector θ).



Latent Dirichlet Allocation

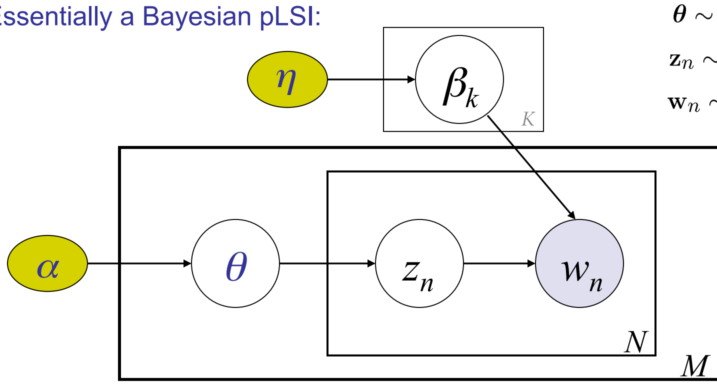
Blei, Ng and Jordan (2003)

Essentially a Bayesian pLSI:

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

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

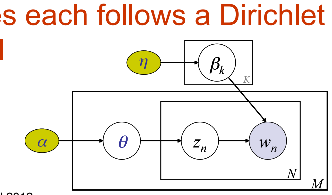
$\mathbf{w}_n \sim p(\mathbf{w}_n | \mathbf{z}_n, \boldsymbol{\beta})$



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

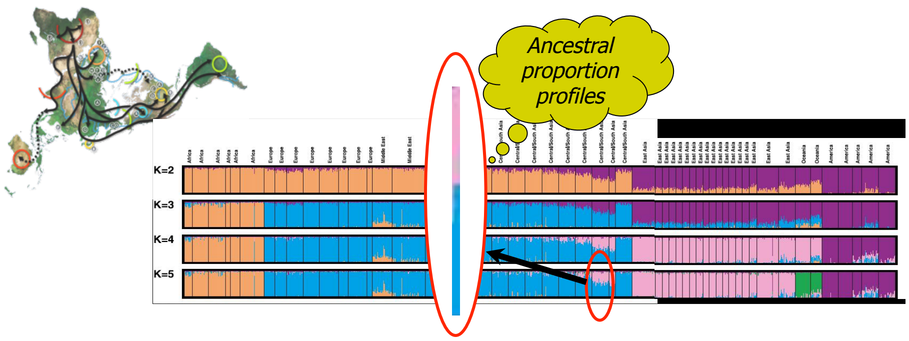
LDA

- Generative model
- Models each word in a document as a sample from a mixture model.
- Each word is generated from a single topic, different words in the document may be generated from different topics.
- A topic is characterized by a distribution over words.
- Each document is represented as a list of admixing proportions for the components (i.e. topic vector).
- The topic vectors and the word rates each follows a Dirichlet prior --- essentially a Bayesian pLSI



LDA was first invented by geneticist …

- How to present population structure?
  - *Structure*



Inference of population structure using multilocus genotype data. J.K. Pritchard, M. Stephens and P. J. Donnelly, 2000. Genetics 155: 945-959.

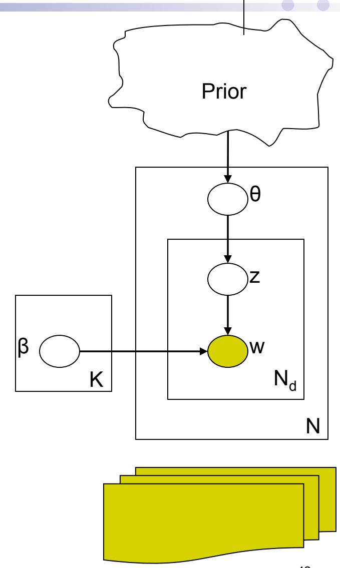
Genetic structure of Human Populations (Rosenberg et al. Science, 2002)

Topic Models = Mixed Membership Models = Admixture

Generating a document

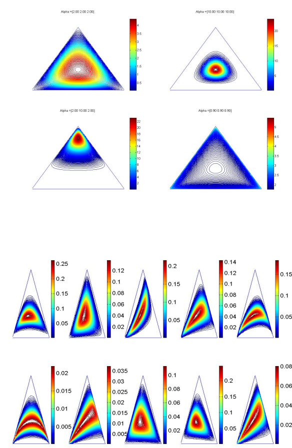
– *Draw* θ from the prior
For each word n
- Draw z_n from $multinomial(\theta)$
- Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$

Which prior to use?



Choices of Priors

- Dirichlet (LDA) (Blei et al. 2003)
  - Conjugate prior means efficient inference
  - Can only capture variations in each topic's intensity independently
- Logistic Normal (CTM=LoNTAM) (Blei & Lafferty 2005, Ahmed & Xing 2006)
  - Capture the intuition that some topics are highly correlated and can rise up in intensity together
  - Not a conjugate prior implies hard inference
- Nested CRP (Blei et al 2005)
  - Defines hierarchy on topics
  - …



Generative Semantic of LoNTAM

Generating a document

– *Draw* θ from the prior

For each word n

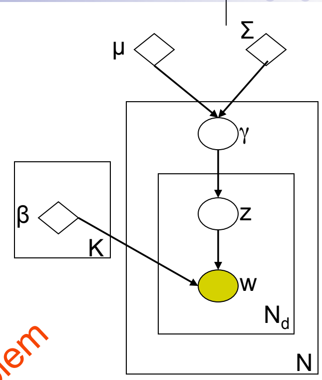
- Draw z_n from $multinomial(\theta)$
- Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$

$$\theta \sim LN_K(\mu, \Sigma)$$

$$\gamma \sim N_{K-1}(\mu, \Sigma) \qquad \gamma_K = 0$$

$$\theta_i = \exp\left\{\gamma_i - \log\left(1 + \sum_{i=1}^{K-1} e^{\gamma_i}\right)\right\}$$

$$C(\gamma) = \log\left(1 + \sum_{i=1}^{K-1} e^{\gamma_i}\right)$$



Problem

- Log Partition Function
- Normalization Constant

Outcomes from a topic model

- The "topics" β in a corpus:

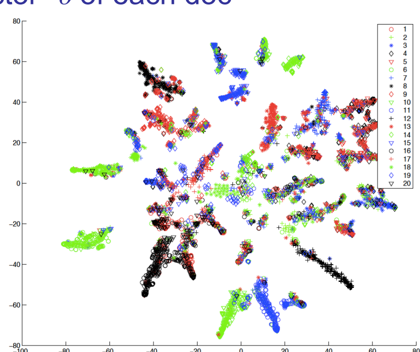
	T 59	T 104	T 31
comp.graphics	image	ftp	card
	jpeg	pub	monitor
	color	graphics	dos
	file	mail	video
	gif	version	apple
	images	tar	windows
	format	file	drivers
	bit	information	vga
	files	send	cards
	display	server	graphics
	T 30	**T 84**	**T 44**
sci.electronics	power	water	sale
	ground	energy	price
	wire	air	offer
	circuit	nuclear	shipping
	supply	loop	sell
	voltage	hot	interested
	current	cold	mail
	wiring	cooling	condition
	signal	heat	email
	cable	temperature	cd

	T 42	T 78	T 47
politics.mideast	israel	jews	armenian
	israeli	jewish	turkish
	peace	israel	armenians
	writes	israeli	armenia
	article	arab	turks
	arab	people	genocide
	war	arabs	russian
	lebanese	center	soviet
	lebanon	jew	people
	people	nazi	muslim
	T 44	**T 94**	**T 49**
misc.forsale	sale	don	drive
	price	mail	scsi
	offer	call	disk
	shipping	package	hard
	sell	writes	mb
	interested	send	drives
	mail	number	ide
	condition	ve	controller
	email	hotel	floppy
	cd	credit	system

  - There is no name for each "topic", you need to name it!
  - There is no objective measure of good/bad
  - The shown topics are the "good" ones, there are many many trivial ones, meaningless ones, redundant ones, … you need to manually prune the results
  - How many topics? …

Outcomes from a topic model

- The "topic vector" θ of each doc

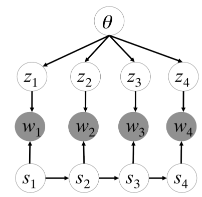


  - Create an embedding of docs in a "topic space"
  - Their no ground truth of θ to measure quality of inference
  - But on θ it is possible to define an "objective" measure of goodness, such as classification error, retrieval of similar docs, clustering, etc., of documents
  - But there is no consensus on whether these tasks bear the true value of topic models …

Outcomes from a topic model

- The per-word topic indicator z:

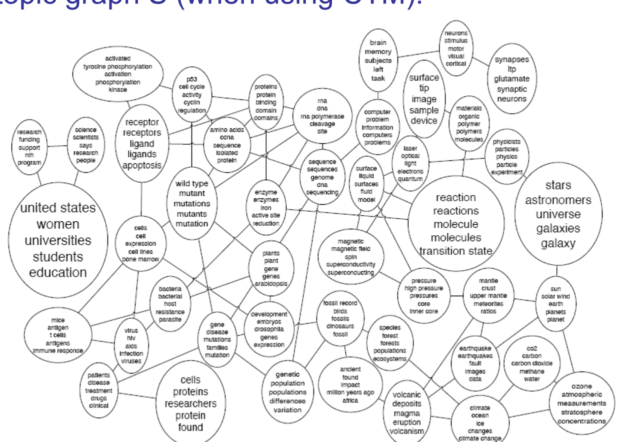
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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.



- Not very useful under the bag of word representation, because of loss of ordering
- But it is possible to define simple probabilistic linguistic constraints (e.g, bi-grams) over z and get potentially interesting results [Griffiths, Steyvers, Blei, & Tenenbaum, 2004]

Outcomes from a topic model

- The topic graph S (when using CTM):

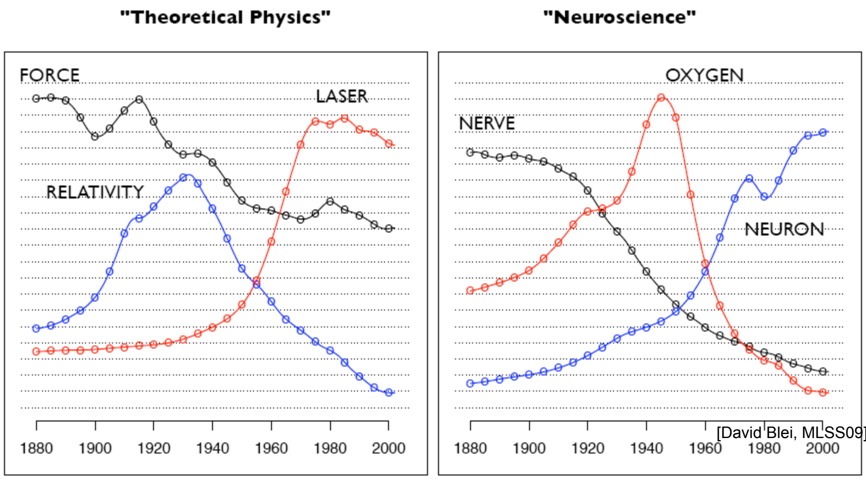


[David Blei, MLSS09]

- Kind of interesting for understanding/visualizing large corpora

Outcomes from a topic model

- Topic change trends



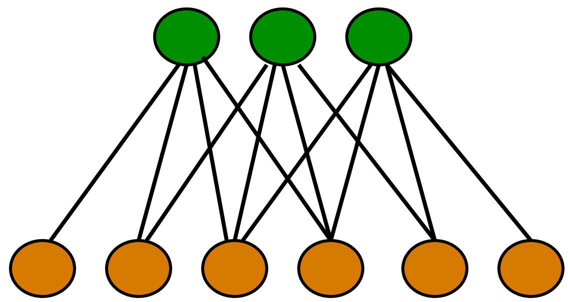
Method Two:

- **Layered Boltzmann machines (an undirected Topic Model)**

The Harmonium

hidden units

visible units



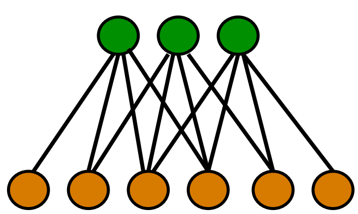
Boltzmann machines:

$$p(x,h \mid \theta) = \exp\left\{ \sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{i,j} \theta_{i,j} \phi_{i,j}(x_i, h_j) - A(\theta) \right\}$$

A Binomial Word-count Model

E.P. Xing, R. Yan and A. G. Hauptmann, UAI 2006

topics



words counts

$h_j = 3$: topic j has strength 3

$h_j \in \mathbf{R}, \qquad \langle h_j \rangle = \sum_i W_{i,j} x_i$

$x_i = \mathrm{n}$: word i has count n

$x_i \in \mathbf{I}$

$$p(\mathbf{h} \mid \mathbf{x}) = \prod_j \mathrm{Normal}_{h_j}\left[\sum_i \vec{W}_{ij} \vec{x}_i, 1 \right]$$

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_i \mathrm{Bi}_{x_i}\left[N, \frac{\exp(\alpha_j + \sum_j W_{ij} h_j)}{1+\exp(\alpha_j + \sum_j W_{ij} h_j)} \right]$$

$$\Rightarrow \quad p(\mathbf{x}) \propto \exp\left\{ \left(\sum_i \alpha_i x_i - \log\Gamma(x_i) - \log\Gamma(N - x_i) \right) + \tfrac{1}{2} \sum_j \left(\sum_i W_{i,j} x_i \right)^2 \right\}$$

The Computational Trade-off

Undirected model: Learning is hard, inference is easy.

Directed Model: Learning is "easier", inference is hard.

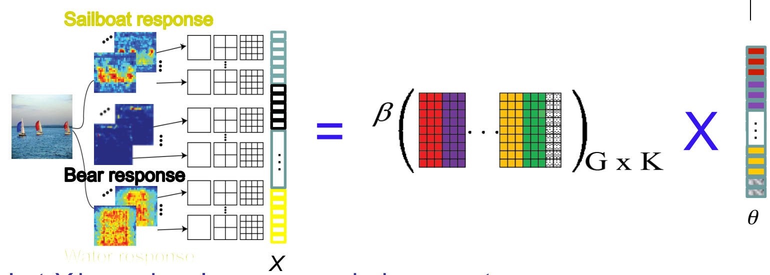
Example: Document Retrieval.

Retrieval is based on comparing (posterior) topic distributions of documents.
- directed models: inference is slow. Learning is relatively "easy".
- undirected model: inference is fast. Learning is slow but can be done offline.

Method Three:

- **Sparse topic coding (a non-probabilistic Topic Model)**
  - **And in this category recently there is also nonnegative matrix factorization (NMF)**

Sparse Coding



- Let X be a signal, e.g., speech, image, etc.
- Let β be a set of normalized "basis vectors"
  - We call it dictionary
- β is "adapted" to x if it can represent it with a few basis vectors
  - There exists a sparse vector θ such that $x \approx \beta\, \theta$
  - We call θ the sparse code

Primer on Sparse Coding

- Sparse Coding with appropriate constraints:

reconstruction loss sparsity-inducing regularizer

$$\min_{\boldsymbol{\theta},\boldsymbol{\beta}} \quad \sum_d \ell(\theta_d, \boldsymbol{\beta}|\mathbf{x}_d) + \lambda\Psi(\boldsymbol{\theta})$$

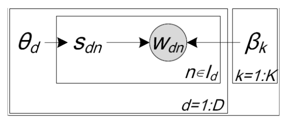
$$\text{s.t.}: \quad \boldsymbol{\beta} \in \Omega_1; \boldsymbol{\theta} \in \Omega_2.$$

- Reconstruction loss can be:
  - the general log-likelihood loss of an exponential family distribution (Lee et al., 2010)
- Sparisty-inducing regularizer can be:
  - the L_0 "pseudo-norm": $\|\theta\|_0 := \sum_i \delta(\theta_i, 0)$ NP-hard
  - the L_1 norm: $\|\theta\|_1 := \sum_i |\theta_i|$ Convex
  - Structured regularizers, e.g., group Lasso (Bengio et al., 2009) $\|\theta\|_{1/2} := \sum_g \|\theta_{\mathcal{I}_g}\|_2$
- Suggests an alternating optimization procedure

Sparse Topical Coding

- Goal: design a non-probabilistic topic model that is amenable to
  - direct control on the posterior sparsity of inferred representations
  - avoid dealing with normalization constant when considering supervision or rich features
  - seamless integration with a convex loss function (e.g., svm hinge loss)

- We extend sparse coding to hierarchical sparse topical coding
  - word code θ
  - document code $\mathbf{s}$



reconstruction loss

sparse codes

$$\min_{\{\boldsymbol{\theta}_d,\mathbf{s}_d\},\boldsymbol{\beta}} \sum_{d,n\in I_d} \ell(w_{dn}, \mathbf{s}_{dn}^{\top}\boldsymbol{\beta}_{.n}) + \lambda \sum_d \|\boldsymbol{\theta}_d\|_1 + \sum_{d,n\in I_d} (\gamma\|\mathbf{s}_{dn} - \boldsymbol{\theta}_d\|_2^2 + \rho\|\mathbf{s}_{dn}\|_1)$$

$$\text{s.t.}: \ \boldsymbol{\theta}_d \geq 0, \ \mathbf{s}_{dn} \geq 0, \ \forall d, n \in I_d; \ \ \boldsymbol{\beta}_k \in \mathcal{P}, \ \forall k,$$

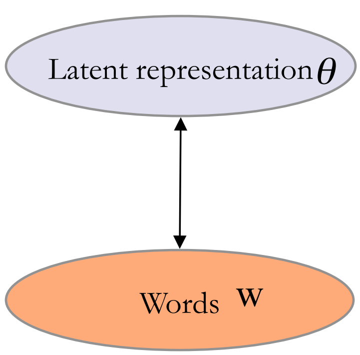
truncated aggregation

non-negative codes

topical bases

J. Zhu, & E.P. Xing. UAI, 2011

Summary: Latent Sub-space Models



The Model:

$$P(\mathbf{w}, \theta; \beta)$$

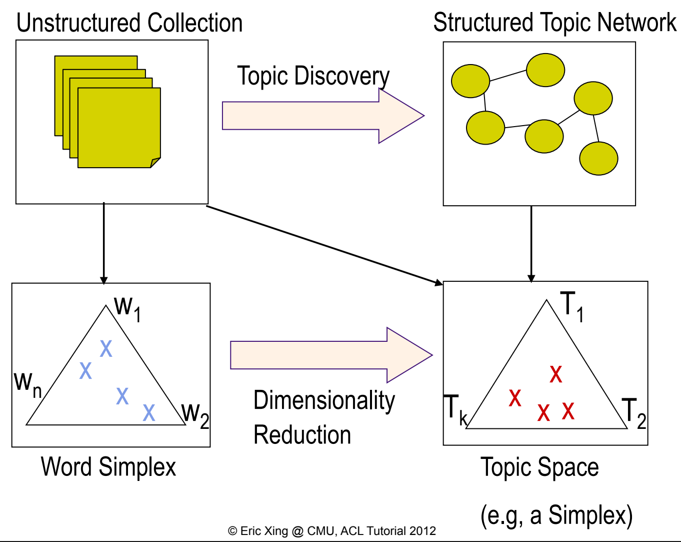
Inferring latent representation:

$$P(\theta \mid \mathbf{w}) = \frac{P(\mathbf{w}, \theta)}{P(\mathbf{w})}$$

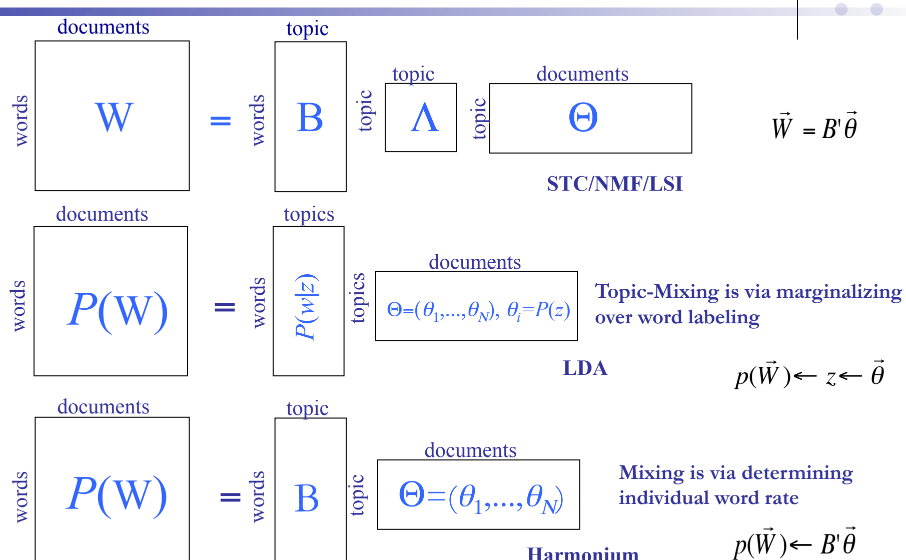
Learning the subspace:

$$\beta = \arg\min f_{\beta}(w, \theta)$$

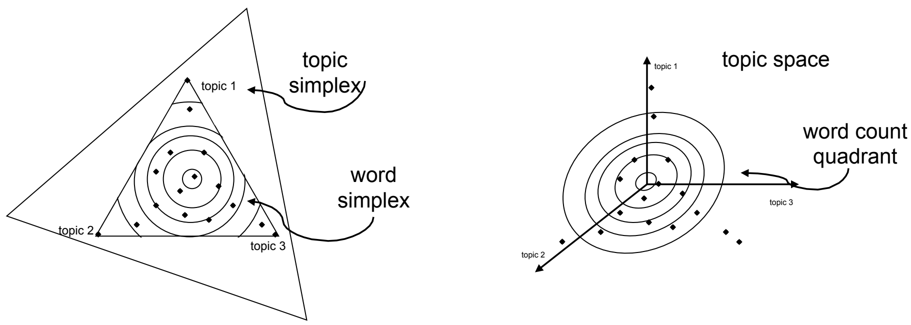
The Big Picture



Comparison of model semantics



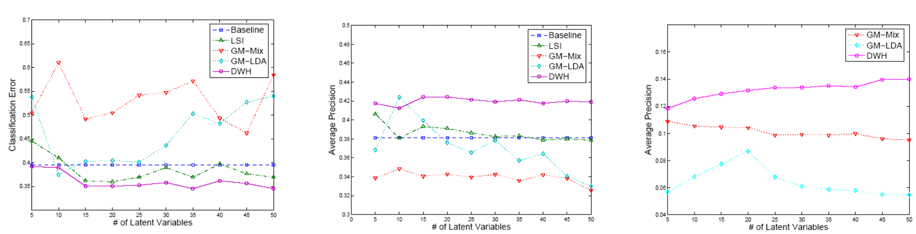
Comparison of topic space



Comparisons

LDA vs. Harmonium

[Xing, Yan, and Hauptman, UAI 2005]

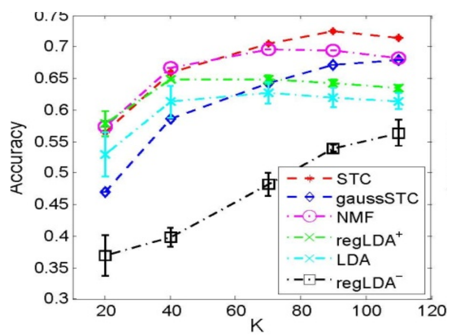


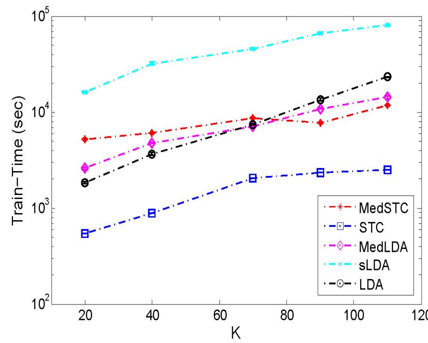
Classification Retrieval Annotation

- LDA is actually doing very poor on several "objectively" evaluatable predictive tasks

Comparisons

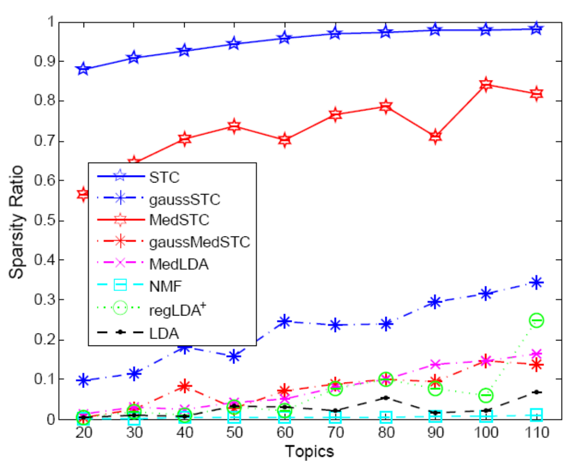
LDA vs. STC
[Zhu and Xing, UAI 2011]





Sparse word codes

- Sparsity ratio: percentage of zeros

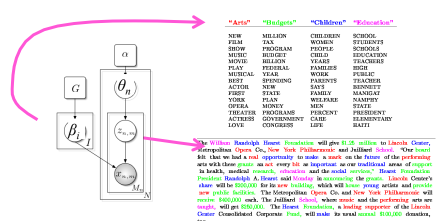


- NMF: non-negative matrix factorization
- MedLDA (Zhu et al., 2009)
- regLDA: LDA with entropic regularizer
- gaussSTC: use L2 rather than L1-norm

2. Computational challenges and three algorithmic paths

Computation on LDA

- Inference
  - Given a Document D
    - Posterior: P(Θ | μ,Σ, β ,D)
    - Evaluation: P(D| μ,Σ, β)



- Learning
  - Given a collection of documents {D_i}
    - Parameter estimation

$$\arg\max_{(\mu,\Sigma,\beta)} \sum \log\left(P\left(D_i \middle| \mu,\Sigma,\beta\right)\right)$$

Exact Bayesian inference on LDA is intractable

- A possible query:

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

  - Close form solution?

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

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

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

Approximate Inference

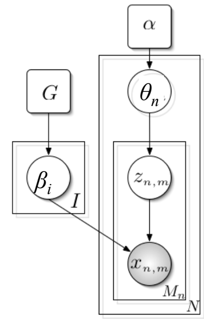
- Variational Inference
  - Mean field approximation (Blei et al)
  - Expectation propagation (Minka et al)
  - Variational 2nd-order Taylor approximation (Ahmed and Xing)

- Markov Chain Monte Carlo
  - Gibbs sampling (Griffiths et al)

Collapsed Gibbs sampling

(Tom Griffiths & Mark Steyvers)

- Collapsed Gibbs sampling
  - Integrate out θ



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

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

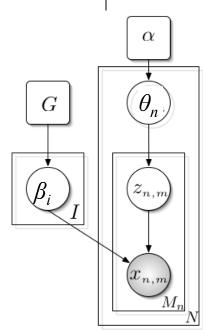
$\mathbf{z}_{-i} = z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_n^{(t)}$

$$\{z^{(1)}, z^{(2)}, \ldots, z^{(T)}\}$$

$$\theta = \frac{1}{T} \sum_t z^{(t)}$$

Gibbs sampling

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



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

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

$n_j^{(w)}$ 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	

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

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	

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

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	

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

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	

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

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	

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

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	

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

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

$$\theta = \frac{1}{T}\sum_t z^{(t)}$$

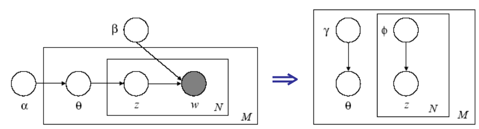
$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + G}{n_{-i,j}^{(\cdot)} + WG} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Variational Inference

(e.g., MF, Jordan et al 1999, GMF, Xing et al 2004)

- Variational approximation

$$q(\theta, z) = q_\theta(\theta) q_z(z)$$
$$= \mathrm{Dir}(\theta \mid \gamma = f(\alpha, \langle z \rangle)) \times \mathrm{Multi}(z \mid \phi = f(\beta_w, \langle \ln\theta \rangle))$$



$$\phi_{ni} \propto \beta_{iw_n} \exp\{E_q[\log(\theta_i) \mid \gamma]\}$$
$$\gamma_i = \alpha_i + \textstyle\sum_{n=1}^N \phi_{ni}.$$

- Data set:
  - 15,000 documents
  - 90,000 terms
  - 2.1 million words
- Model:
  - 100 factors
  - 9 million parameters
- On a single machine MCMC could converge too slowly for this problem, but …

Learning a TM

- Maximum likelihood estimation:

$$\{\beta_1, \beta_2, \ldots, \beta_K\}, \alpha = \arg\max_{(\alpha,\beta)} \sum \log\left(P\left(D_i \middle| \alpha, \beta\right)\right)$$

- Need statistics on topic-specific word assignment (due to z), topic vector distribution (due to θ), etc.
  - E.g,, this is the formula for topic k:

$$\beta_k = \frac{1}{\sum_d N_d} \sum_{d=1}^{D} \sum_{d_n=1}^{N_d} \delta(z_{d,d_n}, k) w_{d,d_n}$$

- These are hidden variables, therefore need an EM algorithm (also known as data augmentation, or DA, in Monte Carlo paradigm)
- This is a “reduce” step in parallel implementation

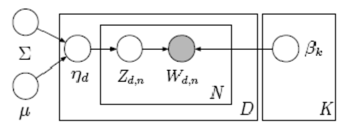
How to evaluate inference/ learning algorithm?

- Empirical performance on, say, clustering, classification, topic saliency, perplexity … ?
- There is no ground truth, poor/good performance may come from model, data, algorithm, parameter tuning …

- In simulation you know the ground, thus you can exclusively compare the difference caused by inference/learning algorithm!

Case study: Correlated Topic Model

- $\mathit{Draw}\ \eta \sim N_{K-1}(\mu, \Sigma)$
- $\theta = \exp\{\eta - c(\eta)\}$
- For each word n in document
  - Draw z_n from $\mathit{multinomial}(\theta)$
  - Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $\mathit{multinomial}(\beta_{z_n})$



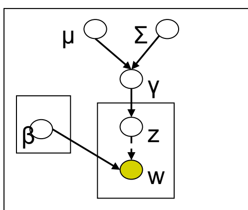
$$C(\eta) = \log\left(1 + \sum_{i=1}^{K-1} e^{\eta_i}\right)$$

Two approaches to approximate it:

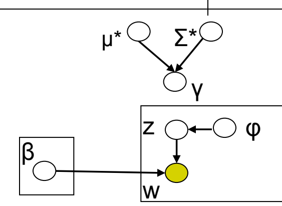
- Blei and Lafferty use tangent
- (Xing 2005) uses second order truncated Taylor approximation

Non-conjugacy comes here

Variational Inference of CTM



$P(\gamma, \{z\} | D)$



$q(\gamma, z_{1:n}) = q(\gamma | \mu^*, \Sigma^*) \prod q(z_n | \phi_n)$

Log Partition Function

$$\log\left(1 + \sum_{i=1}^{K-1} e^{\gamma_i}\right)$$

Σ* is full matrix

Multivariate Quadratic Approx.

Closed Form Solution for μ*, Σ*

Ahmed&Xing 05

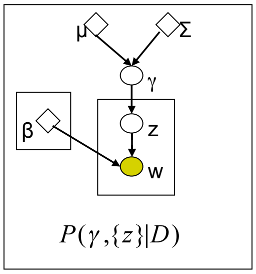
Σ* is assumed to be diagonal

Tangent Approx.

Numerical Optimization to fit μ*, Diag(Σ*)

Blei&Lafferty 05

Variational Inference With no Tears



Iterate until Convergence

- Pretend you know $E[Z_{1:n}]$
  - $P(\gamma|E[z_{1:n}], \mu, \Sigma)$
- Now you know $E[\gamma]$
  - $P(z_{1:n}|E[\gamma], w_{1:n}, \beta_{1:k})$

- More Formally:

$$q^*(X_C) = P\left(X_C \middle| \langle S_Y \rangle_{q_y} : \forall y \in X_{MB} \right)$$

Message Passing Scheme (GMF)

Equivalent to previous method (Xing et. al.2003)

LoNTAM Variations Inference

- Fully Factored Distribution

$$q(\gamma, z_{1:n}) = q(\gamma) \prod q(z_n)$$

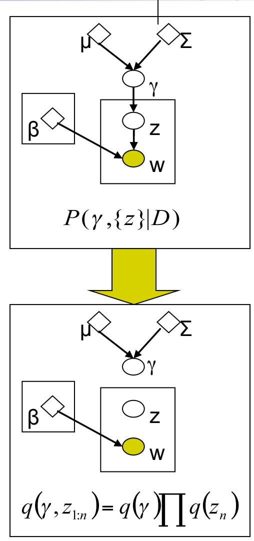
- Two clusters: λ and $Z_{1:n}$

$$q^*(X_C) = P\left(X_C \middle| \langle S_Y \rangle_{q_y} : \forall y \in X_{MB} \right)$$

- Fixed Point Equations

$$q_\gamma^*(\gamma) = P\left(\gamma \middle| \langle S_z \rangle_{q_z}, \mu, \Sigma \right)$$

$$q_z^*(z) = P\left(z \middle| \langle S_\gamma \rangle_{q\gamma}, \beta_{1:k} \right)$$



Variational γ

$$q_\lambda^*(\gamma) = P\left(\gamma \middle| \langle S_z\rangle_{q_z}, \mu, \Sigma\right)$$

$$\propto P(\gamma,\mu,\Sigma)\, P\left(\langle S_z\rangle_{q_z} \middle| \gamma\right)$$

Now what is $\langle S_z\rangle_{q_z}$?

$$S_z = m = \left[\sum_n I(z_n = 1), \ldots, \sum_n I(z_n = k)\right]$$

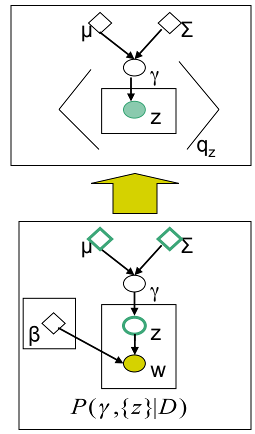
$$\propto N(\gamma,\mu,\Sigma)\exp\left\{\langle m\rangle_{q_z}\gamma - N \times C(\gamma)\right\}$$

$$\propto \exp\left\{\underbrace{-\frac{1}{2}\gamma'\Sigma^{-1}\gamma + \gamma\Sigma^{-1}\mu + \langle m\rangle_{q_z}\gamma} - N \times \underbrace{\boxed{C(\gamma)}}\right\}$$

$$C(\gamma) = C(\gamma_\wedge) + g'_\lambda(\gamma - \gamma_\wedge) + .5\,(\lambda - \gamma_\wedge)\,H\,(\gamma - \gamma_\wedge)$$

$$q_\lambda^*(\gamma) = N(\mu_\gamma, \Sigma_\gamma) \qquad \Sigma_\gamma = inv\left(\Sigma^{-1} + NH\right)$$

$$\mu_\gamma = \Sigma_\gamma\left(\Sigma^{-1}\mu + NH\gamma_\wedge + \langle m\rangle - Ng\right)$$



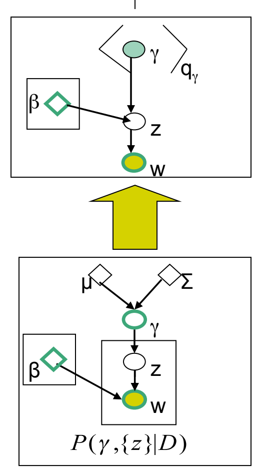
Variational Z

$$q_z^*(z) = P\left(z \middle| \langle S_\gamma\rangle_{q\gamma}, \beta, w\right)$$

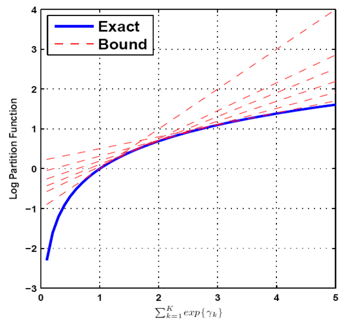
$$\propto P\left(z^k \middle| \langle S_\gamma\rangle_{q\gamma}\right) P\left(w^j \middle| z^k, \beta\right)$$

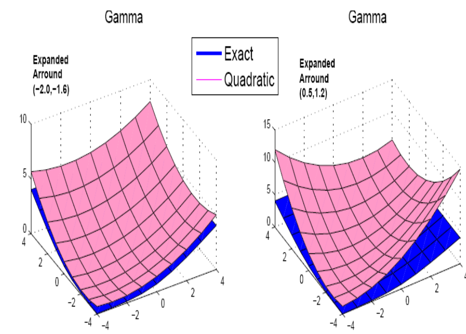
$$\propto P\left(z^k \middle| \langle\gamma\rangle_{q\gamma}\right)\beta_{kj}$$

$$\propto \exp\left\{u_{\gamma,k}\right\}\beta_{kj}$$

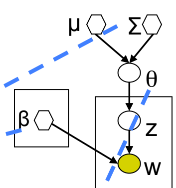


Tangent Approximation

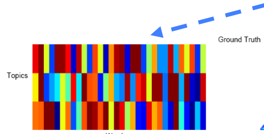


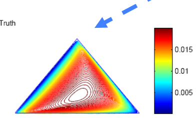


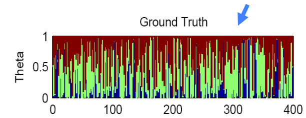
Different Learning/Inference deliver different performance



Test on Synthetic Text (of “known” ground truth):



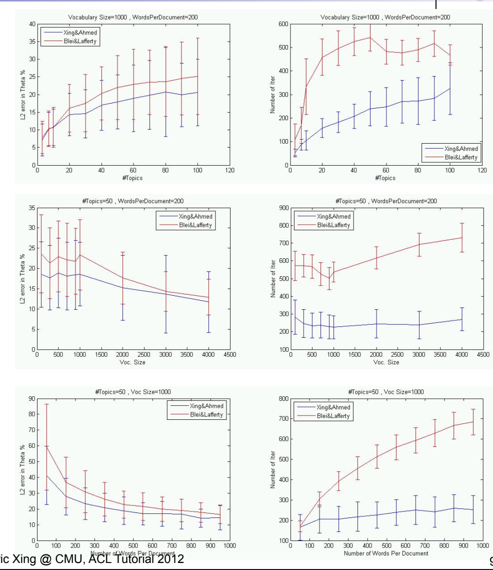




Comparison: accuracy and speed

L2 error in topic vector est. and # of iterations

- Varying Num. of Topics
- Varying Voc. Size
- Varying Num. Words Per Document

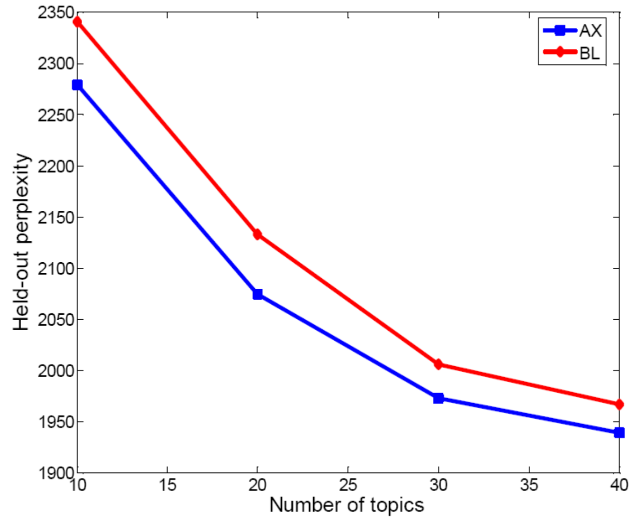


Result on NIPS collection

- NIPS proceeding from 1988-2003
- 14036 words
- 2484 docs
- 80% for training and 20% for testing
- Fit both models with 10,20,30,40 topics
- Compare **perplexity** on held out data
    - The perplexity of a language model with respect to text x is the reciprocal of the geometric average of the probabilities of the predictions in text x. So, if text ***x*** has ***k*** words, then the perplexity of the language model with respect to that text is

$$\Pr(x)^{-1/k}$$

Comparison: perplexity



Classification Result on PNAS collection

- PNAS abstracts from 1997-2002
  - 2500 documents
  - Average of 170 words per document
- Fitted 40-topics model using both approaches
- Use low dimensional representation to predict the abstract category
  - Use SVM classifier
  - 85% for training and 15% for testing

Classification Accuracy

Category	Doc	BL	AX
Genetics	21	61.9	61.9
Biochemistry	86	65.1	77.9
Immunology	24	70.8	66.6
Biophysics	15	53.3	66.6
Total	146	64.3	72.6

-Notable Difference
-Examine the low dimensional representations below

Computation on undirected TM

[Welling et al NIPS 04, Xing et al, UAI 05]

Undirected model: Learning is hard, inference is easy.

Directed Model: Learning is "easier", inference is hard.

Example: Document Retrieval.

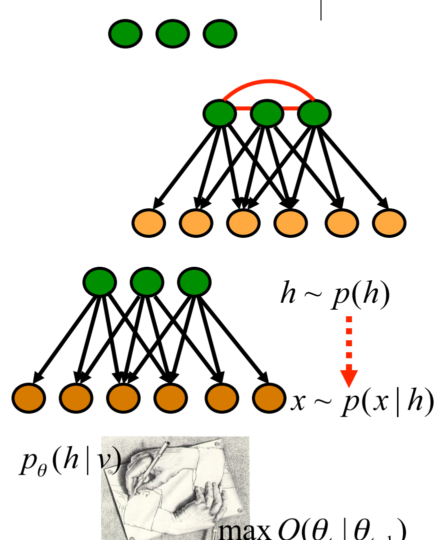
Retrieval is based on comparing (posterior) topic distributions of documents.
- directed models: inference is slow. Learning is relatively "easy".
- undirected model: inference is fast. Learning is slow but can be done offline.

Properties of Directed Networks

- Factors are marginally *independent.*
- Factors are conditionally *dependent* given observations on the visible nodes.

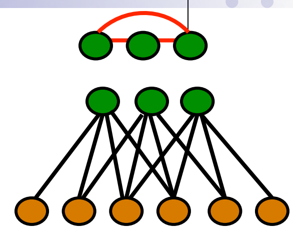
$$P(\ell \mid \mathbf{w}) = \frac{P(\mathbf{w} \mid \ell) P(\ell)}{P(\mathbf{w})}$$

- Easy ancestral sampling.
- Learning with (variational) EM



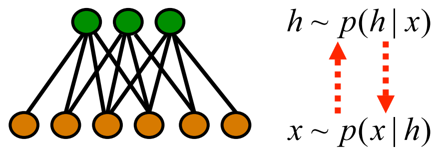
Properties of Harmoniums

- Factors are marginally *dependent*.
- Factors are conditionally *independent* given observations on the visible nodes.

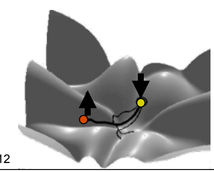


$$P(\ell \mid \mathbf{w}) = \prod_i P(\ell_i \mid \mathbf{w})$$

- Iterative Gibbs sampling.



- Learning with contrastive divergence



Learning and Inference

- Maximal likelihood learning based on gradient ascent.

$$\delta\theta_i \propto \langle f_i(x_i) \rangle_{\text{data}} - \langle f_i(x_i) \rangle_p$$

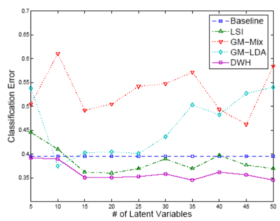
  - gradient computation requires model distribution $p(.)$
  - $p(.)$ is intractable

- Contrastive Divergence
  - approximate $p(.)$ with Gibbs sampling

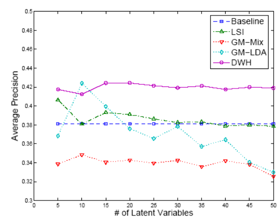
- Variational approximation
  - GMF approximation

$$q(\mathbf{x},\mathbf{z},\mathbf{h}) = \prod_i q(x_i \mid \nu_i) \prod_k q(z_k \mid \mu_k, \sigma_k) \prod_j q(h_j \mid \gamma_j)$$

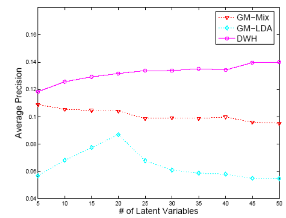
Performance



Classification



Retrieval



Annotation

Computation on STC

[Zhu and Xing, UAI 11]

- Hierarchical sparse coding
  - for each document

$$\min_{\theta,\mathbf{s}} \quad \sum_{n\in I} \ell(w_n, \mathbf{s}_n^\top \beta_n) + \lambda\|\theta\|_1 + \sum_{n\in I}(\gamma\|\mathbf{s}_n - \theta\|_2^2 + \rho\|\mathbf{s}_n\|_1))$$

$$\text{s.t.}: \quad \theta \geq 0; \quad \mathbf{s}_n \geq 0, \ \forall n \in I,$$

  - Word code

$$s_{nk} = \max(0, \nu_k)$$

$$\text{where } 2\gamma\beta_{kn}\nu_k^2 + (2\gamma\mu + \beta_{kn}\eta)\nu_k + \mu\eta - w_n\beta_{kn} = 0$$

  - Document code (truncated averaging)

$$\theta_k = \max(0, \bar{s}_k - \frac{\lambda}{2\gamma|I|}) \text{ where } \bar{s}_k = \frac{1}{|I|}\sum_{n\in I} s_{nk}$$

- Dictionary learning
  - projected gradient descent
  - any faster alternative method can be used

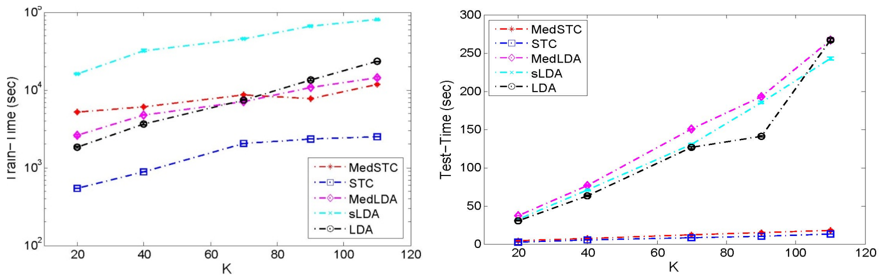
Opt. Algorithm for Sparse Coding

- Much research has been done for optimizing a convex, but non-smooth objective (may subject to some constraints, e.g., non-negativity)
- Greedy algorithm for the non-convex L_0 "pseudo-norm":
  - select the element with maximum correlation with the residual
  - known as "matching pursuit" (Mallat & Zhang, 1993)
- For the convex L_1 norm, many algorithms:
  - Soft-thresholding with coordinate descent (Friedman et al., 2007; Fu, 1998; Zhu & Xing, 2011)
  - Proximal methods (Nesterov, 2007; Jenatton et al., 2010)
  - Active-set methods (Roth & Fischer, 2008)
  - Iterative Re-weighted Least Squares (Daubechies et al., 2008)
  - LARS (Efron et al., 2004) solves for regularization path
  - Online/stochastic variants
  - …

Opt. Algorithm for Dictionary Learning

- Optimize a convex and usually smooth objective w/o (convex) constraints
- General optimization procedure can be applied, less research has been done for this step
  - Projected gradient descent
  - Block-wise coordinate descent
  - …
- A recent progress is made on online/stochastic optimization method (Mairal et al., 2010)

Performance

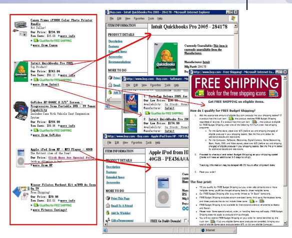


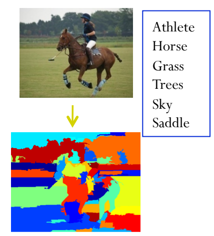
~ 10 times speed up in train &test

3. Scenario I: Multimodal data

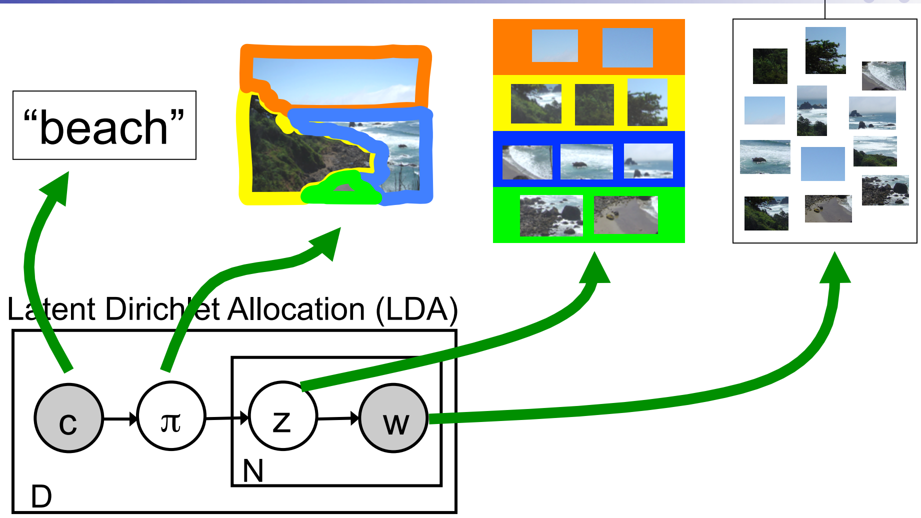
Multi-view analysis

- Two-view webpage documents (Blum & Mitchell, 1998)
  - Contents
  - In-link anchor texts
- Multi-view Images
  - Local features, e.g., color histogram, SIFT bag-of-word, sparse SIFT codes, et al.
  - Global features, e.g., gist (Oliva & Torralba, 2001), et al.
  - Online tags
- Social media and social networks
  - …
- **Problems with a flat model, e.g., SVM**
  - Type difference (from different distributions)
  - Scale, length difference, etc.
  - Incapable of doing view-level prediction
- **Multi-view data analysis via latent sub-space discovery**
  - Provide a good joint distribution
  - Provide good conditional distributions of the description type conditioned on the primary type



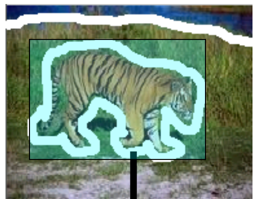


Latent Space Models for Images



Fei-Fei et al. ICCV 2005

Image representation



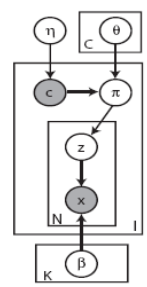
cat, grass, tiger, water

$[r_{11} \cdots r_{1d}]$, $[w_1 \cdots w_{|V|}]$

representation vector (real, 1 per image segment) **:** **annotation vector** (binary, same for each segment)

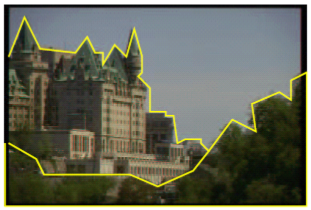
$$[r_{n1} \cdots r_{nd}] \ , \ [w_1 \cdots w_{|V|}]$$

To Generate an Image …



$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, c|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\beta}) &= p(c|\boldsymbol{\eta})p(\boldsymbol{\pi}|c, \boldsymbol{\theta}) \cdot \\
& \prod_{n=1}^{N} p(z_n|\boldsymbol{\pi})p(x_n|z_n, \boldsymbol{\beta}) \\
p(c|\boldsymbol{\eta}) &= \mathrm{Mult}(c|\boldsymbol{\eta}) \\
p(\boldsymbol{\pi}|c, \boldsymbol{\theta}) &= \prod_{j=1}^{C} Dir(\boldsymbol{\pi}|\boldsymbol{\theta}_{j \cdot})^{\delta(c,j)} \\
p(z_n|\boldsymbol{\pi}) &= \mathrm{Mult}(z_n|\boldsymbol{\pi}) \\
p(x_n|z_n, \boldsymbol{\beta}) &= \prod_{k=1}^{K} p(x_n|\boldsymbol{\beta}_{k \cdot})^{\delta(z_n^k, 1)}
\end{aligned}
$$

Annotated images



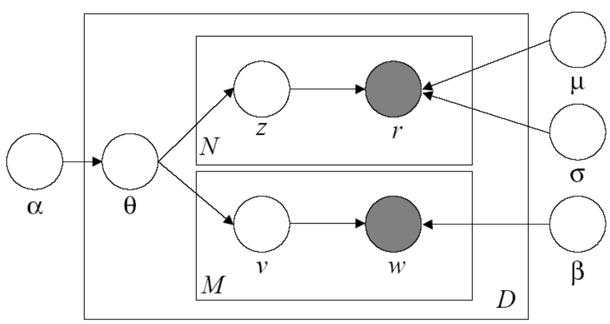
{9.32, 2.44, 0.02, 3.23}
{4.35, 3.12, −0.23, 9.41}
{6.65, 2.11, 1.02, 2.31}

This cozy place is nestled in the heart of the Mission. Easy access to bars, restuarants, and BART.

This, cozy, place, is, nestled, in, the, heart, of, the, Mission, Easy, access, to, bars, restuarants, and, BART

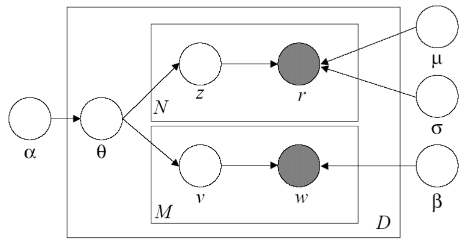
- Forsyth et. al. (2001): images as documents where region-specific feature vectors are like visual words.
- A captioned image can be thought of as annotated data: two documents, one of which describes the other.

Gaussian-multinomial LDA [Bliei et al, JMLR 05]



- A natural next step is to glue two LDA models together.
- Bottom: a traditional LDA model on captions
- Top: a Gaussian-LDA model on images
  - each region is a multivariate Gaussian
- Does not work well

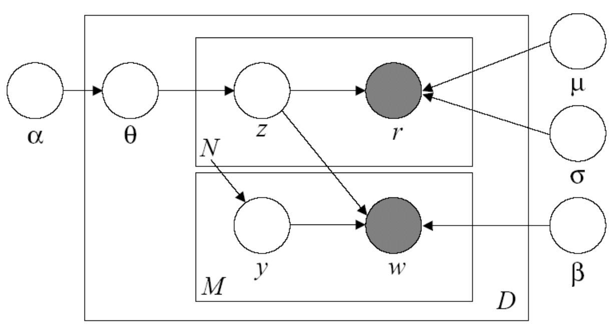
Exchangeability



- Like LDA, GM-LDA implicitly makes an *exchangeability* assumption about words and regions, and their corresponding topics.
- The order in which words and regions are generated does not matter.
- But this is goes against the way we're thinking about the data!
- The words are chosen to describe the image.
- The implicit exchangeability assumptions in the model should reflect this. In other words, we want to model *partial exchangeability*

Corr-LDA

[Bliei et al, JMLR 05]



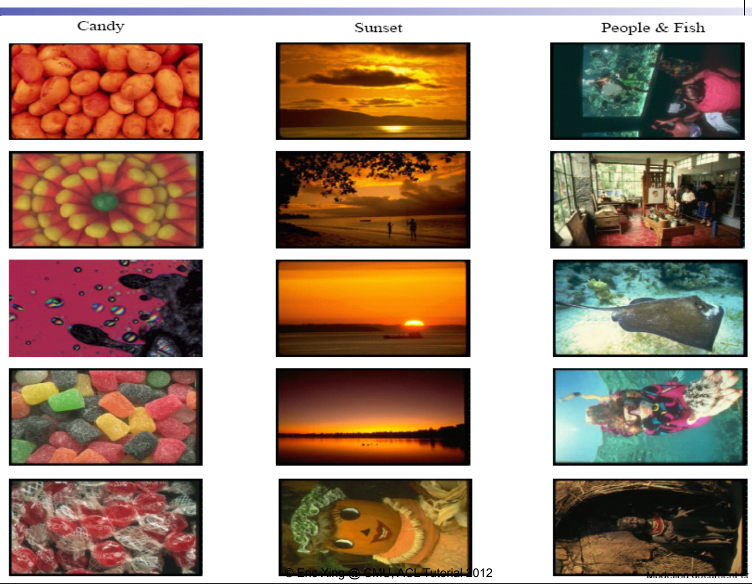
- Since, **w** is conditioned on **z**, the image must be generated first.
- Unlike GM-LDA, the caption is guaranteed to be generated by a subset of the same hidden factors which generated the image.
- The model enforces a correspondence between the latent space associated with images and the latent space associated with captions.

Automatic annotation

True caption
birds tree
Corr−LDA
birds nest leaves branch tree
GM−LDA
water birds nest tree sky
GM−Mixture
tree ocean fungus mushrooms coral

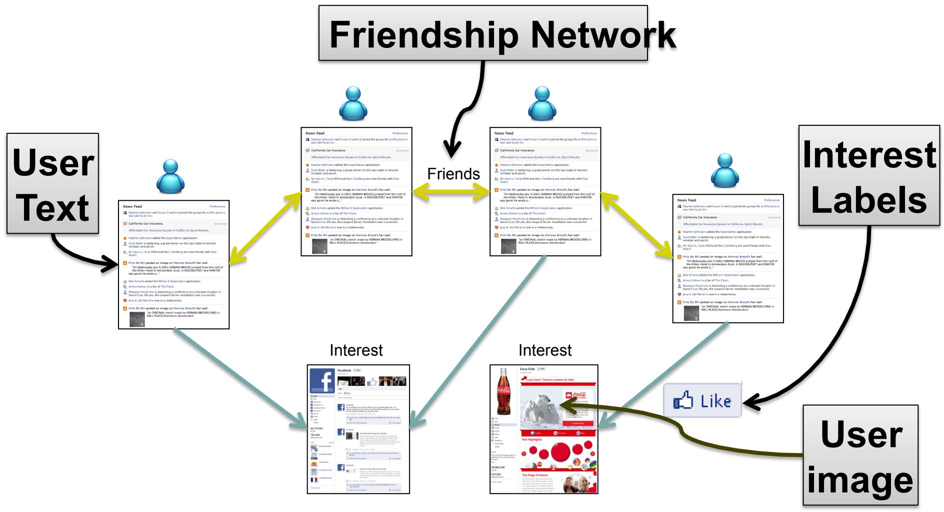
True caption
fish reefs water
Corr−LDA
fish water ocean tree coral
GM−LDA
water sky vegetables tree people
GM−Mixture
fungus mushrooms tree flowers leaves

Text-based image retrieval



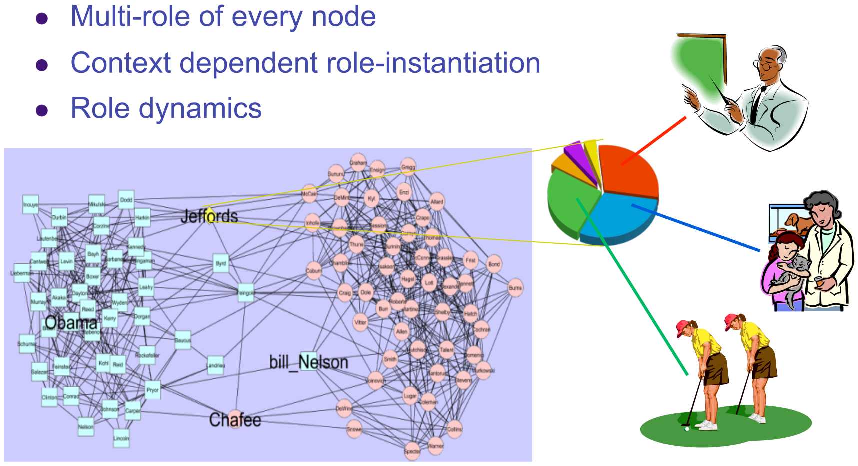
© Eric Xing @ CMU, ACL Tutorial 2012

Multi-view social media data

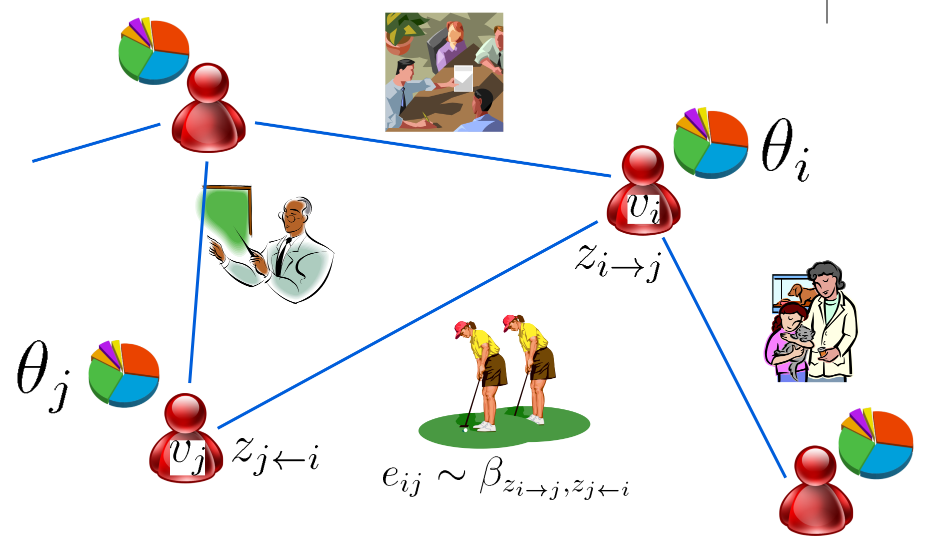


Latent space models for network

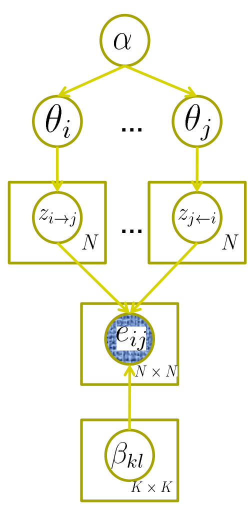
- Micro-inference vs. Meso- or Macro-inference
- Multi-role of every node
- Context dependent role-instantiation
- Role dynamics



Example:

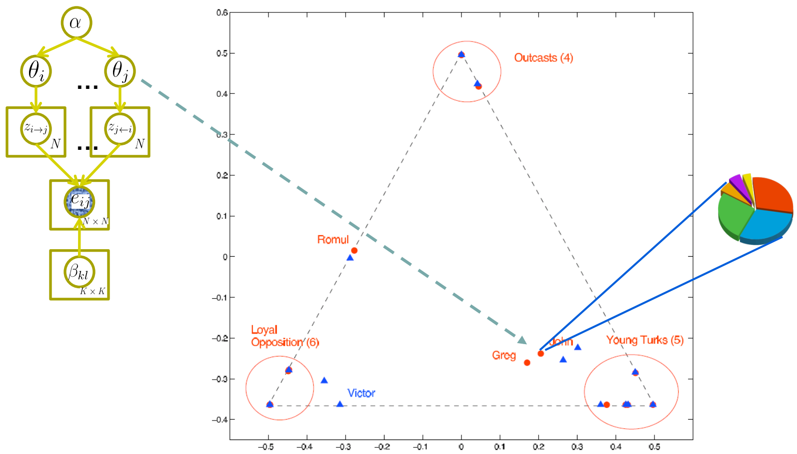


Mixed Membership Stochastic Blockmodel [Airoldi, Blei, Fienberg and Xing, 2008]

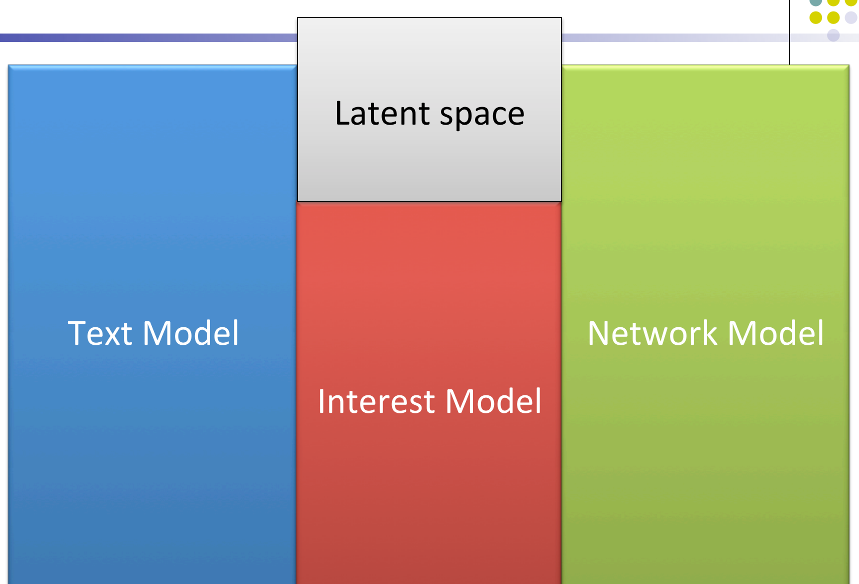


1. $\{\theta_i\}_{i=1}^N \sim p(\theta|\alpha) \equiv \text{Dirichlet}(\theta; \alpha)$ sample mixed membership vectors.
2. For each actor v_j that actor v_i possibly interacts with:
   - $z_{i \to j} \sim \text{Multinomial}(z|\theta_i)$ sample an indicator for v_i;
   - $z_{i \leftarrow j} \sim \text{Multinomial}(z|\theta_j)$ sample an indicator for v_j;
   - $e_{ij} \sim \text{Bernoulli}(e|z_{i \to j}^\top B z_{i \leftarrow j})$ sample a link.

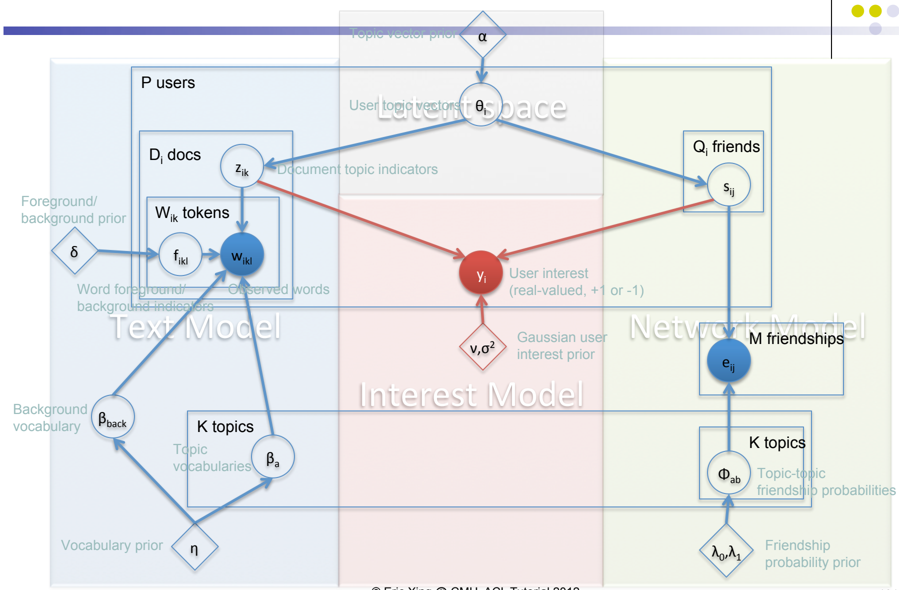
In the mixed-membership simplex [Airoldi, Blei, Fienberg and Xing, 2008]



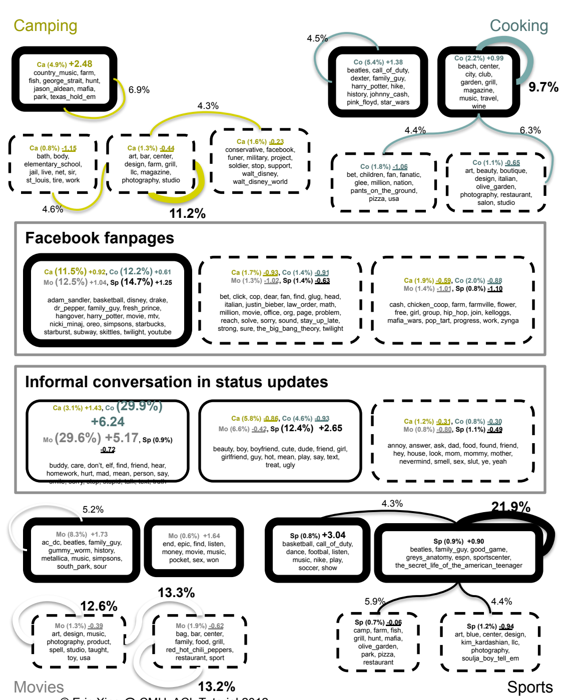
The "Facebook" model



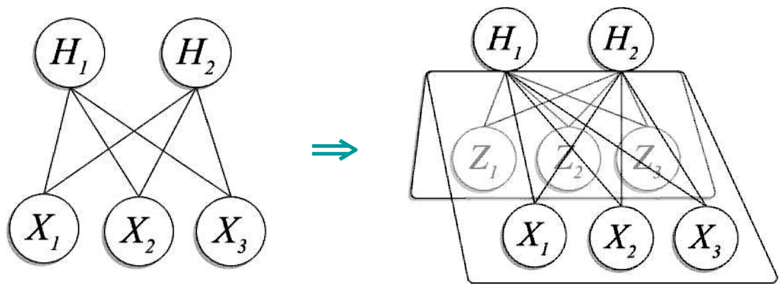
The "Facebook" model



A peep of the Facebook community



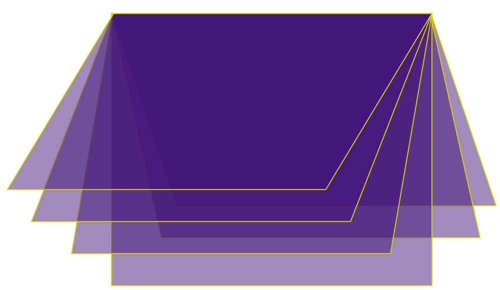
The Harmonium Counterpart [Xing et al, UAI 05]



Just add one more wing:

$$p(\mathbf{z}\mid\mathbf{h}) = \prod_k \text{Normal}\left[\ \sigma^2(\alpha_j + \sum_j W_{ij} h_j), \sigma^2\ \right]\ ,\qquad p(\mathbf{h}\mid\mathbf{x},\mathbf{z}) = \prod_j \text{Normal}\left[\ \sum_i W_{ij} x_i + \sum_k U_{kj} z_k, 1\ \right]$$

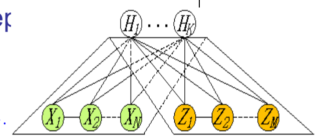
Multi-wing Harmoniums



Multi-view Markov Networks

- An simple undirected GM with conditional indep

$$p(\mathbf{x},\mathbf{z},\mathbf{h}) \propto \exp\Big\{ \sum_i \theta_i^\top \phi(x_i,x_{i+1}) + \sum_j \eta_j^\top \psi(z_j,z_{j+1}) + \sum_k \lambda_k^\top \varphi(h_k)$$
$$+ \sum_{ik} \phi(x_i,x_{i+1})^\top \mathbf{W}_i^k \varphi(h_k) + \sum_{jk} \psi(z_j,z_{j+1})^\top \mathbf{U}_j^k \varphi(h_k) \Big\}.$$



- Local conditional Markov networks (CRFs conditioned on latent H):

$$p(\mathbf{x}|\mathbf{h}) = \exp\Big\{ \sum_i \hat{\theta}_i^\top \phi(x_i,x_{i+1}) - A(\hat{\theta}) \Big\}, \text{ where } \hat{\theta}_i = \theta_i + \sum_k \mathbf{W}_i^k \varphi(h_k);$$

$$p(\mathbf{z}|\mathbf{h}) = \exp\Big\{ \sum_j \hat{\eta}_j^\top \psi(z_j,z_{j+1}) - B(\hat{\eta}) \Big\}, \text{ where } \hat{\eta}_j = \eta_j + \sum_k \mathbf{U}_j^k \varphi(h_k);$$

- Conditionally independent latent variables

$$p(\mathbf{h}|\mathbf{x},\mathbf{z}) = \prod_k \exp\Big\{ \hat{\lambda}_k^\top \varphi(h_k) - C_k(\hat{\lambda}_k) \Big\}, \text{ where } \hat{\lambda}_k = \lambda_k + \sum_i \mathbf{W}_i^k \phi(x_i,x_{i+1}) + \sum_j \mathbf{U}_j^k \psi(z_j,z_{j+1})$$

  - very efficient for fully observed view input data; potentially scale up to large data (e.g., millions of images) (Weston et al, 2010)

Examples of Latent Topics

T_1 storms gulf hawaii low forecast southeast showers

T_2 rebounds 14 shouting tests guard cut hawks

T_3 engine flying craft asteroid say hour aerodynamic

T_4 safe cross red sure dry providing services

T_5 losing jersey sixth antonio david york orlando

Are we done?

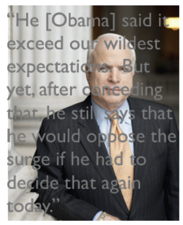
- What was our task?
  - Embedding (lower dimensional representation): yes, Doc → θ
  - Distillation of semantics: kind of, we've learned "topics" β
  - Classification: is it good?
  - Clustering: is it reasonable?
  - Other predictive tasks?

4. Scenario II: when supervision is available

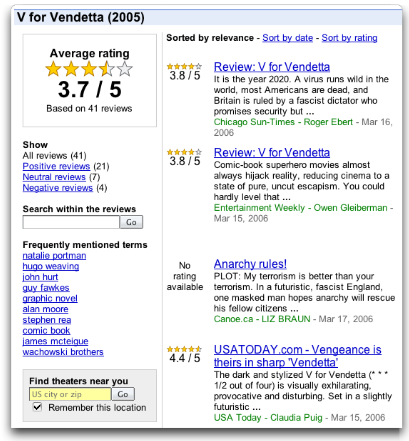
Problem: Discriminative topic models for text classification/scoring

- Democratic or republican?
- Movie review/scoring

Barack Obama



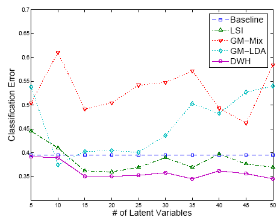
John McCain



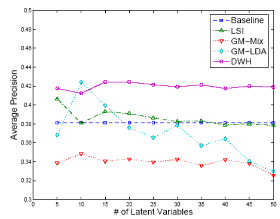
We want to answer …

- Are we satisfied with the conventional topic models and the MLE method for PREDICTION?
- Can we learn a PREDICTIVE model better?

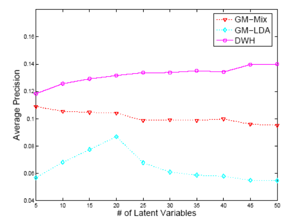
The shocking results on LDA



Classification



Retrieval

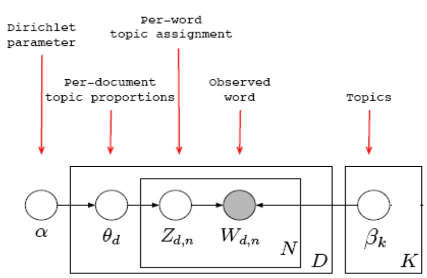


Annotation

- LDA is actually doing very poor on several "objectively" evaluatable predictive tasks

Why?

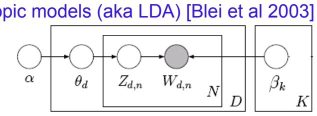
- LDA is not designed, nor trained for such tasks, such as classification, there is not warrantee that the estimated topic vector θ is good at discriminating documents



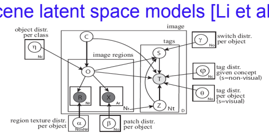
$$\overbrace{\begin{matrix}\theta_1 & \dots & \theta_D \\ 0.8 & \dots & 0.3 \\ 0.2 & \dots & 0.7\end{matrix}}^{\theta} \qquad \beta : \begin{pmatrix} 0.70 & 0.05 & 0.03 & \dots \\ 0.12 & 0.52 & 0.05 & \dots \end{pmatrix}$$

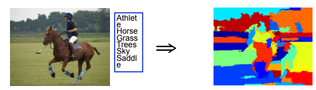
Unsupervised Latent Subspace Discovery

- Finding latent subspace representations (an old topic)
  - Mapping a high-dimensional representation into a latent low-dimensional representation, where each dimension can have some interpretable meaning, e.g., a semantic topic
- Examples:
  - Topic models (aka LDA) [Blei et al 2003]

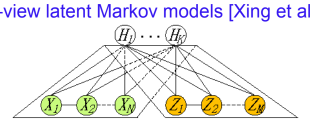


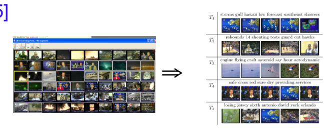
  - Total scene latent space models [Li et al 2009]





  - Multi-view latent Markov models [Xing et al 2005]





  - PCA, CCA, …

Predictive Subspace Learning with *Supervision*

- Unsupervised latent subspace representations are generic but can be sub-optimal for predictions
- Many datasets are available with supervised side information
  - Tripadvisor Hotel Review (http://www.tripadvisor.com)

  - LabelMe http://labelme.csail.mit.edu/

woman entering shop
man walking towards camera
balcony rail
building
sidewalk
manhole
light
person woman walking
ice cream
head
arm
torso

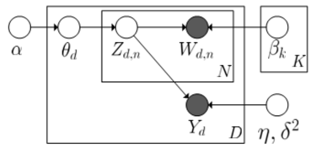
  - Many others

Flickr (http://www.flickr.com/)

IMAGENET

- Can be noisy, but not random noise (Ames & Naaman, 2007)
  - labels & rating scores are usually assigned based on some intrinsic property of the data
  - helpful to suppress noise and capture the most useful aspects of the data
- **Goals:**
  - **Discover latent subspace representations that are both *predictive* and *interpretable* by exploring weak supervision information**

I. Supervised Topic Model



(Blei & McAuliffe, 2007)

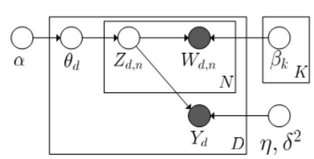
- How to integrate the max-margin principle into a probabilistic latent variable model?

Max-Likelihood Estimation	Max-Margin and Max-Likelihood
sLDA	MedLDA

(Zhu et al, ICML 2009)

Supervised Topic Model

- LDA ignores documents' side information (e.g., categories or rating score), thus lead to suboptimal topic representation for supervised tasks
- Supervised Topic Models handle such problems, e.g., sLDA (Blei & McAuliffe, 2007) and DiscLDA(Simon et al., 2008)
  - Generative Procedure (sLDA):
    - For each document d:
      - Sample a topic proportion $\theta_d \sim \mathrm{Dir}(\alpha)$
      - For each word:
        - Sample a topic $Z_{d,n} \sim \mathrm{Mult}(\theta_d)$
        - Sample a word $W_{d,n} \sim \mathrm{Mult}(\beta_{z_{d,n}})$
      - Sample y_d
        - $y_d \sim \mathcal{N}(\eta^\top \bar{Z}_d, \delta^2)$ ← Continuous (regression)
        - $y_d \sim \mathrm{GLM}(\bar{Z}_d, \eta^\top, \delta^2)$ ← Discrete (classification)



(Blei & McAuliffe, 2007)

How to train sLDA?

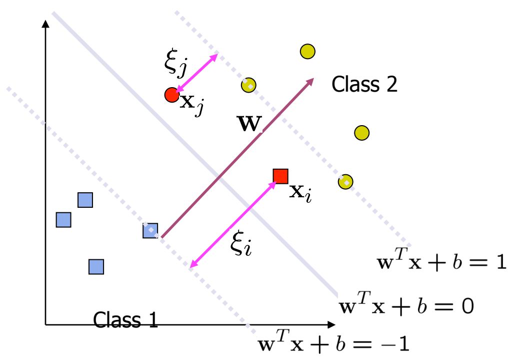
- Maximize $P(Y, W)$?

- Maximize

$$P(Y|W)?$$

- Support vector machines

Support vector machines



$$\min_{w,b} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$$

$$\text{s.t} \quad \begin{array}{l} y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i \\ \xi_i \geq 0, \quad \forall i \end{array}$$

SVM using VC-dimension

VC Theory

(Vapnik, 1982)
Given $x_1, \ldots, x_n \in \mathbb{R}^d$ iid and $||x_i||_2 \leq D$, if $\mathcal{H}_\gamma$ is the hypothesis space of linear classifiers in $\mathbb{R}^d$ with margin γ,

$$\boxed{VC(\mathcal{H}_\gamma) \leq \min\left\{d, \left\lceil \frac{4D^2}{\gamma^2} \right\rceil \right\}.}$$

$$\boxed{error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H)(\ln \frac{2m}{VC(H)} + 1) + \ln \frac{4}{\delta}}{m}}}$$

SVM using VC-dimension

- Thus large-margin → small VC-dim → better generalization bound
- Recall that d+1 is the upper bound for a linear classifier in d-space

$$\boxed{VC(\mathcal{H}_\gamma) \leq \min\left\{d, \left\lceil \frac{4D^2}{\gamma^2} \right\rceil \right\}.}$$

$$\boxed{error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H)(\ln \frac{2m}{VC(H)} + 1) + \ln \frac{4}{\delta}}{m}}}$$

MLE versus max-margin learning

- Likelihood-based estimation
  - Probabilistic (joint/conditional likelihood model)
  - Easy to perform Bayesian learning, and incorporate prior knowledge, latent structures, missing data
  - Bayesian regularization!!

- Max-margin learning
  - Non-probabilistic (concentrate on input-output mapping)
  - Not obvious how to perform Bayesian learning or consider prior, and missing data
  - Sound theoretical guarantee with limited samples

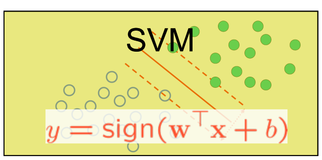
- Maximum Entropy Discrimination (MED) (Jaakkola, et al., 1999)
  - Model averaging $\hat{y} = \mathrm{sign} \int p(\mathbf{w}) F(x; \mathbf{w})\, \mathrm{d}\mathbf{w}$ $\quad (y \in \{+1, -1\})$
  - The optimization problem (binary classification)

$$\min \; KL(p(\Theta) || p_0(\Theta))$$
$$\text{s.t.} \int p(\Theta)[y_i F(x; \mathbf{w}) - \xi_i]\, \mathrm{d}\Theta \geq 0, \forall i,$$

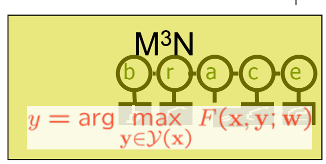
MED subsumes SVM.

where Θ is the parameter $\mathbf{w}$ when ξ are kept fixed or the pair $(\mathbf{w}, \xi)$ when we want to optimize over ξ

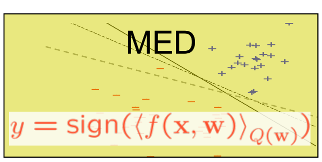
A road map for max-margin learning



$$\min_{\mathbf{w},\xi} \; \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} \xi_i$$
$$y^i(\mathbf{w}^\top \mathbf{x}^i + b) \geq 1 - \xi_i, \quad \forall i$$



$$\min_{\mathbf{w},\xi} \; \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} \xi_i$$
$$\mathbf{w}^\top[\mathbf{f}(\mathbf{x}^i, \mathbf{y}^i) - \mathbf{f}(\mathbf{x}^i, \mathbf{y})] \geq \ell(\mathbf{y}^i, \mathbf{y}) - \xi_i, \quad \forall i, \forall \mathbf{y} \neq \mathbf{y}^i$$



$$\min_{Q} \; \mathrm{KL}(Q || Q_0)$$
$$y^i \langle f(\mathbf{x}^i) \rangle_Q \geq \xi_i, \quad \forall i$$

MED-MN ?

= SMED + "Bayesian" M3N

Primal and Dual Sparse!

MaxEnt Discrimination Markov Network

- Structured MaxEnt Discrimination (SMED):

$$\text{P1}: \quad \min_{p(\mathbf{w}),\xi} \boxed{KL(p(\mathbf{w})||p_0(\mathbf{w})) + U(\xi)}$$
$$\text{s.t.} \;\; p(\mathbf{w}) \in \mathcal{F}_1, \;\; \xi_i \geq 0, \forall i.$$

generalized maximum entropy or *regularized* KL-divergence

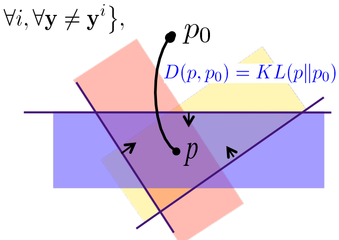
- Feasible subspace of weight distribution:

$$\mathcal{F}_1 = \left\{ p(\mathbf{w}) : \boxed{\int p(\mathbf{w})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta\ell_i(\mathbf{y})]\, \mathrm{d}\mathbf{w} \geq -\xi_i,} \; \forall i, \forall \mathbf{y} \neq \mathbf{y}^i \right\},$$

expected margin constraints.

- Average from distribution of M³Ns

$$h_1\big(\mathbf{x}; p(\mathbf{w})\big) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \int p(\mathbf{w}) F(\mathbf{x},\mathbf{y};\mathbf{w}) d\mathbf{w}$$



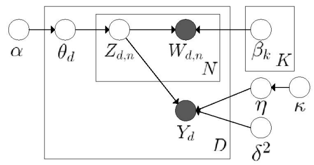
MedLDA: a max-margin approach

- Big picture of supervised topic models
  - sLDA: optimizes the joint likelihood for regression and classification
  - DiscLDA: optimizes the conditional likelihood for classification ONLY
  - MedLDA: based on max-margin learning for both regression and classification

MedLDA Regression Model

(Zhu et al, ICML 2009)

- Bayesian sLDA:



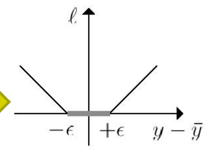
- MED Estimation:

$$\text{P1}(\text{MedLDA}^r): \quad \min_{q,\alpha,\beta,\delta^2,\xi,\xi^\star} \mathcal{L}(q) + C\sum_{d=1}^{D}(\xi_d + \xi_d^\star)$$

$$\text{s.t. } \forall d: \begin{cases} y_d - E[\eta^\top \bar{Z}_d] \leq \epsilon + \xi_d, & \mu_d \\ -y_d + E[\eta^\top \bar{Z}_d] \leq \epsilon + \xi_d^\star, & \mu_d^\star \\ \xi_d \geq 0, & v_d \\ \xi_d^\star \geq 0, & v_d^\star \end{cases}$$

model fitting

predictive accuracy



  - Variational bound $q(\theta, \mathbf{z}, \eta|\gamma, \phi) \sim p(\theta, \mathbf{z}, \eta|\alpha, \beta, \delta^2, \mathbf{y}, \mathbf{W})$

$$\mathcal{L}(q) \triangleq -E[\log p(\theta, \mathbf{z}, \eta, \mathbf{y}, \mathbf{W}|\alpha, \beta, \delta^2)] - \mathcal{H}(q(\mathbf{z}, \theta, \eta)) \geq -\log p(\mathbf{y}, \mathbf{W}|\alpha, \beta, \delta^2)$$

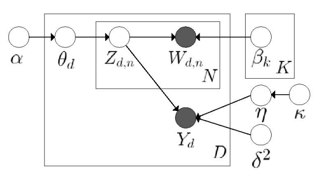
- Predictive Rule:

$$\bar{y} = E[Y|w_{1:N}, \alpha, \beta, \delta^2] = E_{q(Z,\eta)}[\eta^\top \bar{Z}|w_{1:N}, \alpha, \beta, \delta^2]$$

MedLDA Classification Model

(Zhu et al, ICML 2009)

- Bayesian sLDA:



- Multiclass MedLDA Classification Model:

$$\text{P2}(\text{MedLDA}^c): \quad \min_{q,q(\eta),\alpha,\beta,\xi} \mathcal{L}(q) + C\sum_{d=1}^{D}\xi_d$$

$$\text{s.t. } \forall d, \ y \neq y_d: \ E[\eta^\top \Delta\mathbf{f}_d(y)] \geq 1 - \xi_d; \ \xi_d \geq 0,$$

  - Variational bound $q(\theta, \mathbf{z}, \eta|\gamma, \phi) \sim p(\theta, \mathbf{z}, \eta|\alpha, \beta, \delta^2, \mathbf{y}, \mathbf{W})$

$$\mathcal{L}(q) \triangleq -E[\log p(\theta, \mathbf{z}, \eta, \mathbf{y}, \mathbf{W}|\alpha, \beta, \delta^2)] - \mathcal{H}(q(\mathbf{z}, \theta, \eta)) \geq -\log p(\mathbf{y}, \mathbf{W}|\alpha, \beta, \delta^2)$$

- Predictive Rule:

$$y^\star = \arg\max_y E[\eta^\top \mathbf{f}(y, \bar{Z})|\alpha, \beta]$$

Variational EM Alg.

- **E-step**: infer the posterior distribution of hidden r.v. $(\theta,\ \mathbf{z},\ \eta)$
- **M-step**: estimate unknown parameters $(\alpha,\ \beta,\ \delta^2)$

- Independence assumption: $q(\theta, \mathbf{z}, \eta|\gamma, \phi) = q(\eta) \prod_{d=1}^{D} q(\theta_d|\gamma_d) \prod_{n=1}^{N} q(z_{dn}|\phi_{dn})$

$$L(\gamma, \phi, q(\eta),\ \alpha, \beta, \delta^2, \xi, \xi^\star, \mu, \mu^\star, v, v^\star) = \mathcal{L}(q) + C\sum_{d=1}^{D}(\xi_d + \xi_d^\star) - \sum_{d=1}^{D}\sum_{i=1}^{N} c_{di}(\sum_{j=1}^{K}\phi_{dij} - 1)$$
$$- \sum_{d=1}^{D}\mu_d(\epsilon + \xi_d - y_d + E[\eta^\top \bar{Z}_d]) - \sum_{d=1}^{D}(\mu_d^\star(\epsilon + \xi_d^\star + y_d - E[\eta^\top \bar{Z}_d]) + v_d\xi_d + v_d^\star\xi_d^\star)$$

- Optimize *L* over ϕ:

$$\phi_{di} \propto \exp\left(\underline{E[\log\theta|\gamma] + E[\log p(w_{di}|\beta)]} + \underline{\frac{y_d}{N\delta^2}E[\eta] - \frac{2E[\eta^\top\phi_{d,-i}\eta] + E[\eta\circ\eta]}{2N^2\delta^2}} + \underline{\frac{E[\eta]}{N}(\mu_d - \mu_d^\star)}\right)$$

  - The first two terms are the same as in LDA
  - The third and fourth terms are similar to those of sLDA, but in expected version. The variance matters!
  - The last term is a regularizer. Only support vectors affect the topic proportions
- Optimize *L* over other variables. See the paper for details!

MedTM: a general framework

- MedLDA can be generalized to arbitrary topic models:
  - Unsupervised or supervised
  - Generative or undirected random fields (e.g., Harmoniums)

- MED Topic Model (MedTM):

$$\mathrm{P(MedTM)}: \min_{q(H),q(\Upsilon),\Psi,\xi} \underbrace{\mathcal{L}(q(H))}_{\text{model fitting}} + \underbrace{KL(q(\Upsilon)\|p_0(\Upsilon)) + U(\xi)}_{\text{predictive accuracy}}$$
$$\text{s.t. } \textit{expected} \text{ margin constraints}$$

- H: hidden r.v.s in the underlying topic model, e.g., $(\theta, \mathbf{z})$ in LDA
- Υ: parameters in predictive model, e.g., η in sLDA
- Ψ: parameters of the topic model, e.g., α in LDA
- $\mathcal{L}$: an variational upper bound of the log-likelihood
- U: a convex function over slack variables

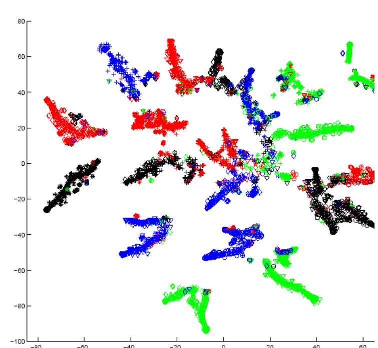
Experiments

- Goal:
  - To qualitatively and quantitatively evaluate how the max-margin estimates of MedLDA affect its topic discovering procedure

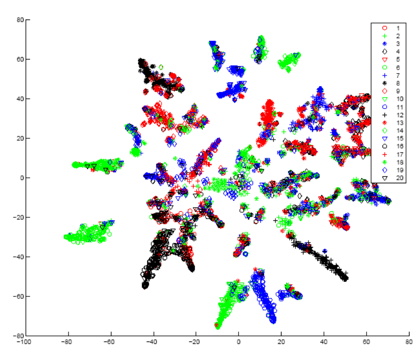
- Data Sets:
  - 20 Newsgroups (classification)
    - Documents from 20 categories
    - ~ 20,000 documents in each group
    - Remove stop word as listed in UMASS Mallet
  - Movie Review (regression)
    - 5006 documents, and 1.6M words
    - Dictionary: 5000 terms selected by tf-idf
    - Preprocessing to make the response approximately normal (Blei & McAuliffe, 2007)

Document Modeling

- Data Set: 20 Newsgroups
- 110 topics + 2D embedding with t-SNE (var der Maaten & Hinton, 2008)



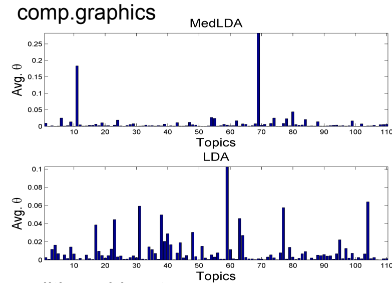
MedLDA



LDA

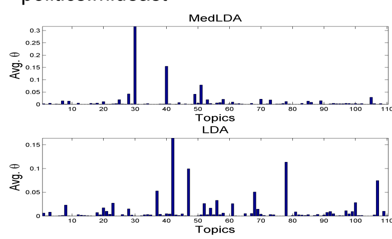
Document Modeling (cont')

comp.graphics



MedLDA			LDA		
T 69	T 11	T 80	T 59	T 104	T 31
image	graphics	db	image	ftp	card
jpeg	image	key	jpeg	pub	monitor
gif	data	chip	color	graphics	dos
file	ftp	encryption	file	mail	video
color	software	clipper	gif	version	apple
files	pub	system	images	tar	windows
bit	mail	government	format	file	drivers
images	package	keys	bit	information	vga
format	fax	law	files	send	cards
program	images	escrow	display	server	graphics

politics.mideast

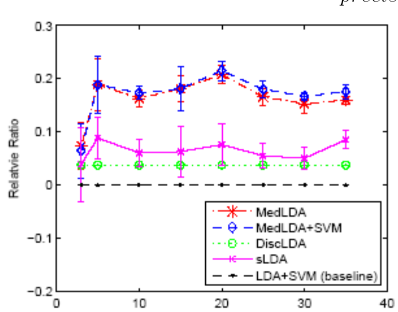


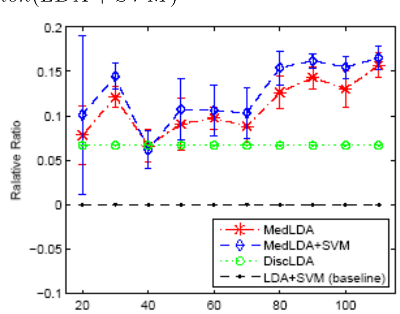
T 30	T 40	T 51	T 42	T 78	T 47
israel	turkish	israel	israel	jews	armenian
israeli	armenian	lebanese	israeli	jewish	turkish
jews	armenians	israeli	peace	israel	armenians
arab	armenia	lebanon	writes	israeli	armenia
writes	people	people	article	arab	turks
people	turks	attacks	arab	people	genocide
article	greek	soldiers	war	arabs	russian
jewish	turkey	villages	lebanese	center	soviet
state	government	peace	lebanon	jew	people
rights	soviet	writes	people	nazi	muslim

Classification

- **Data Set:** 20Newsgroups
  - Binary classification: "alt.atheism" and "talk.religion.misc" (Simon et al., 2008)
  - Multiclass Classification: all the 20 categories
- **Models**: DiscLDA, sLDA (Binary ONLY! Classification sLDA (Wang et al., 2009)), LDA+SVM (baseline), MedLDA, MedLDA+SVM
- **Measure**: Relative Improvement Ratio

$$RR(\mathcal{M}) = \frac{precision(\mathcal{M})}{precision(LDA + SVM)} - 1$$

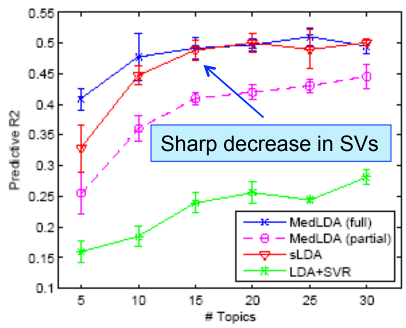


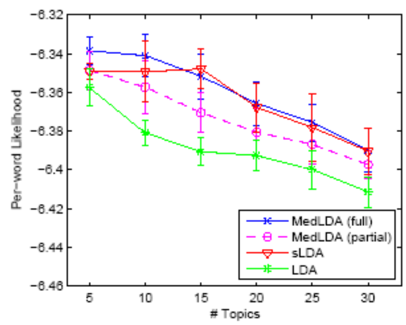


Regression

- **Data Set**: Movie Review (Blei & McAuliffe, 2007)
- **Models**: MedLDA(*partial*), MedLDA(*full*), sLDA, LDA+SVR
- **Measure**: predictive R^2 and per-word log-likelihood

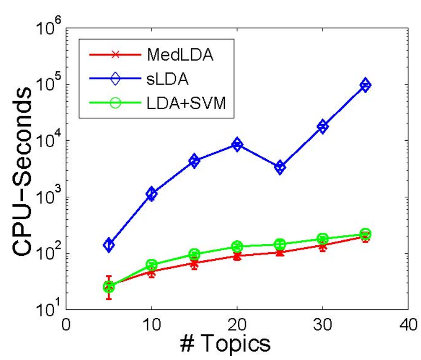
$$pR^2 = 1 - \frac{\sum_d (y_d - \hat{y}_d)^2}{\sum_d (y_d - \bar{y}_d)^2}$$





Time Efficiency

- Binary Classification



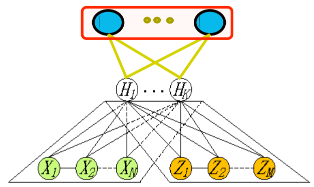
- Multiclass:
  - MedLDA is comparable with LDA+SVM
- Regression:
  - MedLDA is comparable with sLDA

II. Supervised Multi-view MNs

- A probabilistic method with an additional view of response variables Y

$$p(y|\mathbf{h}) = \frac{\exp\{\mathbf{V}^\top \mathbf{f}(\mathbf{h}, y)\}}{Z(V, \mathbf{h})}$$

normalization factor

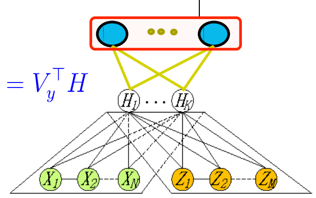


- Parameters can be learned with maximum likelihood estimation, e.g., special supervised Harmonium (Yang et al., 2007)
  - contrastive divergence is the commonly used approximation method in learning undirected latent variable models (Welling et al., 2004; Salakhutdinov & Murray, 2008).

Max-margin learning of MNs

- Expected discriminant function:

$$F(y; V) \triangleq \mathbb{E}_H[F(y, H; V)], \text{ where } F(y, H; V) = V_y^\top H$$



- Prediction rule:

$$y^* = \arg\max_y \mathbb{E}_{\mathbf{H}}[F(y, H; V)]$$

- Hinge loss:

$$\mathcal{R}_{hinge}(V) = \frac{1}{D}\sum_d \max_y[\Delta\ell_d(y) - V^\top \mathbb{E}_H[\Delta f_d(y)]],$$

- Joint max-margin and max-likelihood estimation:

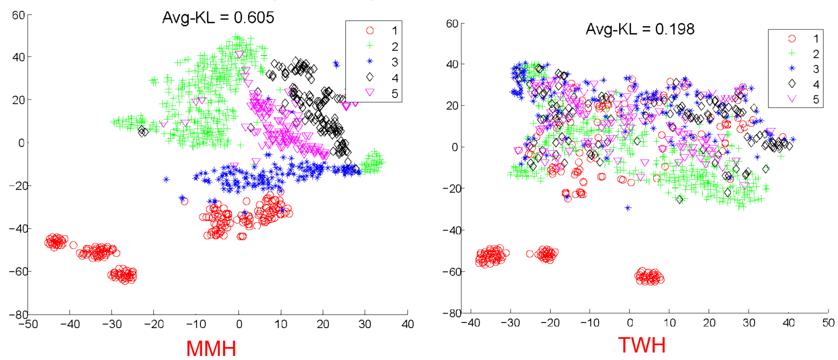
$$\min_{\Theta, V} \; -L(\Theta) + \frac{1}{2}C_1\|V\|_2^2 + C_2\mathcal{R}_{hinge}(V)$$

  - where $L(\theta) := \sum_d \log p(x_d, z_d)$ is data likelihood

- The rationale is: we want to find a latent representation and a prediction model , which on one hand tend to predict as accurate as possible on training data, while on the other hand tend to explain the data well.

Predictive Latent Representation

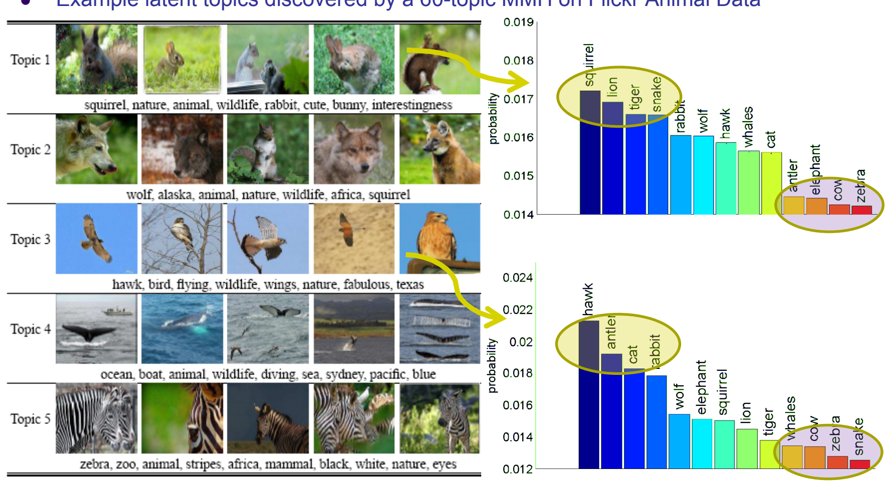
- t-SNE (van der Maaten & Hinton, 2008) 2D embedding of the discovered latent space representation on the TRECVID 2003



- Avg-KL: average pair-wise divergence

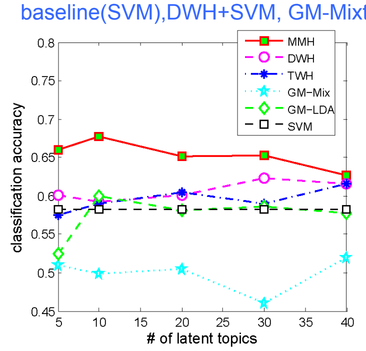
Predictive Latent Representation

- Example latent topics discovered by a 60-topic MMH on Flickr Animal Data

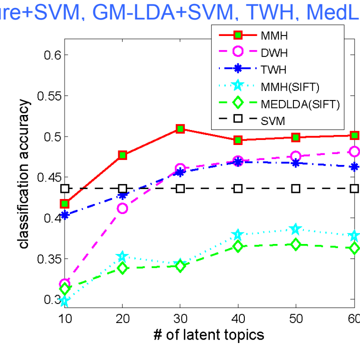


Classification Results

- Data Sets:
  - (Left) TRECVID 2003: (text + image features)
  - (Right) Flickr 13 Animal: (sift + image features)
- Models:
  - baseline(SVM),DWH+SVM, GM-Mixture+SVM, GM-LDA+SVM, TWH, MedLDA



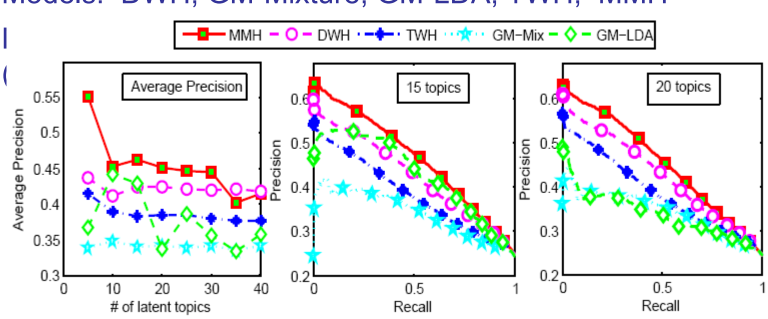
TRECVID



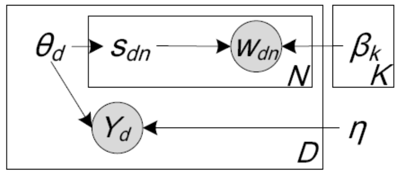
Flickr

Retrieval Results

- Data Set: TRECVID 2003
  - Each test sample is treated as a query, training samples are ranked based on the cosine similarity between a training sample and the given query
  - Similarity is computed based on the discovered latent topic representations
- Models: DWH, GM-Mixture, GM-LDA, TWH, MMH
- I
  (



III. Supervised STC



- Joint loss minimization

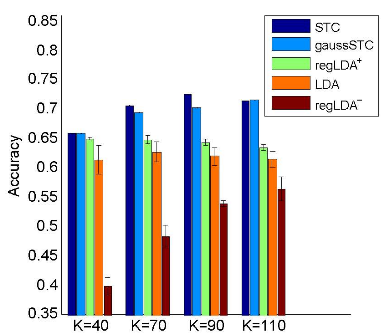
$$\min_{\{\theta_d\},\{\mathbf{s}_d\},\beta,\eta} \quad f(\{\theta_d\},\{\mathbf{s}_d\},\beta) + C\mathcal{R}_h(\{\theta_d\},\eta) + \frac{1}{2}\|\eta\|_2^2$$

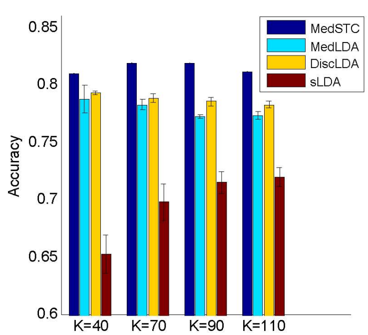
$$\text{s.t.}: \quad \theta_d \geq 0,\ \forall d;\ \mathbf{s}_{dn} \geq 0,\ \forall d, n \in I_d;\ \beta_k \in \mathcal{P},\ \forall k,$$

  - coordinate descent alg. applies with closed-form update rules
  - No sum-exp function; seamless integration with non-probabilistic large-margin principle

Classification accuracy

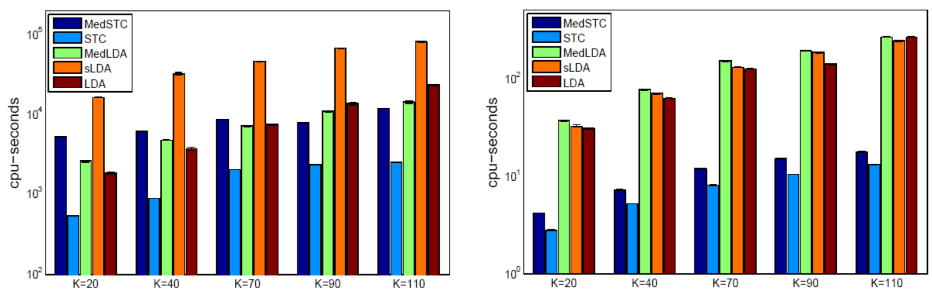
- 20 newsgroup data:





Time efficiency

- training & testing time



- No calls of digamma function
- Converge faster with one additional dimension of freedom

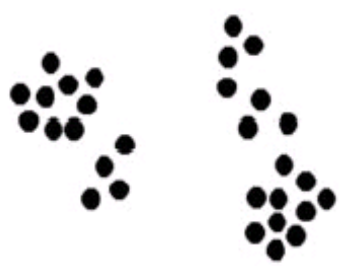
Summary

- Max-margin, instead of max-likelihood learning of supervised topic models (MedLDA, MMH, MedSTC)
  - Explicit interpretation of effects by support vectors
  - MedLDA can discover discriminative topic representations that are more suitable for supervised tasks
  - The classification model is efficient and can avoid dealing with the normalization factor of a GLM

- The same principle can be applied to a wide variety of probabilistic (MedTM) and non-probabilistic latent variable models

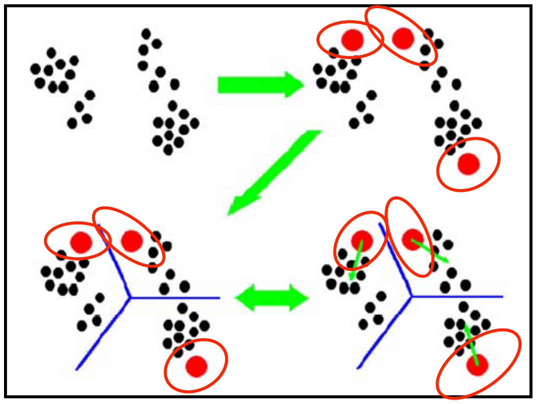
5. Scenario III: what if I don't know the total number of topics?

Clustering



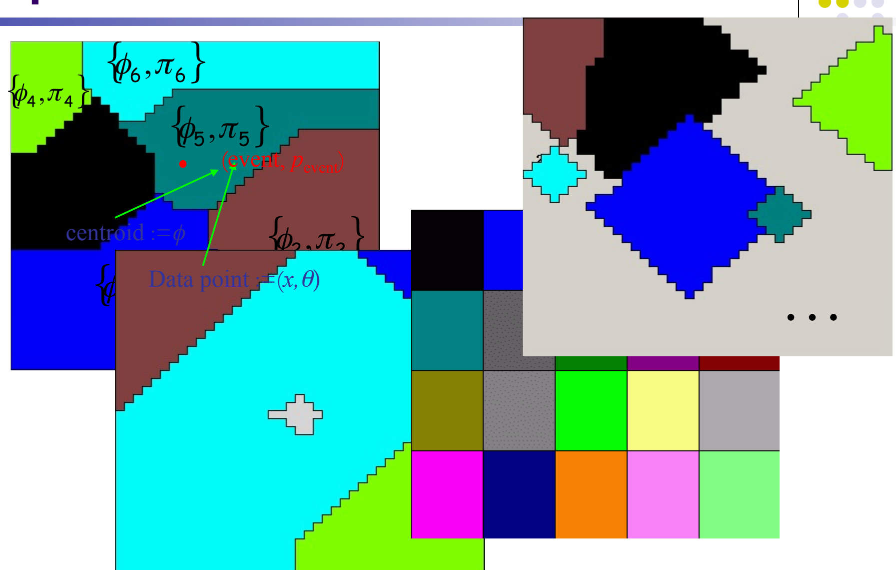
A Classical Approach

- Clustering as Mixture Modeling

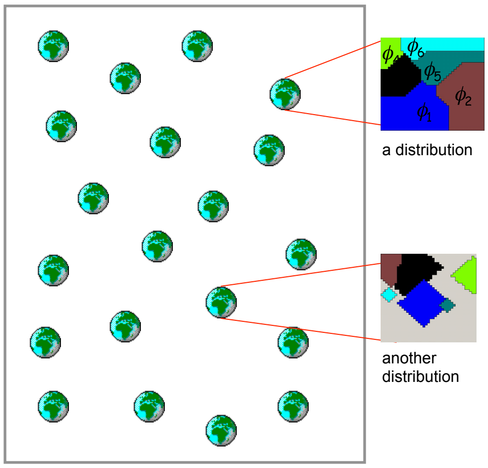


- Then "model selection"

Random Partition of Probability Space



Dirichlet Process



- A *CDF*, G, on possible worlds of random partitions follows a Dirichlet Process if for any measurable finite partition $(\phi_1, \phi_2, .., \phi_m)$:

$$(G(\phi_1), G(\phi_2), \dots, G(\phi_m)) \sim \text{Dirichlet}(\alpha G_0(\phi_1), \dots, \alpha G0(\phi_m))$$

where G_0 is the base measure and α is the scale parameter

Thus a Dirichlet Process G defines a distribution of distribution

Stick-breaking Process

$$G \sim \text{DP}(\alpha, G_0)$$

$$G = \prod_{k=1} \pi_k \delta(\theta_k)$$

$$\theta_k \sim G_0$$

Location

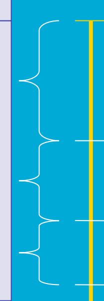
$$\sum_{k=1} \pi_k = 1$$

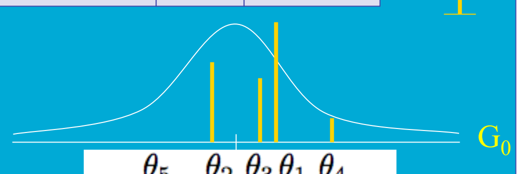
$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_k)$$

Mass

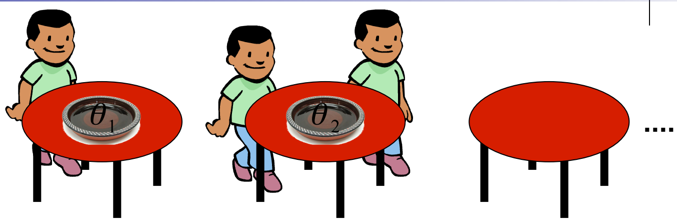
$$\beta_k \sim \text{Beta}(1, \alpha)$$

$\prod_{j=1}^{k-1}(1-\beta_j)$	β_k	π_k
0	0.4	0.4
0.6	0.5	0.3
0.3	0.8	0.24





Chinese Restaurant Process



$P(c_i = k \mid \mathbf{c}_{-i}) =$

1	0		0
$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$		0
$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$		$\frac{\alpha}{2+\alpha}$
$\frac{1}{3+\alpha}$	$\frac{2}{3+\alpha}$		$\frac{\alpha}{3+\alpha}$
$\frac{m_1}{i+\alpha-1}$	$\frac{m_2}{i+\alpha-1}$	....	$\frac{\alpha}{i+\alpha-1}$

This is a Dirichlet Process mixture

MCMC for CRP

- Gibbs sampling for exploring the posterior distribution under the proposed model
  - Under the CRP metaphor, due to exchangeability, every sample can be treated as the LAST sample!

$$p(c_i = k \mid \mathbf{c}_{[-i]}, \mathbf{x}, \theta) \propto p(c_i = k \mid \mathbf{c}_{[-i]})\, p(x_i \mid \theta_k, \mathbf{h}_{[-i]}, \mathbf{c}_{[-i]})$$

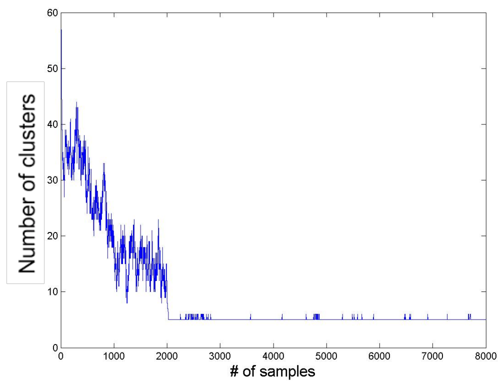
Posterior Prior x Likelihood

⋮

CRP

  - One can also integrate out the parameters such as θ and perform collapse Gibbs sampling
  - Gibbs sampling algorithm: draw samples of each random variable to be sampled given values of all the remaining variables

Convergence of Ancestral Inference



Variational Inference [Blei & Jordan 2005, Kurihara et al 2007]

- On a single machine Gibbs sampling solution is not efficient enough to scale up to the large scale problems.
- Truncated stick-breaking approximation can be formulated in the space of explicit, non-exchangeable cluster labels.
- Variational inference can now be applied to such a finite-dimensional distribution

- Variational Inference:
  - For a complicated $P(X_1, X_2, \dots X_n)$, approximate it with $Q(X)$:

$$Q(\mathbf{X}) = \prod_i Q(\mathbf{X}_{C_i})$$

$$\{Q^*(\mathbf{X}_{C_i})\} = \arg\min KL(Q(\mathbf{X})|P(\mathbf{X}))$$

Approximations to DP

- Truncated stick-breaking representation

$$\begin{aligned} &v_i \sim \mathcal{B}(v_i; 1, \alpha) && i = 1, ..., T-1 \\ &v_T = 1 && \\ &\pi_i = v_i \prod_{j<i}(1 - v_j) && i = 1, ..., T \\ &\pi_i = 0 && i > T \end{aligned}$$

- Finite symmetric Dirichlet approximation

$$\boldsymbol{\pi} \sim \mathcal{D}(\boldsymbol{\pi}; \tfrac{\alpha}{K}, ..., \tfrac{\alpha}{K})$$

- The joint distribution can be expressed as:

$$P(X, \mathbf{z}, \mathbf{v}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi}(\mathbf{v}))\right] \left[\prod_{i=1}^{T} p(\eta_i)\mathcal{B}(v_i; 1, \alpha)\right]$$

- The joint distribution can be expressed as:

$$P(X, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi})\right] \left[\prod_{i=1}^{K} p(\eta_i)\right] \mathcal{D}(\boldsymbol{\pi}; \tfrac{\alpha}{K}, ..., \tfrac{\alpha}{K})$$

VB inference

- We can then apply the VB inference on the four approximations

$$\{Q^*(\mathbf{X}_{C_i})\} = \arg\min KL(Q(\mathbf{X})|P(\mathbf{X}))$$

The approximated posterior distribution for TSB and FSD are

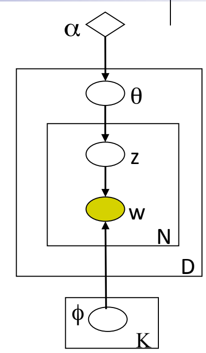
$$Q_{\text{TSB}}(\mathbf{z}, \boldsymbol{\eta}, \mathbf{v}) = \left[\prod_{n}^{N} q(z_n)\right] \left[\prod_{i=1}^{T} q(\eta_i) q(v_i)\right] \qquad Q_{\text{FSD}}(\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\pi}) = \left[\prod_{n}^{N} q(z_n)\right] \left[\prod_{k=1}^{K} q(\eta_k)\right] q(\boldsymbol{\pi})$$

Depending on marginalization or not, v and π may be integrated out.

How to build an infinite LDA?

Generative Process

-For each document d
- Sample $\theta_d \sim$ Dirichlet(α)
- For each word *w* in d
  - Sample z ~ Multi(θ_d)
  - Sample *w* ~ Multi(ϕ_z)



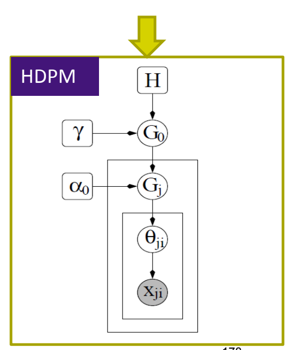
Topics' trends evolve over time?

Topics' distributions evolve over time?

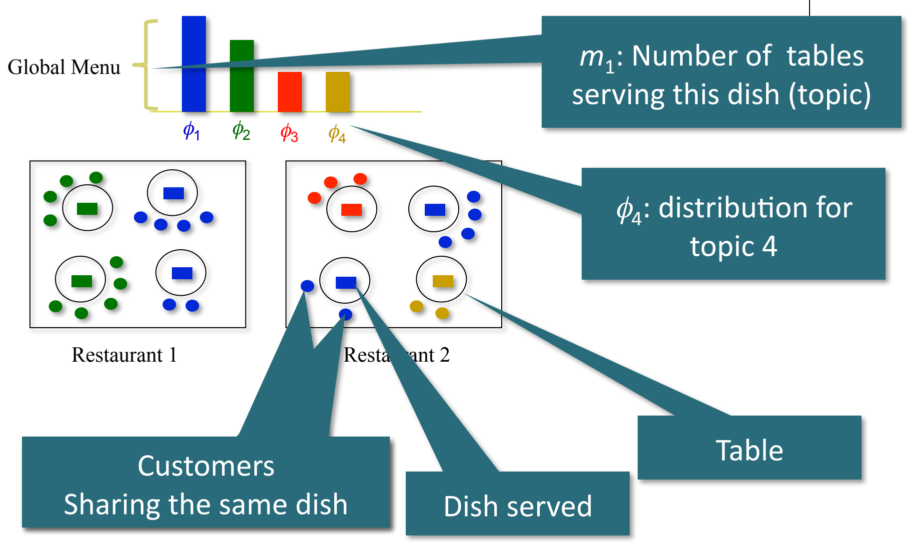
Number of topics grow with the data?

The Chinese Restaurant Franchise Process

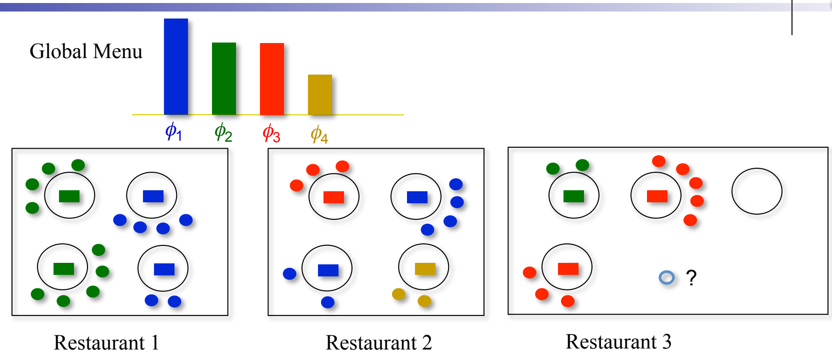
- Hierarchical Dirichlet Process Mixture (HDPM) automatically determines number of topics in LDA
- We will focus on the Chinese Restaurant Franchise process construction
  - A set of restaurants that share a global menu
- Metaphor
  - Restaurant = documents
  - Customer = word
  - Dish = topic
  - Global Menu = Set of topics



The Chinese Restaurant Franchise Process

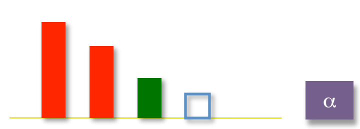


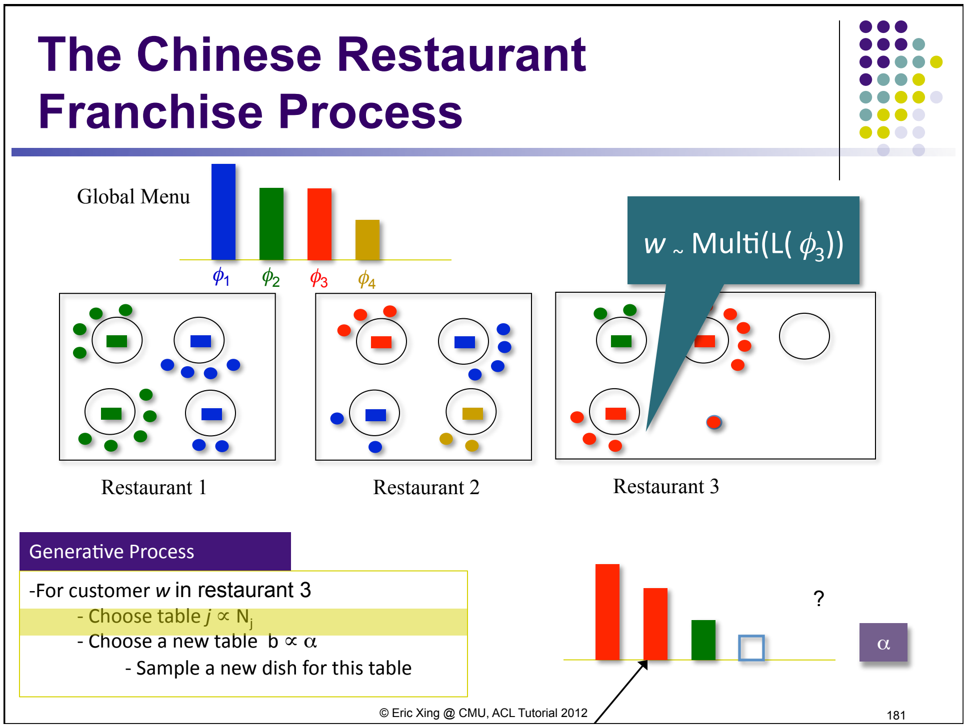
The Chinese Restaurant Franchise Process

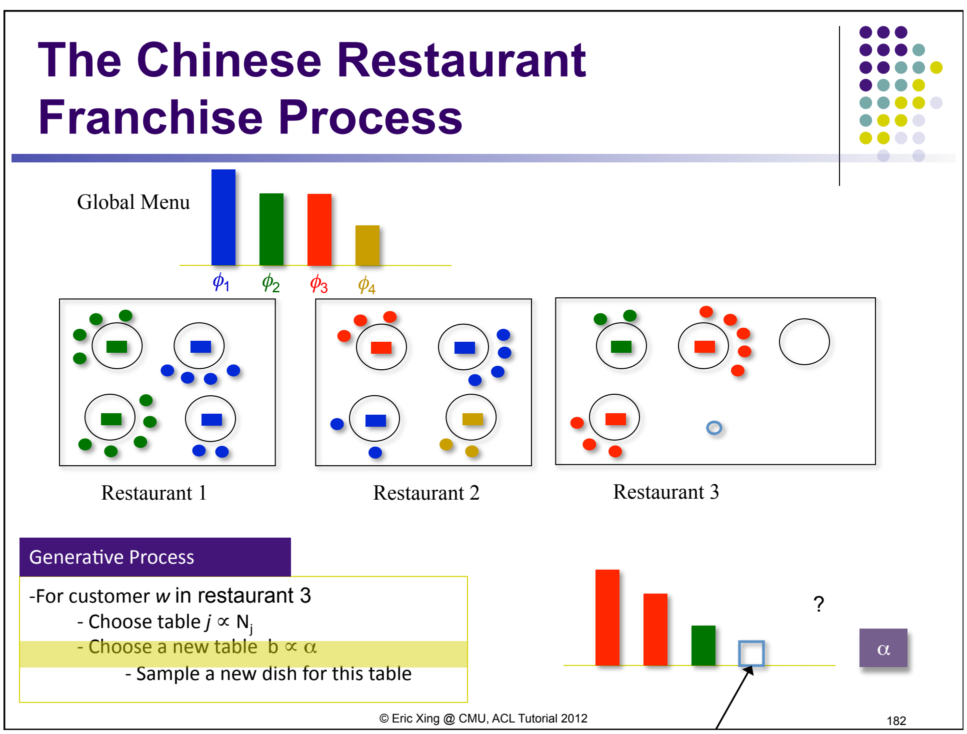


Generative Process

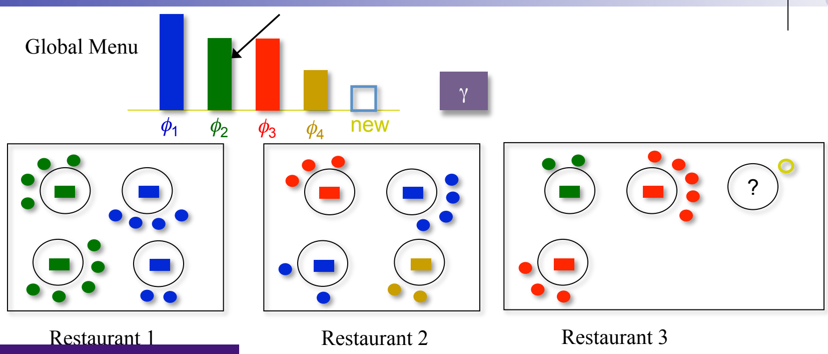
-For customer w in restaurant 3
- Choose table $j \propto N_j$
- Choose a new table $b \propto \alpha$
  - Sample a new dish for this table




The Chinese Restaurant Franchise Process
Global Menu
ϕ_1 ϕ_2 ϕ_3 ϕ_4
$w \sim$ Multi(L(ϕ_3))
Restaurant 1
Restaurant 2
Restaurant 3
Generative Process
-For customer w in restaurant 3
- Choose table $j \propto N_j$
- Choose a new table b $\propto \alpha$
- Sample a new dish for this table
?
α
© Eric Xing @ CMU, ACL Tutorial 2012
181

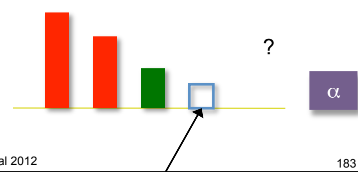

The Chinese Restaurant Franchise Process
Global Menu
ϕ_1 ϕ_2 ϕ_3 ϕ_4
Restaurant 1
Restaurant 2
Restaurant 3
Generative Process
-For customer w in restaurant 3
- Choose table $j \propto N_j$
- Choose a new table b $\propto \alpha$
- Sample a new dish for this table
?
α
© Eric Xing @ CMU, ACL Tutorial 2012
182

The Chinese Restaurant Franchise Process



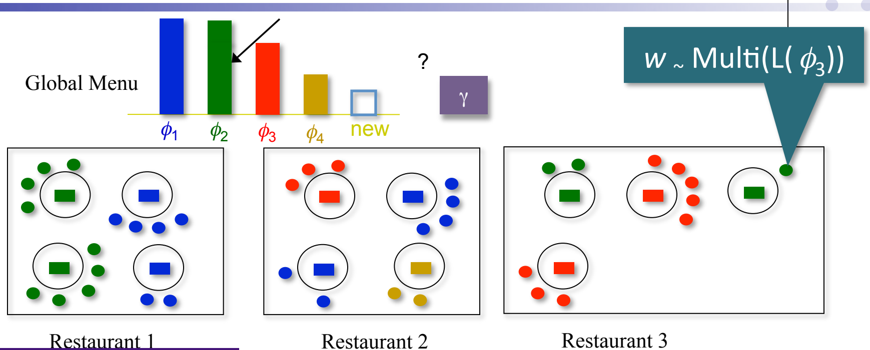
Generative Process

- For customer w in restaurant 3
  - Choose table $j \propto N_j$
  - Choose a new table $b \propto \alpha$
    - Sample a new dish for this table
    - Existing dish $k \propto m_k$
    - A new dish $\propto \gamma$



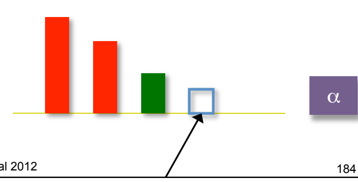
183

The Chinese Restaurant Franchise Process



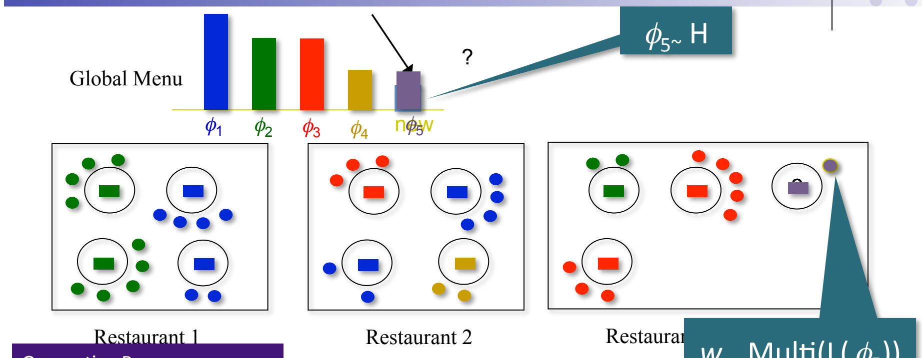
Generative Process

- For customer w in restaurant 3
  - Choose table $j \propto N_j$
  - Choose a new table $b \propto \alpha$
    - Sample a new dish for this table
    - Existing dish $k \propto m_k$
    - A new dish $\propto \gamma$



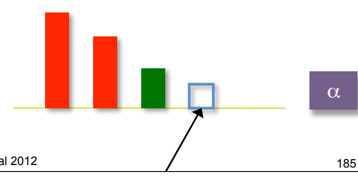
184

The Chinese Restaurant Franchise Process

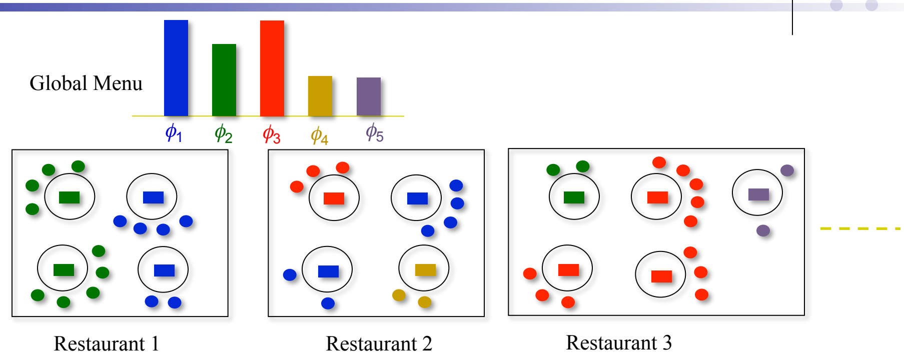


Generative Process

- For customer w in restaurant 3
  - Choose table $j \propto N_j$
  - Choose a new table $b \propto \alpha$
    - Sample a new dish for this table
    - Existing dish $k \propto m_k$
    - A new dish $\propto \gamma$

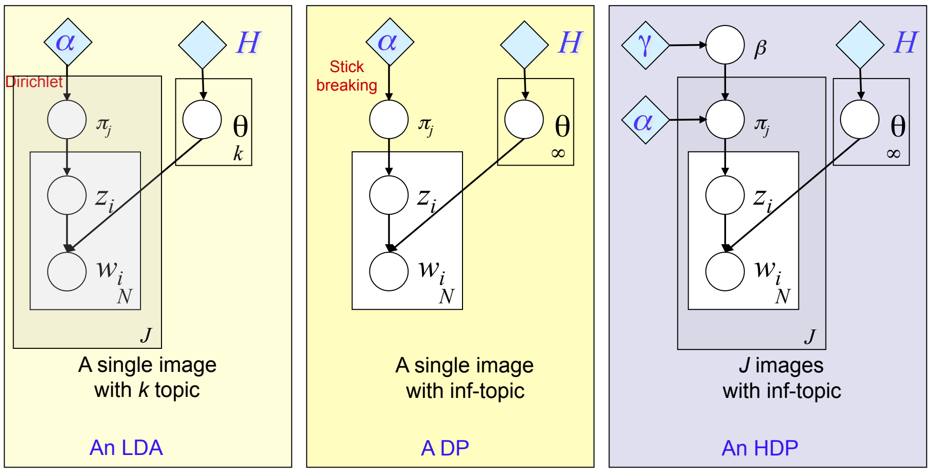


The Chinese Restaurant Franchise Process



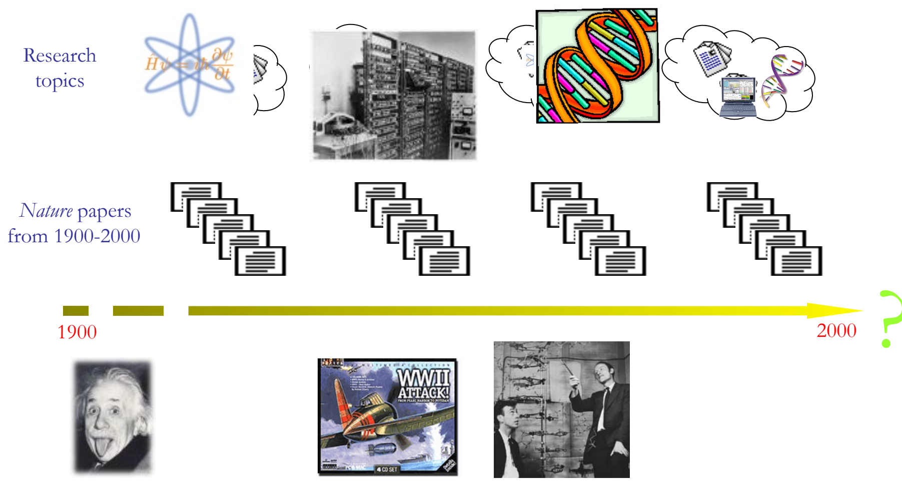
Topics' trends evolve over time?	✗
Topics' distributions evolve over time?	✗
Number of topics grow with the data?	✓

Summary: From LDA to Infinite Topic Models

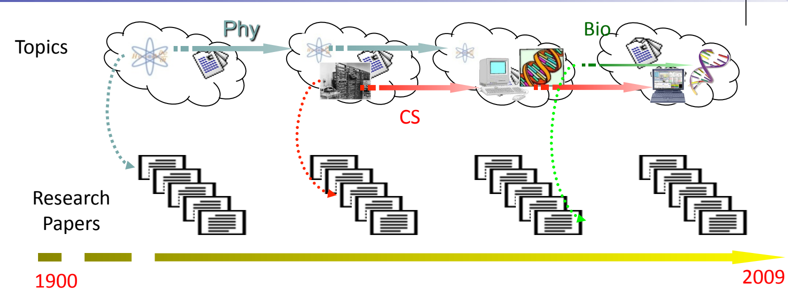


6. Scenario IV: Topic evolution in Streaming Corpus

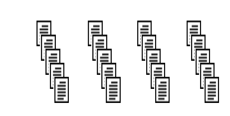
How to model topic evolution?



Problem Statement

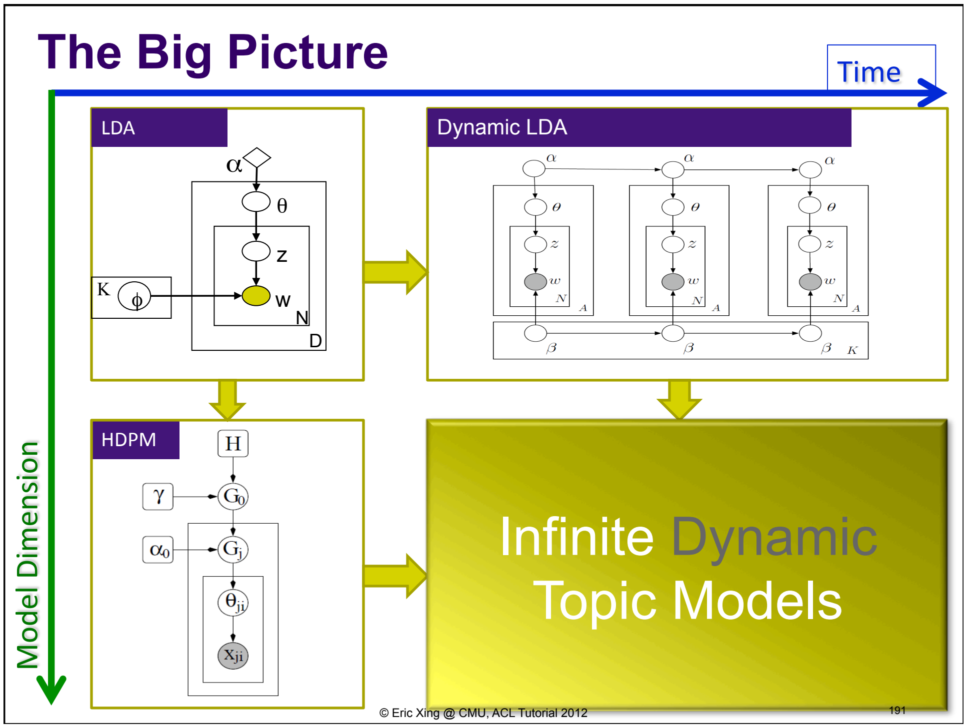


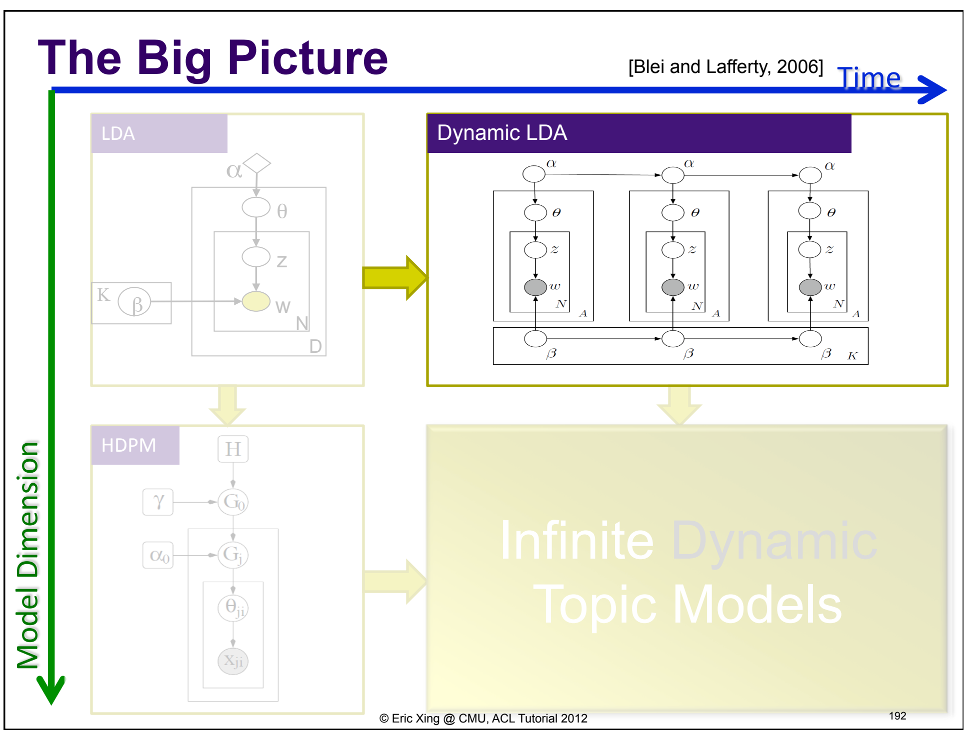
given



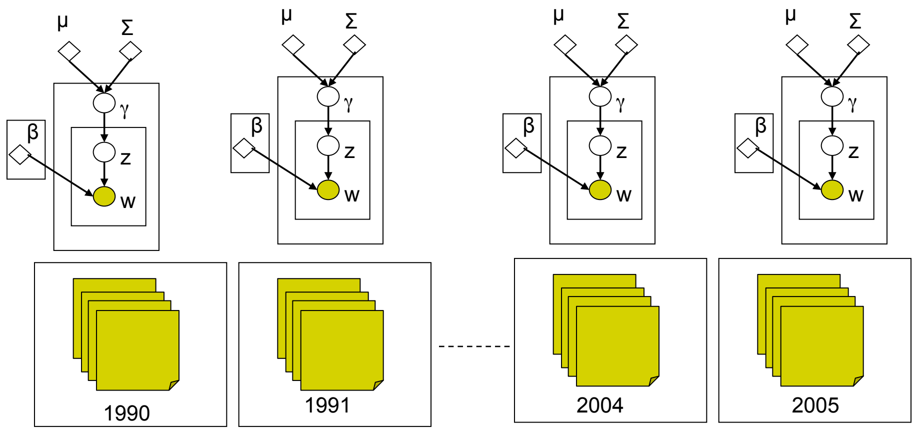
Discover

- Potentially infinite number of topics
  - With time-varying trends
  - And time-varying distributions
  - And variable durations
    - Topics can die
    - New topics can born

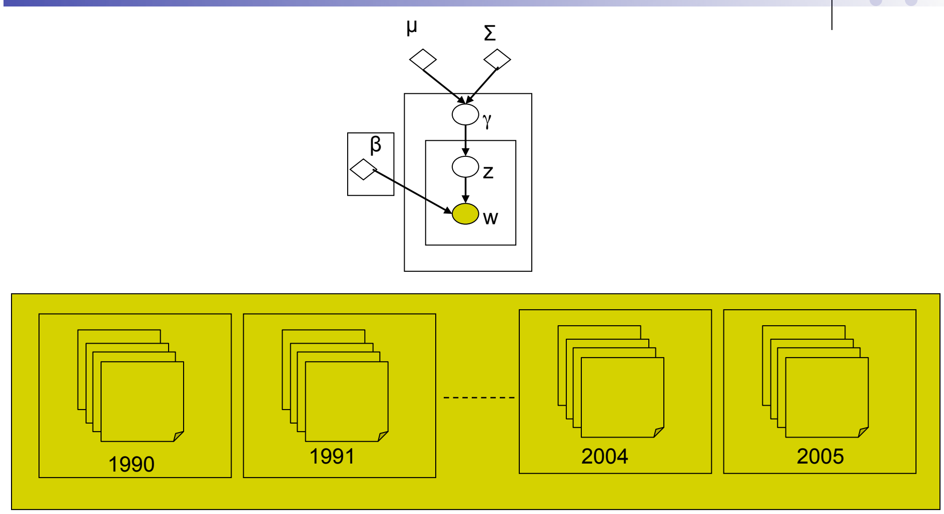

The Big Picture
Time
Model Dimension
LDA
Dynamic LDA
HDPM
Infinite Dynamic
Topic Models
© Eric Xing @ CMU, ACL Tutorial 2012
191


The Big Picture
[Blei and Lafferty, 2006]
Time
Model Dimension
LDA
Dynamic LDA
HDPM
Infinite Dynamic
Topic Models
© Eric Xing @ CMU, ACL Tutorial 2012
192

Text Stream



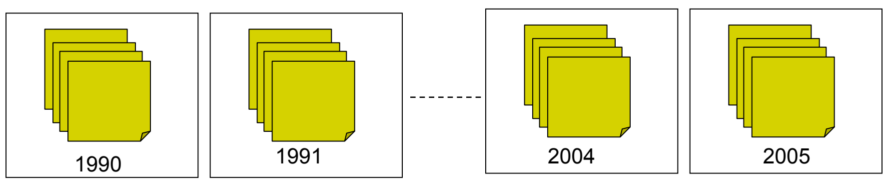
Text Stream



How to Model Topic Evolution

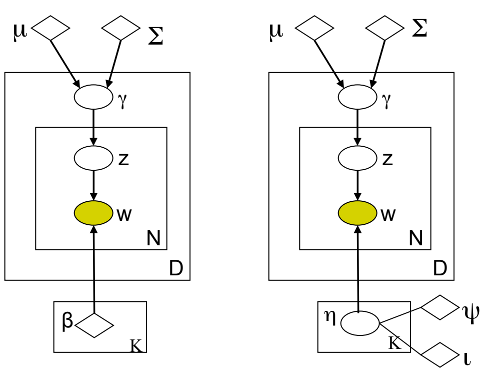
- Topic Trends
- Topic Keywords
- Topic correlations
- ~~Number of topics~~

The Dynamic Correlated Topic model

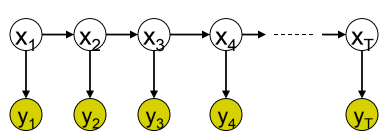


Building Blocks

CTMs



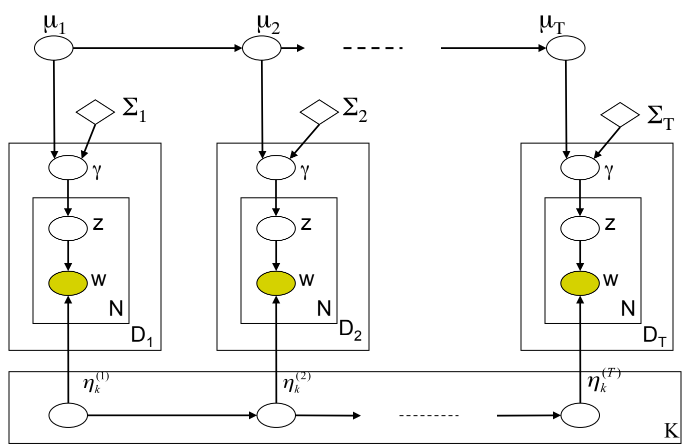
Kalman Filters



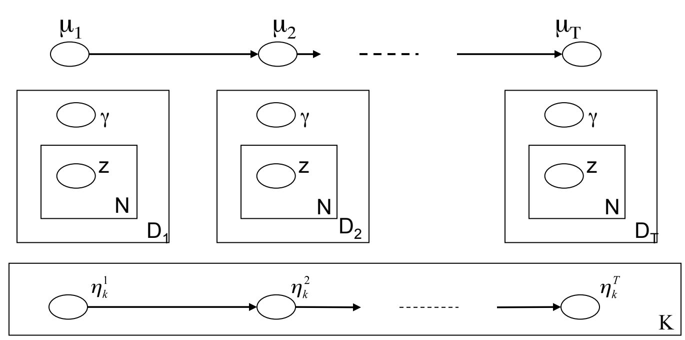
$$\begin{aligned} X_t &= AX_{t-1} + \xi \\ Y_t &= CX_t + \delta_t \end{aligned}$$

$$\begin{aligned} X_t | X_{t-1} &\sim \mathrm{N}(AX_t, \Phi) \\ Y_t | X_t &\sim \mathrm{N}(CX_t, \psi_t) \end{aligned}$$

The Dynamic CTM



Generalized Mean Field Inference



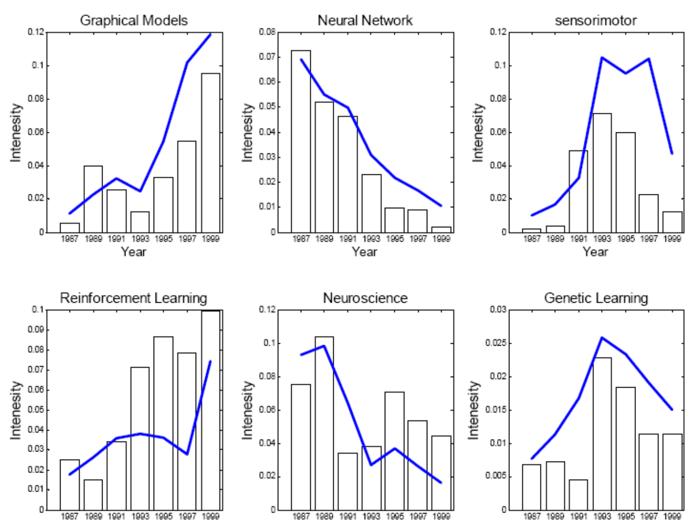
Generalized Mean Field Inference:

$$q(X) = P\left(X \mid \langle S_y \rangle_{q_y} : \forall y \in X_{MB}\right)$$

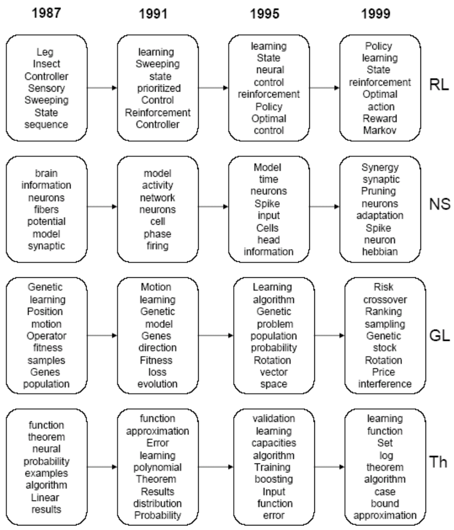
Experimental Results

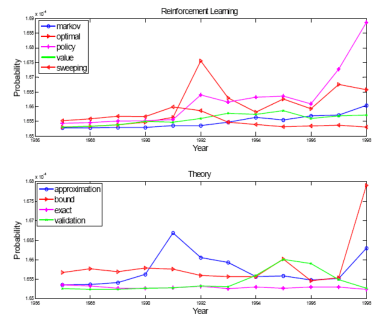
- NIPS data set
  - 12 years
  - 14036 words
  - 2484 docs
  - 90% for training and 10% for testing

Topic Trends

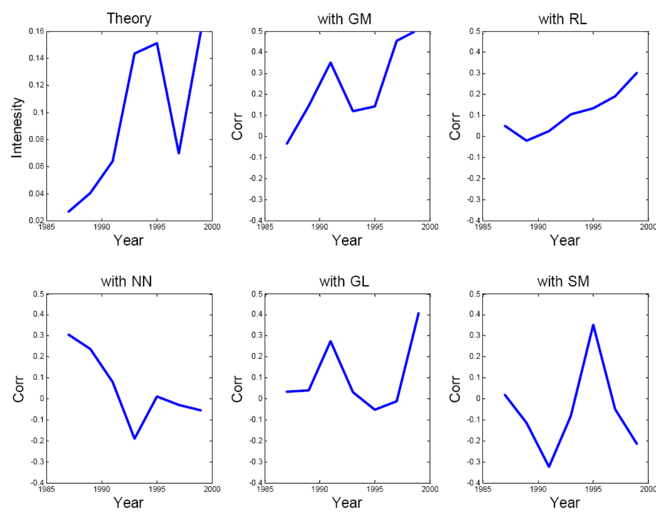


Topic Words over Time



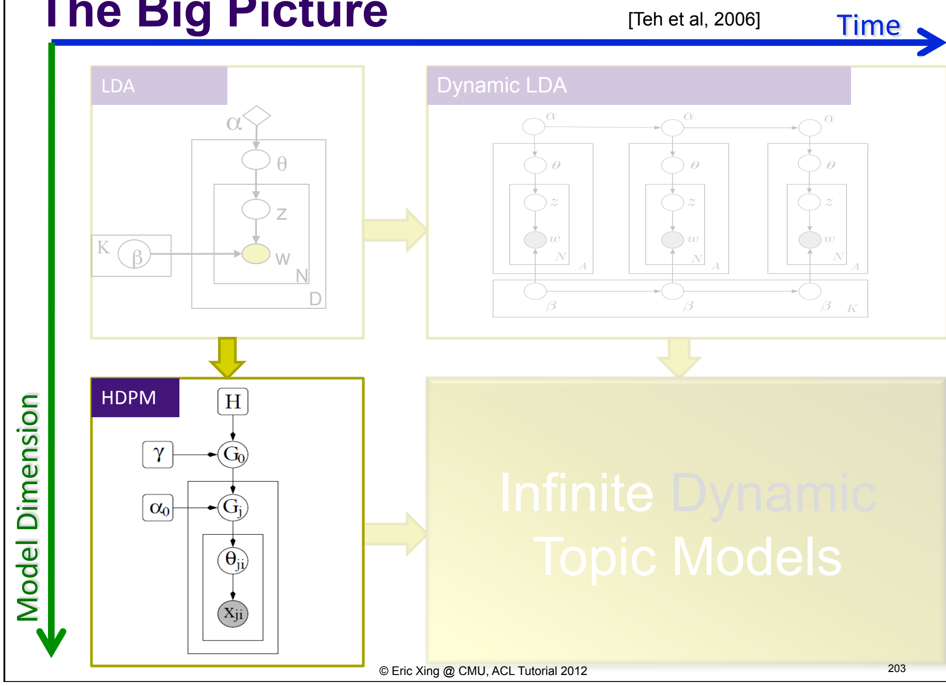


Topic Correlations Over Time



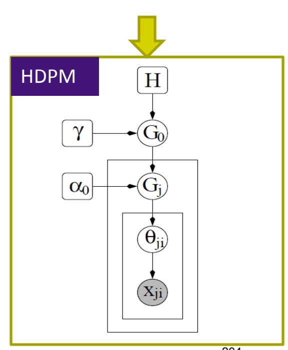
The Big Picture

[Teh et al, 2006]



The Chinese Restaurant Franchise Process

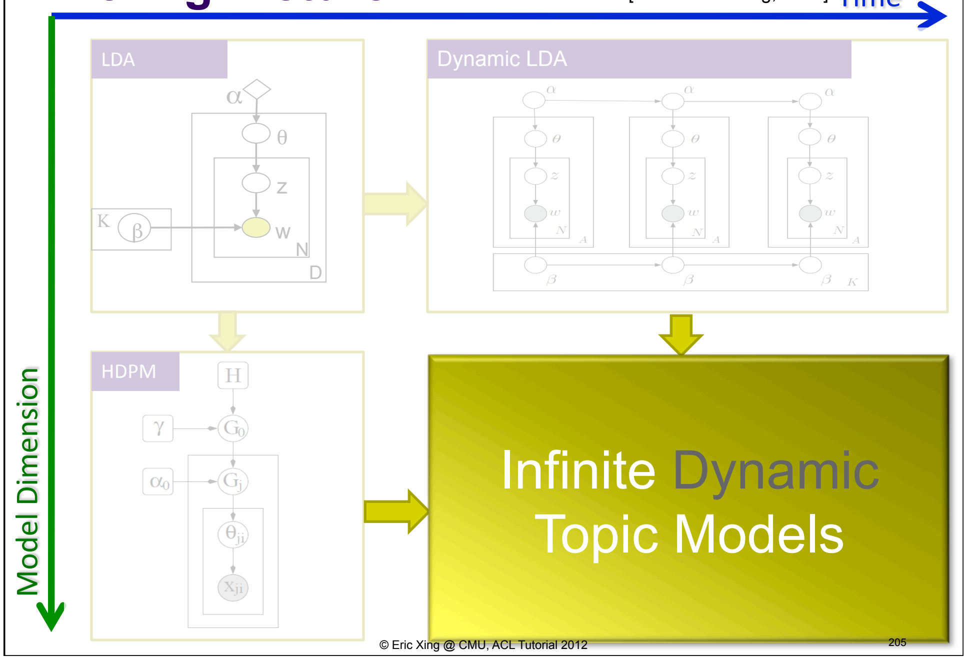
- HDPM automatically determines number of topics in LDA
- We will focus on the Chinese Restaurant Franchise process construction
  - A set of restaurants that share a global menu
- Metaphor
  - Restaurant = documents
  - Customer = word
  - Dish = topic
  - Global Menu = Set of topics



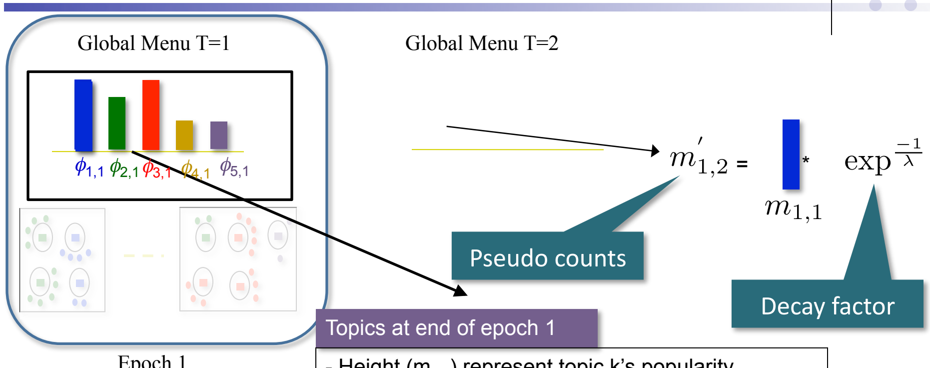
We have covered it already!

The Big Picture

[Ahmed and Xing, 2010]



The "Evolving" Chinese Restaurant Franchise Process



Topics at end of epoch 1

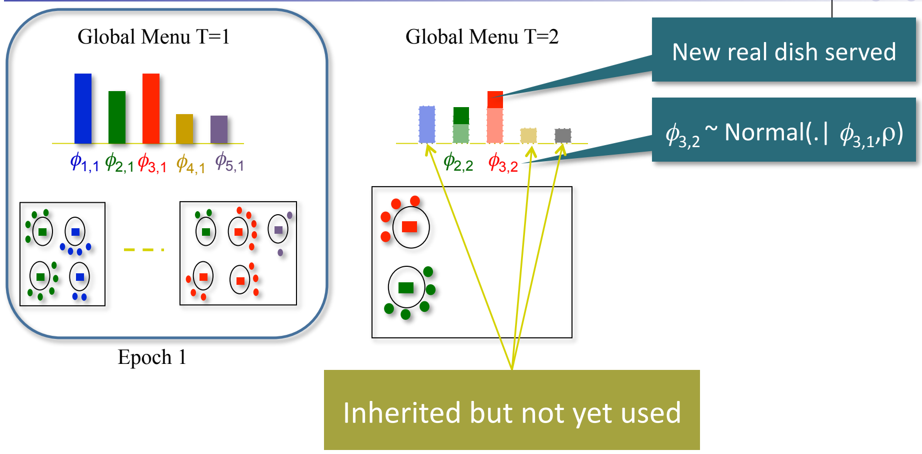
- Height ($m_{k,1}$) represent topic k's popularity
- $\phi_{k,1}$ represents topic k's word distribution

Documents in epoch 1 are generated as before

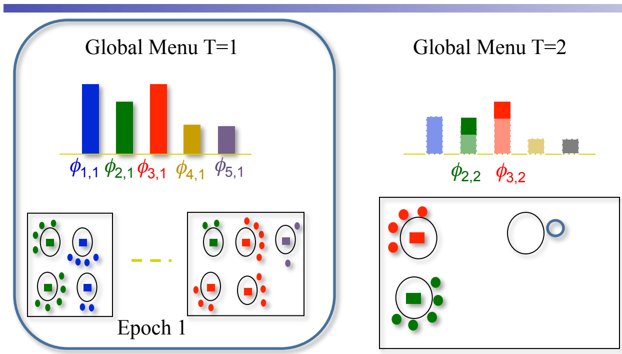
Observations

- Popular topics at epoch 1 are likely to be popular at epoch 2
- $\phi_{k,2}$ is likely to smoothly evolve from $\phi_{k,1}$

The "Evolving" Chinese Restaurant Franchise Process



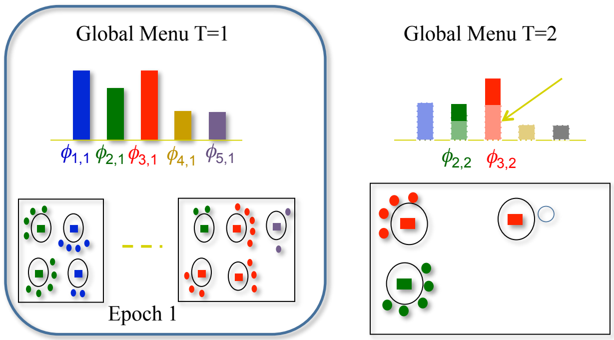
The "Evolving" Chinese Restaurant Franchise Process



Generative Process

-For customer *w* in restaurant 1
- -[as in static case] Choose table $j \propto N_j$
- - Choose a new table $b \propto \alpha$
  - -Sample a new dish for this table
  - - Existing and inherited dish $k \propto m`_{k,2} + m_{k,2}$
  - - Existing but NOT inherited dish $k \propto m`_{k,2}$ Then $\phi_{k,2}$ ~ Normal(.| $\phi_{k,1}$,ρ)
  - - A new dish $\propto \gamma$ Then ϕ_{new} ~ H

The "Evolving" Chinese Restaurant Franchise Process



Generative Process

-For customer w in restaurant 1

-[as in static case] Choose table $j \propto N_j$

- Choose a new table $b \propto \alpha$

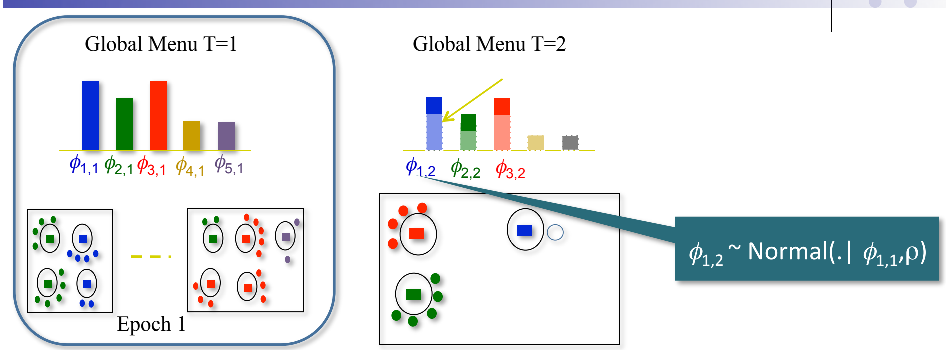
- Sample a new dish for this table

- Existing and inherited dish $k \propto m`_{k,2} + m_{k,2}$

- Existing but NOT inherited dish $k \propto m`_{k,2}$ Then $\phi_{k,2} \sim \mathrm{Normal}(.|\ \phi_{k,1},\rho)$

- A new dish $\propto \gamma$ Then $\phi_{new} \sim H$

The "Evolving" Chinese Restaurant Franchise Process



Generative Process

-For customer w in restaurant 1

-[as in static case] Choose table $j \propto N_j$

- Choose a new table $b \propto \alpha$

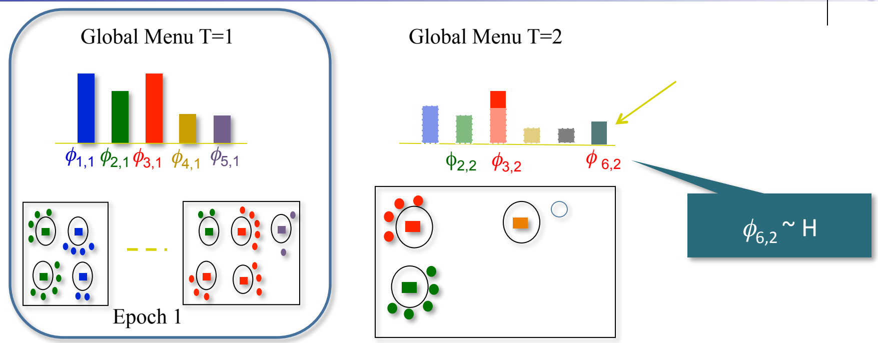
- Sample a new dish for this table

- Existing and inherited dish $k \propto m`_{k,2} + m_{k,2}$

- Existing but NOT inherited dish $k \propto m`_{k,2}$ Then $\phi_{k,2} \sim \mathrm{Normal}(.|\ \phi_{k,1},\rho)$

- A new dish $\propto \gamma$ Then $\phi_{new} \sim H$

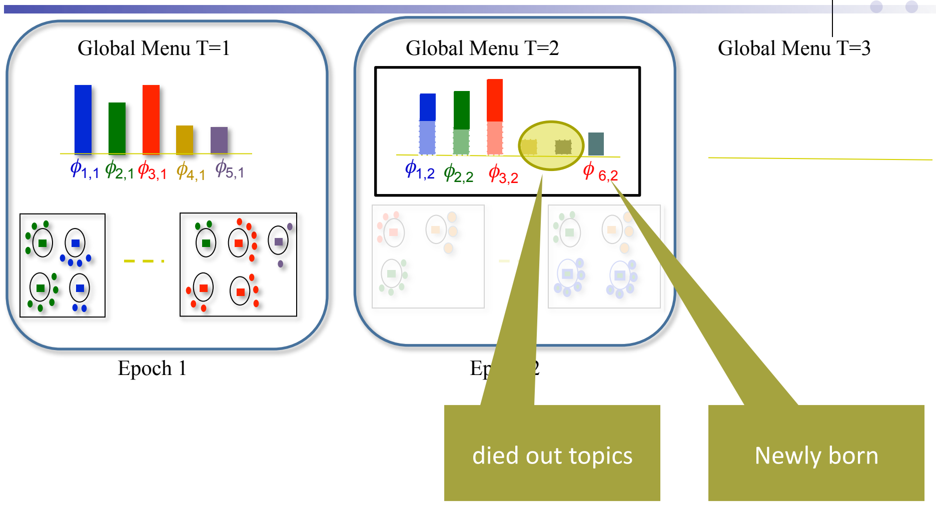
The "Evolving" Chinese Restaurant Franchise Process



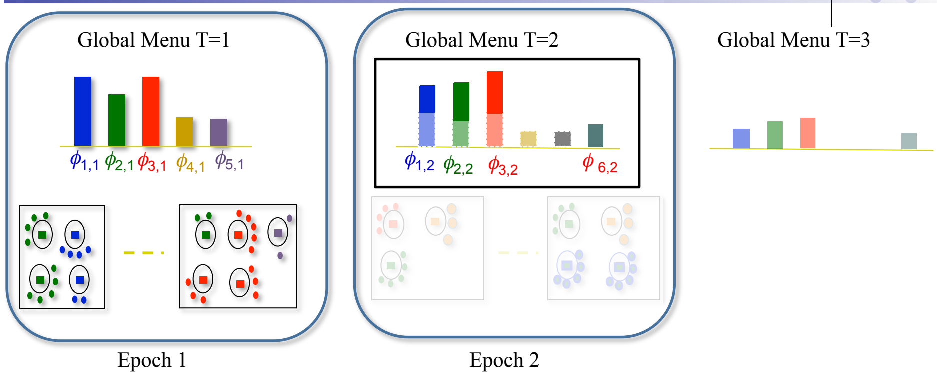
Generative Process

-For customer w in restaurant 1
- -[as in static case] Choose table $j \propto N_j$
- - Choose a new table $b \propto \alpha$
  - -Sample a new dish for this table
  - - Existing and inherited dish $k \propto m`_{k,2} + m_{k,2}$
  - - Existing but <u>NOT inherited</u> dish $k \propto m`_{k,2}$ <u>Then</u> $\phi_{k,2} \sim \text{Normal}(.| \phi_{k,1},\rho)$
  - - A new dish $\propto \gamma$ <u>Then</u> $\phi_{new} \sim H$

The "Evolving" Chinese Restaurant Franchise Process

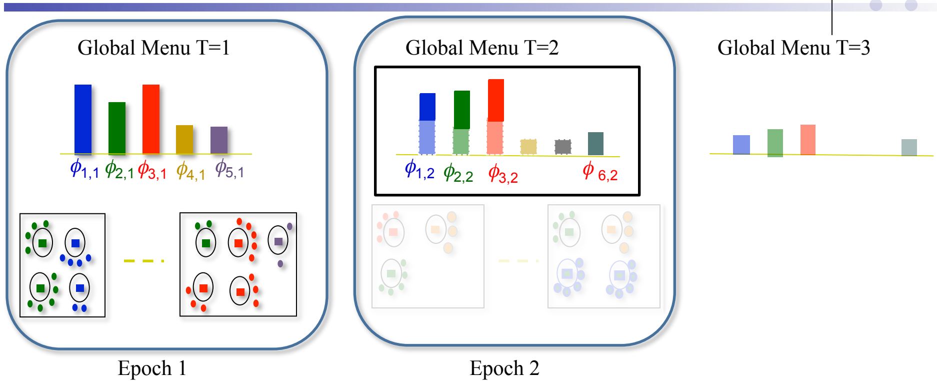


The “Evolving” Chinese Restaurant Franchise Process



- Topics’ trends evolve over time? ✔
- Topics’ distributions evolve over time? ✔
- Number of topics grow with the data? ✔

The “Evolving” Chinese Restaurant Franchise Process



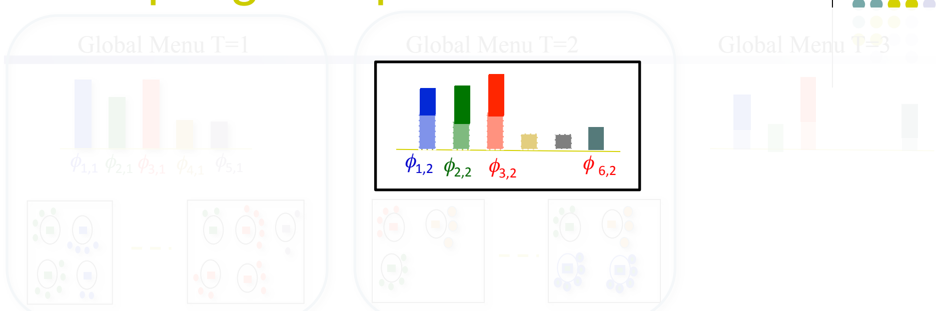
-We just described a first order RCRF process
- for a general Δ-order process

$$m'_{kt} = \sum_{\delta=1}^{\Delta} \exp^{\frac{-\delta}{\lambda}} m_{k,t-\delta}$$

Inference

- Gibbs Sampling
  - Sample a table for each word
  - Sample a topic for each table
  - Sample the topic parameter over time
  - Sample hyper-parameters
- How to deal with non-conjugacy
  - Algorithm 8 in Neal's 1998 + Metropolis-Hasting
- Efficiency
  - The Markov blanket contains the previous and following Δ epochs

Sampling a Topic for a Table

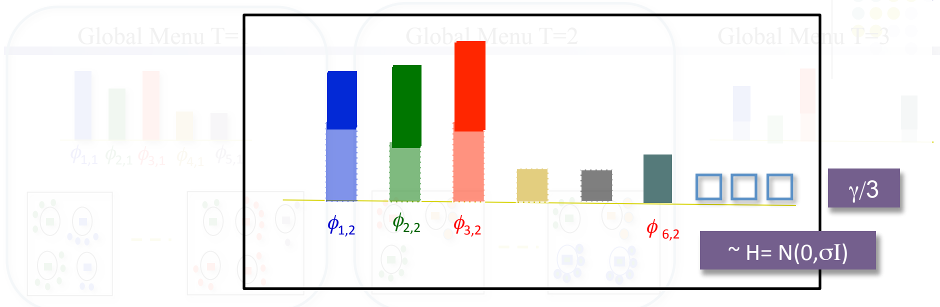


$$\mathrm{P}(k_{tdb} = k | \boldsymbol{k}^{-tdb}_{t-\Delta:t+\Delta}, \boldsymbol{b}_{td}, \boldsymbol{\phi}, \boldsymbol{w}_t) \propto$$

$$\mathrm{P}(k_{tdb} = k | \boldsymbol{k}^{-tdb}_{t-\Delta:t}) \, \mathrm{P}(\boldsymbol{v}_{tdb} | \boldsymbol{\phi}, k_{tdb} = k) \prod_{\delta=1}^{\Delta} \mathrm{P}(\boldsymbol{k}_{t+\delta} | \boldsymbol{k}^{-tdb \to k}_{t+\delta-\Delta:t+\delta-1})$$

Past	Emission	Future
	Non-Conjugacy	Efficiency

Sampling a Topic for a Table



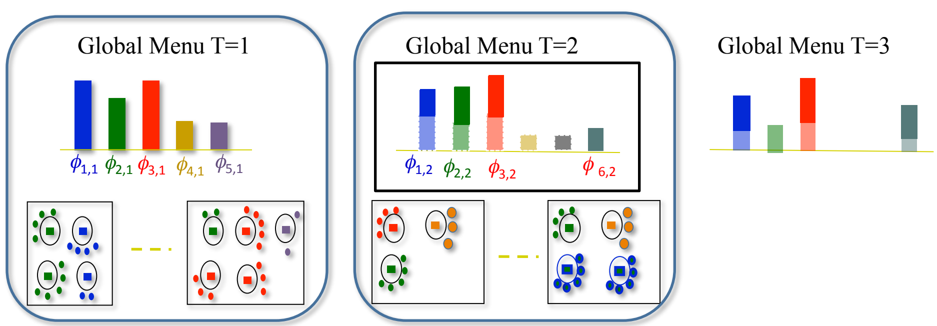
$$P(k_{tdb}=k|\boldsymbol{k}^{-tdb}_{t-\Delta:t+\Delta}, \boldsymbol{b}_{td}, \boldsymbol{\phi}, \boldsymbol{w}_t) \propto$$

$$P(k_{tdb}=k|\boldsymbol{k}^{-tdb}_{t-\Delta:t})\, P(\boldsymbol{v}_{tdb}|\boldsymbol{\phi}, k_{tdb}=k) \prod_{\delta=1}^{\Delta} P(\boldsymbol{k}_{t+\delta}|\boldsymbol{k}^{-tdb\to k}_{t+\delta-\Delta:t+\delta-1})$$

Past | Emission | Future

Non-Conjugacy | Efficiency

Sampling a Topic for a Table



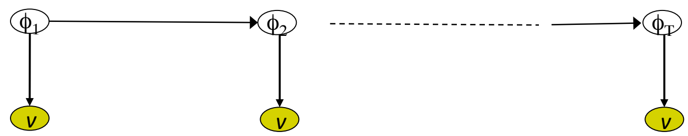
$$P(k_{tdb}=k|\boldsymbol{k}^{-tdb}_{t-\Delta:t+\Delta}, \boldsymbol{b}_{td}, \boldsymbol{\phi}, \boldsymbol{w}_t) \propto$$

$$P(k_{tdb}=k|\boldsymbol{k}^{-tdb}_{t-\Delta:t})\, P(\boldsymbol{v}_{tdb}|\boldsymbol{\phi}, k_{tdb}=k) \prod_{\delta=1}^{\Delta} P(\boldsymbol{k}_{t+\delta}|\boldsymbol{k}^{-tdb\to k}_{t+\delta-\Delta:t+\delta-1})$$

Past | Emission | Future

Non-Conjugacy | Pre-compute And update

Sampling Topic Parameters

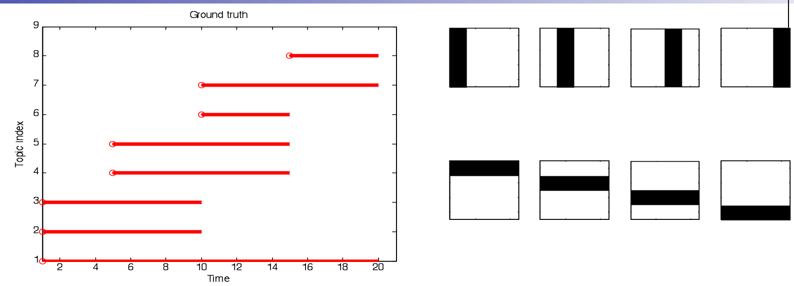


- $V|\phi \sim$ Mult(Logistic(ϕ))
- Linear-State space model with non-Gaussian emission
- Use Laplace approximation inside the Forward-Backward algorithm
- Use the resulting distribution as a proposal

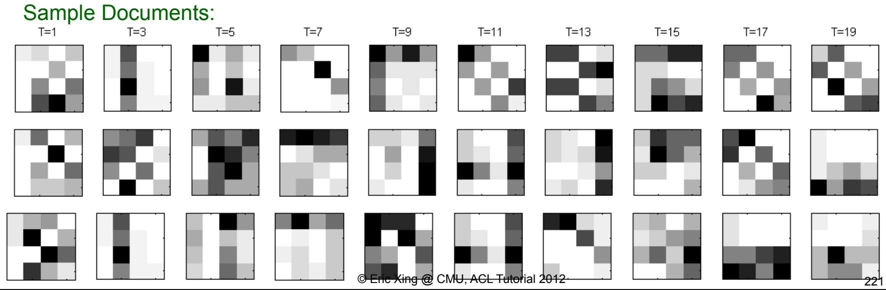
Experiments

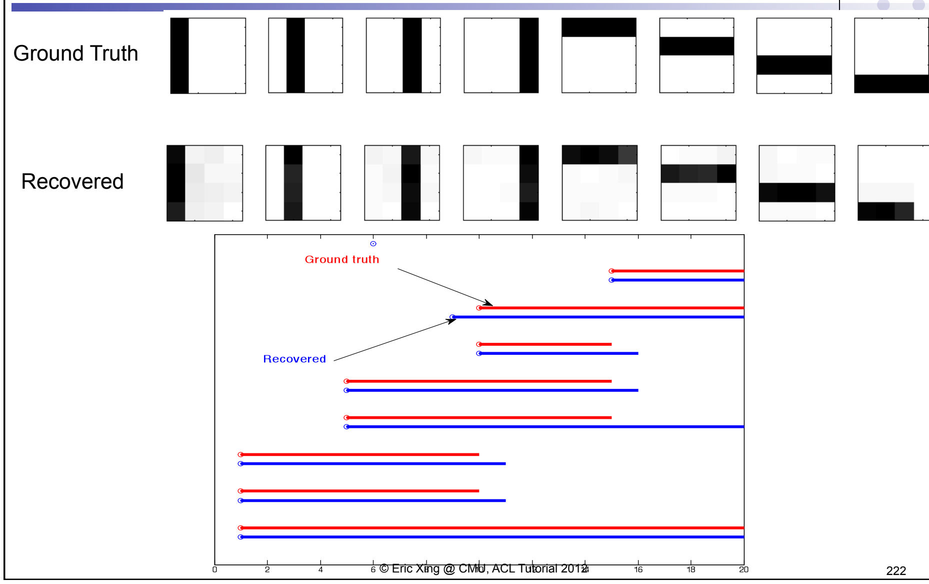
- Simulated data
  - Simulated 20 epochs with 100 data points in each epoch
- Timeline of the NIPS conference
  - 13 years
  - 1740 documents
  - 950 words per document
  - ~3500 vocabulary

Simulation Experiment

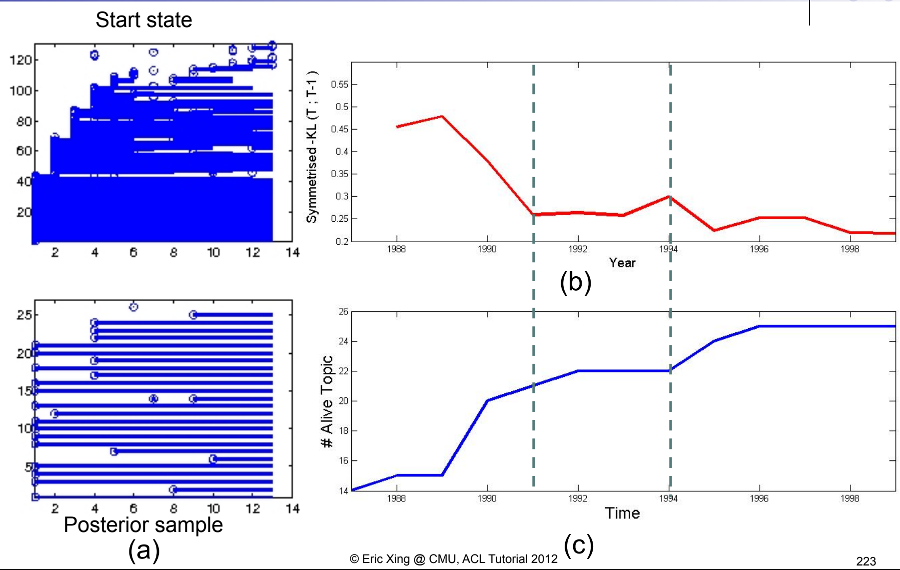


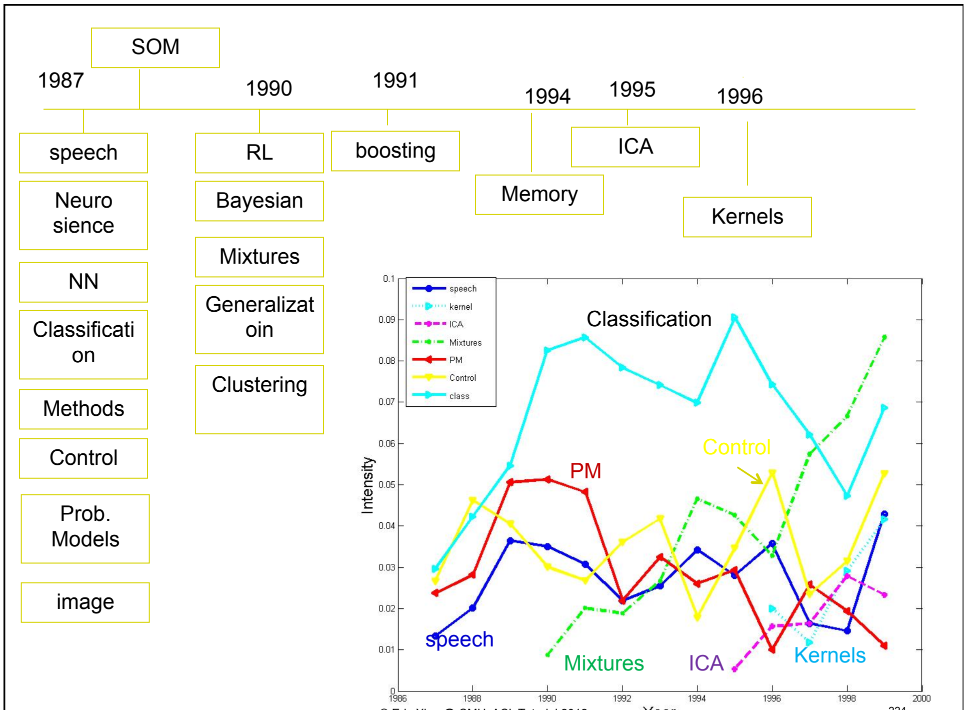
Sample Documents:

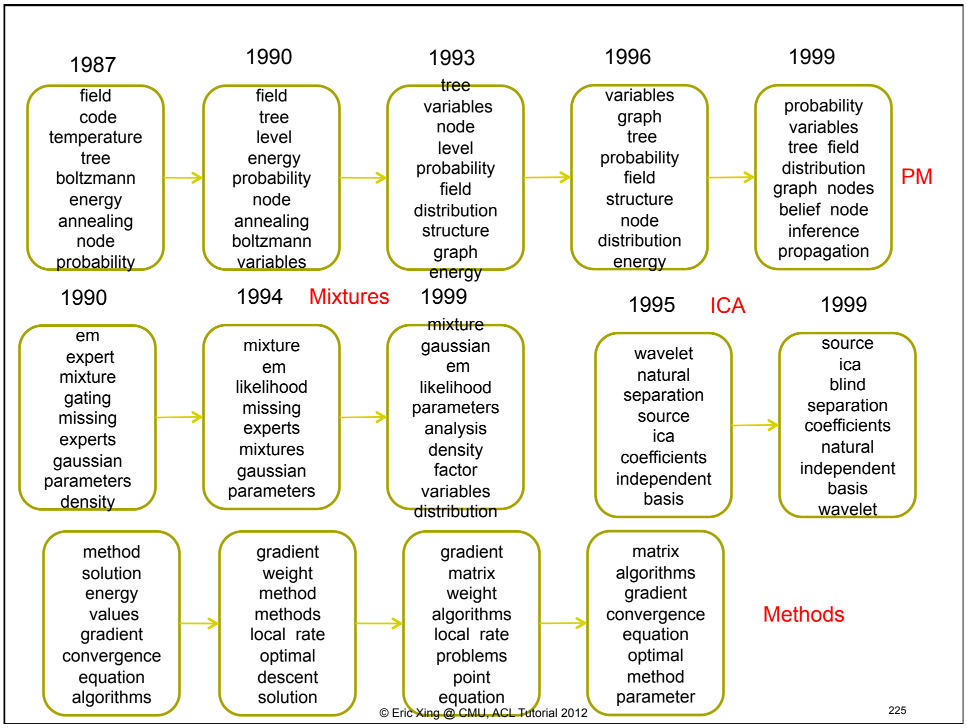


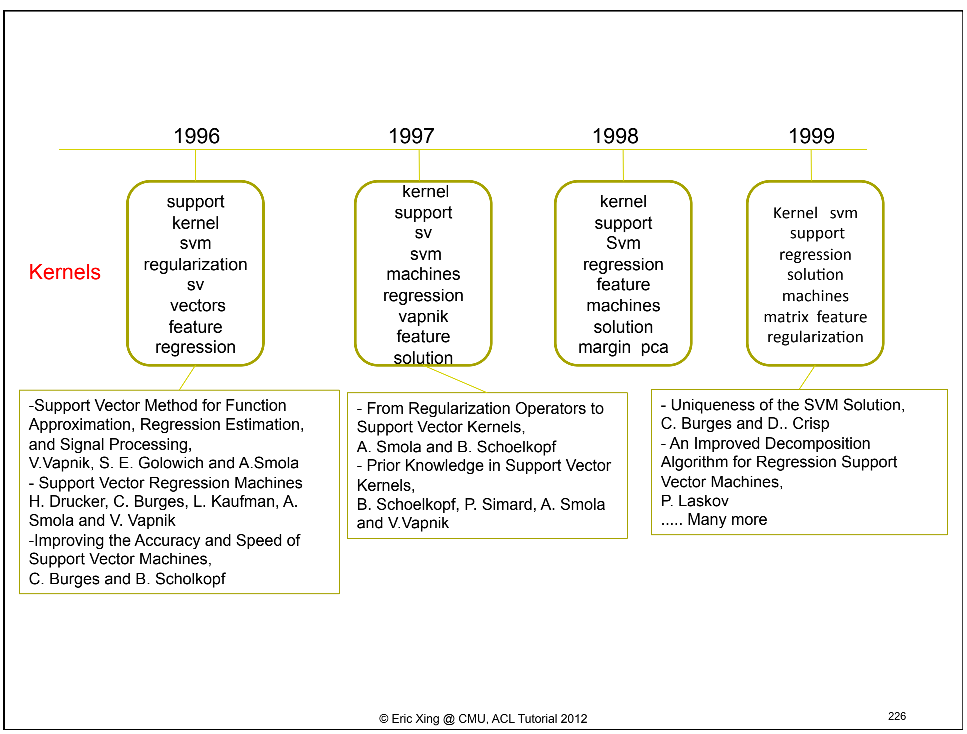


Analyzing the NIPS Corpus

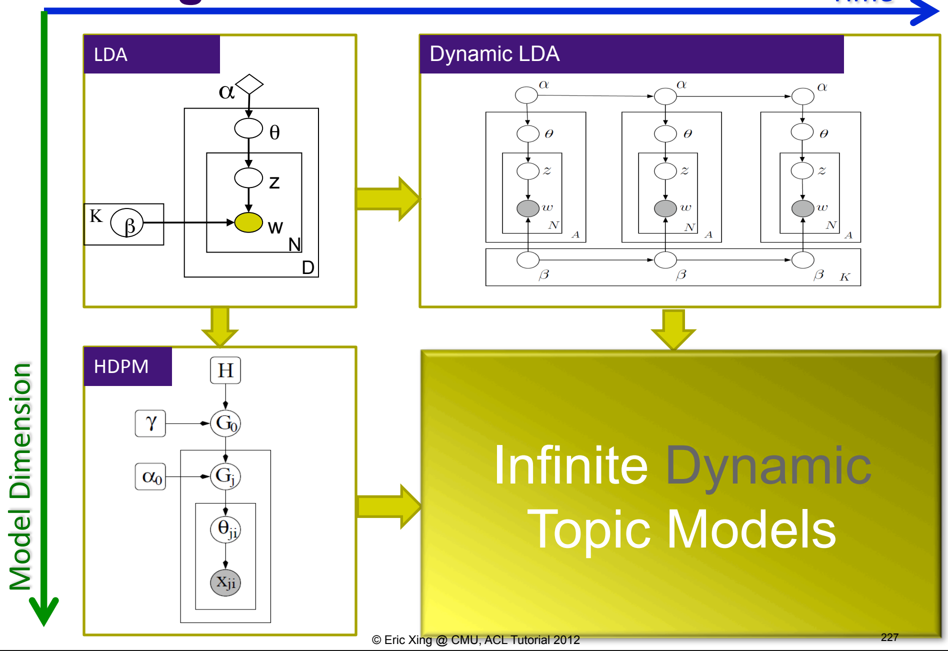





1987
field code temperature tree boltzmann energy annealing node probability
1990
field tree level energy probability node annealing boltzmann variables
1993
tree variables node level probability field distribution structure graph energy
1996
variables graph tree probability field structure node distribution energy
1999
probability variables tree field distribution graph nodes belief node inference propagation
PM
1990
em expert mixture gating missing experts gaussian parameters density
1994
Mixtures
mixture em likelihood missing experts mixtures gaussian parameters
1999
mixture gaussian em likelihood parameters analysis density factor variables distribution
1995
ICA
wavelet natural separation source ica coefficients independent basis
1999
source ica blind separation coefficients natural independent basis wavelet
method solution energy values gradient convergence equation algorithms
gradient weight method methods local rate optimal descent solution
gradient matrix weight algorithms local rate problems point equation
matrix algorithms gradient convergence equation optimal method parameter
Methods
© Eric Xing @ CMU, ACL Tutorial 2012
225

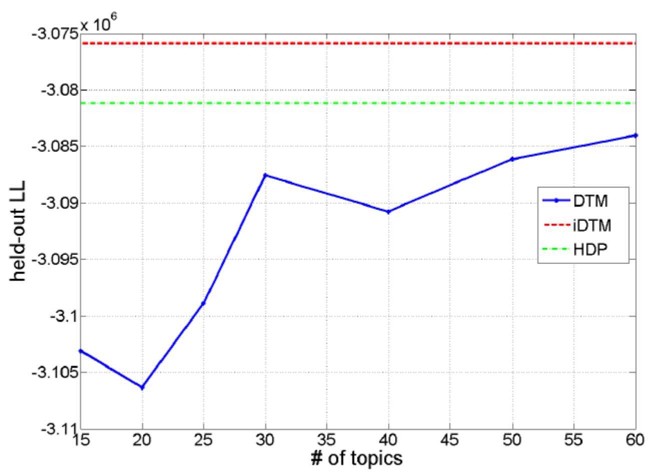

1996
1997
1998
1999
Kernels
support kernel svm regularization sv vectors feature regression
kernel support sv svm machines regression vapnik feature solution
kernel support Svm regression feature machines solution margin pca
Kernel svm support regression solution machines matrix feature regularization
-Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing,
V.Vapnik, S. E. Golowich and A.Smola
- Support Vector Regression Machines
H. Drucker, C. Burges, L. Kaufman, A. Smola and V. Vapnik
-Improving the Accuracy and Speed of Support Vector Machines,
C. Burges and B. Scholkopf
- From Regularization Operators to Support Vector Kernels,
A. Smola and B. Schoelkopf
- Prior Knowledge in Support Vector Kernels,
B. Schoelkopf, P. Simard, A. Smola and V.Vapnik
- Uniqueness of the SVM Solution,
C. Burges and D.. Crisp
- An Improved Decomposition Algorithm for Regression Support Vector Machines,
P. Laskov
..... Many more
© Eric Xing @ CMU, ACL Tutorial 2012
226

The Big Picture



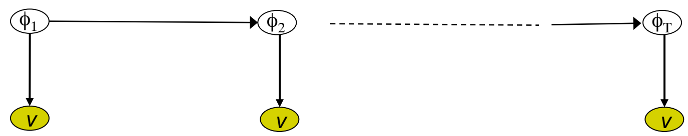
© Eric Xing @ CMU, ACL Tutorial 2012 227

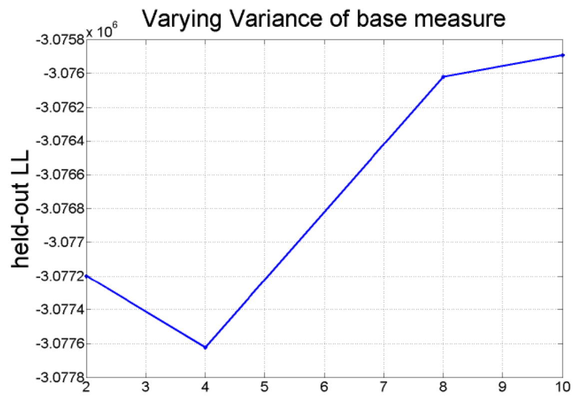
Quantitative Analysis



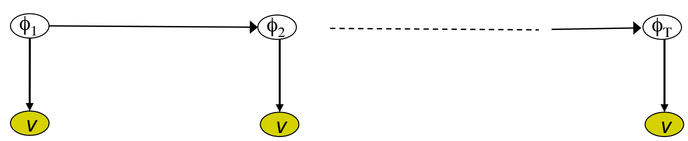
© Eric Xing @ CMU, ACL Tutorial 2012 228

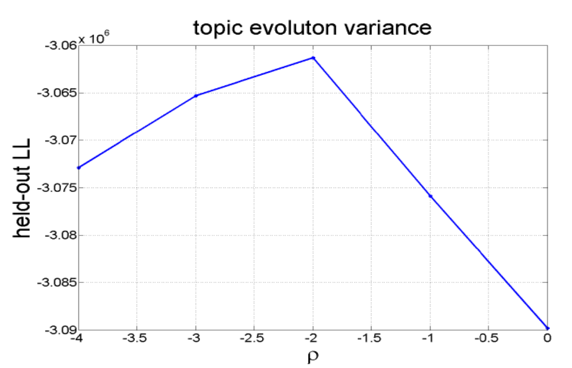
Hyper-parameter Sensitivity





Hyper-parameter Sensitivity



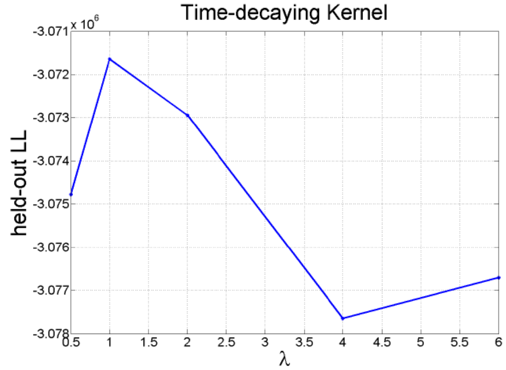


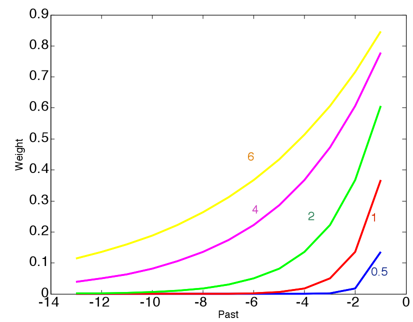
Hyper-parameter Sensitivity

$$m'_{kt} = \sum_{\delta=1}^{\Delta} \exp^{\frac{-\delta}{\lambda}} m_{k,t-\delta}$$

Global Menu T=3

Time-decaying Kernel





8. Algorithmic Scalability and Large-Scale Learning

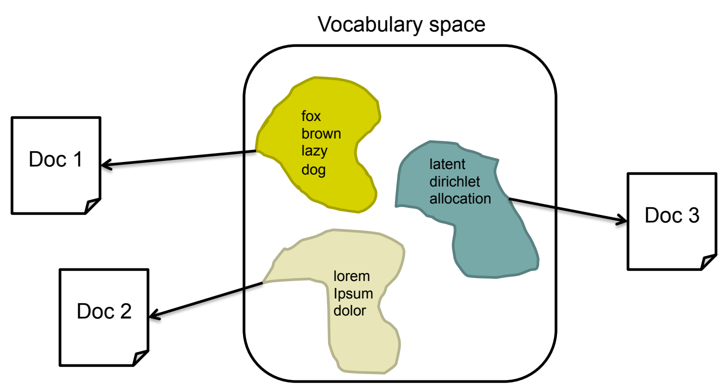
Scaling topic models to large document collections

- Large-scale corpora have millions, even billions of documents, with vocabulary sizes in the millions
  - Runtime and memory challenges!

- Scaling to such corpora requires techniques such as:
  1. Efficient data representations and algorithms
  2. Parallelization over multiple CPUs
  3. Online inference, to handle incoming documents one-at-a-time

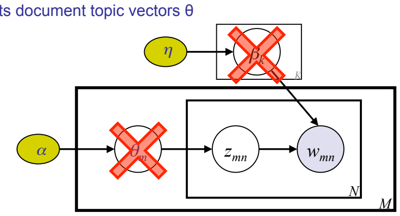
Efficient data representations and algorithms

- Key observation: most documents contain just a small fraction of the words in the vocabulary
  - We say that the documents are sparse



Efficient data representations and algorithms

- Collapsed Gibbs Sampling is a very popular inference algorithm for topic models
  - CGS samples just the word-topic indicators z, without having to sample document topic vectors θ or topic vocabularies β
  - CGS requires us to track of two types of counts:
    - Topic-word: For each topic k, # of times it is assigned to vocabulary word v
      - i.e. for all documents m and words n, # of z_{mn} s.t. z_{mn} = k and w_{mn} = v
      - Represents topic vocabularies β
    - Document-topic: In each document m, # of words n assigned to topic k
      - i.e. for all words n in document m, # of z_{mn} s.t. z_{mn} = k
      - Represents document topic vectors θ



Efficient data representations and algorithms

- Store both word-topic and document-topic counts using a dictionary (key-value) data structure
  - Take advantage of sparsity to save memory!
  - Savings can be very large:
    - e.g. you have 500 topics but each document uses only 5 on average
    - e.g. you have 1 million vocabulary words, but each topic uses only 10,000 on average

Document-topic counts for document m

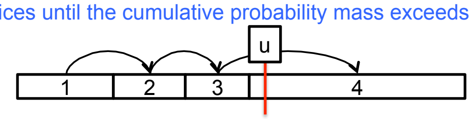
Topic 1: 5 words
Topic 6: 3 words
Topic 8: 1 word
…

Topic-word counts for topic k

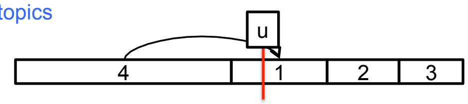
dog: 10 occurrences
cat: 15 occurrences
mouse: 3 occurrences
…

Efficient data representations and algorithms

- It's not just about saving memory
- We can speed up Collapsed Gibbs sampling by exploiting sparsity in word-topic and document-topic counts
  - Notice that each word-topic indicator z follows a discrete distribution over K topics
  - We sample z by drawing u ~ Uniform(0,1), and then iterating through each of the K topic choices until the cumulative probability mass exceeds u



  - If we consider the topic choices with the largest probability mass first, we'll stop after fewer topics



  - See Yao, Mimno and McCallum (2009) for details

Efficient data representations and algorithms

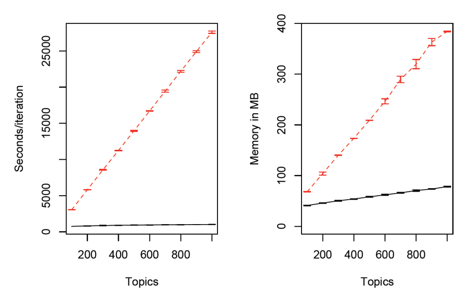


Figure 2: A comparison of time and space efficiency between standard Gibbs sampling (dashed red lines) and the SparseLDA algorithm and data structure presented in this paper (solid black lines). Error bars show the standard deviation over five runs.

Efficient Methods for Topic Model Inference on Streaming Document Collections (Yao, Mimno and McCallum 2009)

Efficient data representations and algorithms

- What about variational inference?
- We can apply stochastic gradient descent to speed up variational updates
  - Instead of computing the full gradient, just subsample terms from the gradient!
  - Specifically, we subsample random documents, and then keep gradient terms belonging to those documents
    - This works because the topic model log-likelihood decomposes as a sum over documents!
  - Often, we can obtain good performance with just a small fraction of the terms

Parallelization over multiple CPUs

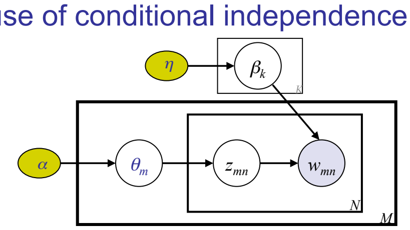
- Efficient data/algorithms can only get you so far on one CPU
  - The processor industry is no longer focused on single-CPU performance
    - In 5 years, new processors will not be much faster than today's processors
    - But they will have many more CPU cores for parallel programming!
  - On the other hand, text corpora are growing rapidly
    - English Wikipedia has nearly 4 million articles
    - Blogosphere generates 900 thousand posts per day in 2008, and almost certainly more today (Source: Technorati)
  - We must parallelize – both now, and in the future as well

Parallelization over multiple CPUs

- Three common strategies for parallel inference:
  - Apply “standard” variational inference (VB) or Gibbs sampling, but distribute the documents over the CPUs
    - Advantages: easy extension to standard variational/MCMC algorithms
    - Drawbacks: convergence no longer guaranteed under some situations, like Collapsed Gibbs sampling
  - Use particle filtering (aka Sequential Monte Carlo sampling) with P particles, and split the particles over the CPUs
    - Advantages: convergence is always guaranteed; can pick any number of particles P
    - Drawbacks: naïve sampler can lead to very poor results, thus care is needed in designing the sampler
  - Use Auxiliary variables to distribute inference
    - Advantages: Convergence guaranteed, collapsed Gibbs sampling on individual CPUs.
    - Drawbacks: Latency introduced by network if entire data cannot reside on a single machine.

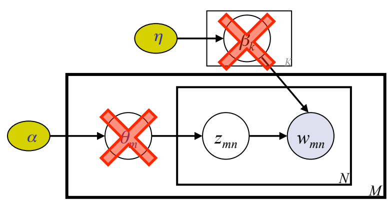
Distributing documents

- By distributing documents over CPUs, we can:
  - Infer the word-topic indicators z and document topic vectors in θ in parallel (conditioned on the topic vocabularies β)
  - Easily implemented as a map operation
- To infer the topic vocabularies β:
  - Consolidate statistics from the word-topic indicators z into one CPU
  - Have that CPU infer the β’s (conditioned on z’s)
  - Easily implemented as a map-reduce operation
- Works because of conditional independence in the model!



Distributing documents

- In Collapsed Gibbs sampling however, the conditional independence assumptions break down
  - Integration of θ, β makes the z's depend on each other
  - MCMC convergence guarantees no longer hold when parallel sampling the z's
- In practice however, parallel CGS sampling does produce good results (Asuncion, Smyth and Welling 2008)
  - Though this may or may not generalize to more complex topic models



Distributing documents

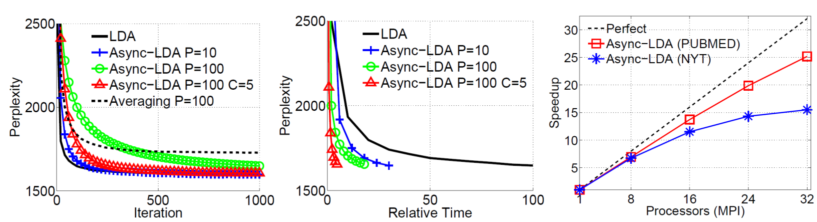


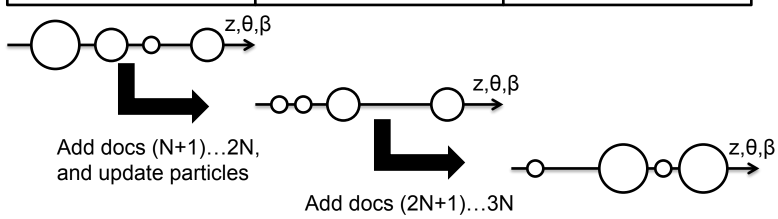
Figure 3: (a) Left: Convergence plot for Async-LDA on KOS, K=16. (b) Middle: Same plot with x-axis as relative time. (c) Right: Speedup results for NYT and PUBMED on a cluster, using Message Passing Interface.

Asynchronous Distributed Learning of Topic Models (Asuncion, Smyth and Welling 2008)

Sequential Monte Carlo

- Under SMC (particle filtering), we "evolve" the posterior distribution one set of documents at a time
  - Represent the posterior as a set of weighted samples (called "particles")
    - Within a set of documents, particles are evolved according to some proposal distribution
    - Each particle can be evolved independently in parallel
  - In the illustration below, each circle is a particle. Axis locations represent latent variable values, and sizes represent particle weights.

Documents 1…N	Documents (N+1)…2N	Documents (2N+1)…3N



Sequential Monte Carlo

Table 1: Details of Yahoo! News dataset and corresponding clustering accuracies of the baseline (LSHC) and our method (Story), $K = 100$.

Sample No.	Sample size	Num Words	Num Entities	Story Acc.	LSHC Acc.
1	111,732	19,218	12,475	**0.8289**	0.738
2	274,969	29,604	21,797	**0.8388**	0.791
3	547,057	40,576	32,637	**0.8395**	0.800

SMC inference performs well on real world datasets

Table 5: Number of particles, sample-1, $K = 100$.

#Particles	4	8	16	32	50
Accuracy	0.8101	08289	0.8299	0.8308	0.8358

Increasing the number of particles improves performance

Online Inference for the Infinite Topic-Cluster Model: Storylines from Streaming Text (Ahmed, Ho, Teo, Eisenstein, Smola, Xing 2011)

Online Inference

- Often, we want to add new documents to the model incrementally
  - But we can't afford to rerun inference all documents, especially for huge corpora!
  - How can we insert the new documents in a statistically principled manner?

- Online inference allows us to incorporate the influence of new documents
  - Sequential Monte Carlo (already explained)
  - Online variational inference

Online Variational Inference

- Key idea: split the set of docs into smaller "minibatches", similar to Sequential Monte Carlo

  - In each minibatch:
    - Perform variational inference on word-topic indicators z and document topic vectors θ, for all docs in the minibatch
    - Perform gradient steps on the topic vocabularies β, using only terms corresponding to docs in the minibatch

  - The use of minibatches is equivalent to stochastic gradient updates, which are guaranteed to converge (Hoffman, Blei and Bach 2010)

- We can process docs as they arrive, one minibatch at a time

Online Variational Inference

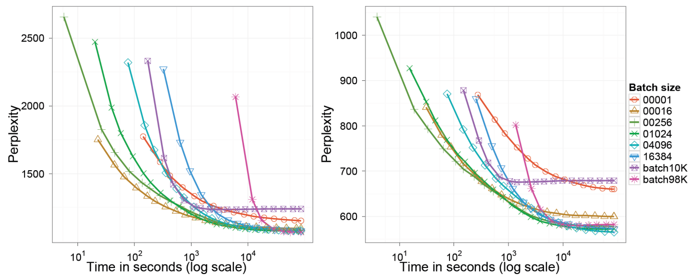


Figure 2: Held-out perplexity obtained on the *Nature* (left) and Wikipedia (right) corpora as a function of CPU time. For moderately large mini-batch sizes, online LDA finds solutions as good as those that the batch LDA finds, but with much less computation. When fit to a 10,000-document subset of the training corpus batch LDA's speed improves, but its performance suffers.

Online Learning for Latent Dirichlet Allocation
(Hoffman, Blei and Bach 2009)

Auxiliary Variable Representation

- Key Idea "Dirichlet mixtures of Dirichlet processes are Dirichlet processes"
- We can re-write the generative process of DPMM as:-

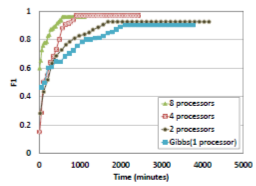
$$D_j \sim DP\left(\tfrac{\alpha}{P}, H\right), \quad \phi \sim Dirichlet\left(\tfrac{\alpha}{P}, \ldots, \tfrac{\alpha}{P}\right), \quad \pi_i \sim \phi, \quad \theta_i \sim D_{\pi_i}, \quad x_i \sim f(\theta_i)$$
$$j = 1, \ldots, P \textit{ and } i = 1, \ldots, N.$$

- By adding additional constrain on the concentration parameter of bottom level DP we can re-write HDP as:-

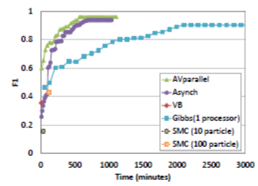
$$\zeta_j \sim \text{Gamma}\left(\frac{\alpha}{P}\right), \quad j = 1, \ldots, P$$
$$D_{0j} \sim \text{DP}\left(\frac{\alpha}{P}\right), \quad j = 1, \ldots, P$$
$$\nu_m \sim \text{Dirichlet}(\zeta_1, \ldots, \zeta_P)$$

$$D_{mj} \sim \text{DP}(\zeta_j, D_{0j}), \quad m = 1, \ldots, M, \; j = 1, \ldots, P$$
$$\pi_{mi} \sim \nu_m, \quad m = 1, \ldots, M, \; i = 1, \ldots, N_m$$
$$\theta_{mi} \sim D_{m\pi_{mi}}$$
$$x_{mi} \sim f(\theta_{mi})$$

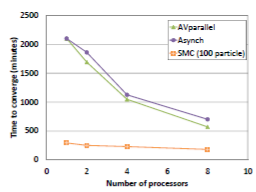
Auxiliary Variable Representation



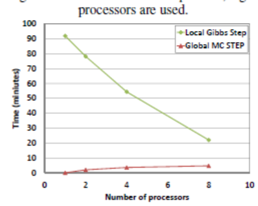
(a) F1 score against run time for AVparallel



(b) F1 score against run time for various algorithms. Unless otherwise specified, eight processors are used.



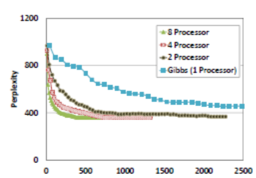
(c) Time taken to reach convergence (< 0.1% chaange in F1)



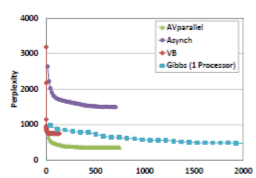
(d) Time spent in global and local steps for AVparallel, over 500 iterations.

Figure 1: Synthetic data modeled using a DPMM.

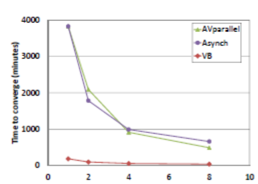
Auxiliary Variable Representation



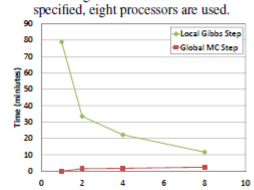
(a) Test set perplexity against run time for AVparallel



(b) Test set perplexity against run time for various algorithms. Unless otherwise specified, eight processors are used.



(c) Time taken to reach convergence (< 0.1% change in perplexity)



(d) Time spent in global and local steps for AVparallel, over 20 iterations.

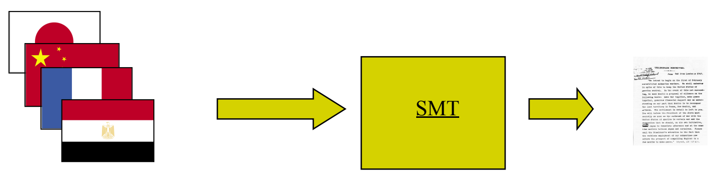
Figure 2: NIPS corpus, modeled using an HDP.

Algorithmic Scalability and Large-Scale learning

- Data structures and algorithms matter
  - Dictionaries to exploit vocabulary/topic sparsity
  - Faster sampling by reordering topics
- Parallelization is great, but one needs to be careful
  - Multiple parallel inference approaches, with their own pros/cons
  - Conditional independence allows us to divide documents among processors
  - Auxiliary variables can provide conditional independence in collapsed samplers
- Online inference is possible
  - For the same reasons that parallelization is possible
  - Use incoming document minibatches to update topic vocabulary distributions

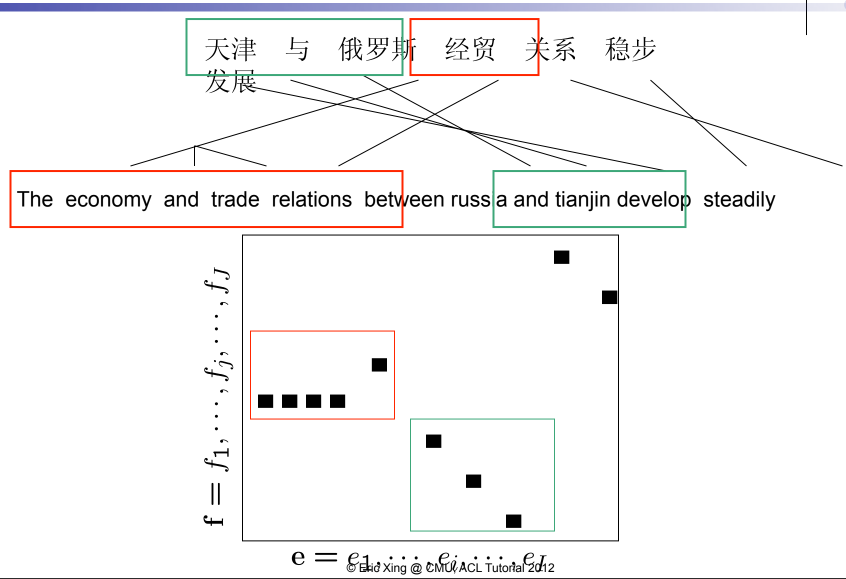
9: Other apps (Optional)

I. Machine translation

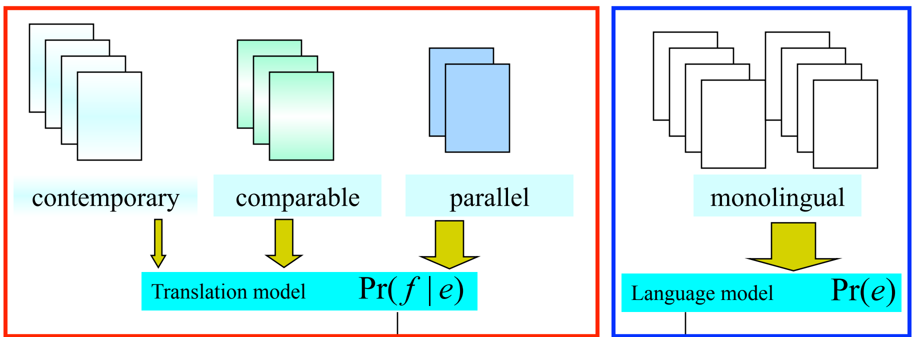


B. Zhao and E.P Xing, ACL 2006

Word Alignment



The Statistical Formulation



$$\hat{a} = \arg\max_{a} \Pr(f \mid e, a) \Pr(e)$$

BiTAM: From monolingual to bilingual topic models (Zhao & Xing, ACL/Coling 2006)

- Monolingual space, a unigram LM $p(w|z)$
  - A topic corresponding to a point in the *word simplex*.
  - *AdMixture of unigrams* (Blei, et al. 2003)
- Bilingual space, a translation lexicon $p(f|e, z)$
  - Given a topic z, a word usually has limited translations.
  - Topic-specific translation lexicons are sharper
  - Each topic is a point in the *conditional simplex*
  - *AdMixture of topic-specific translation lexicons* (Zhao & Xing, ACL/Coling 2006)
- Example
  - A Chinese word "**club**", the translations can be:

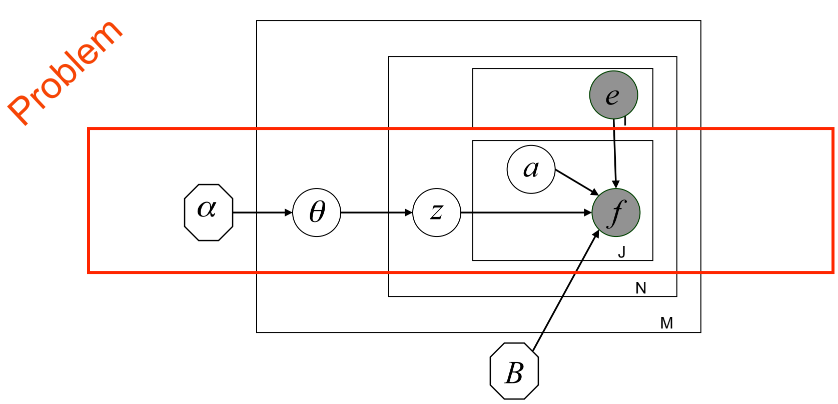
ogre	war	socialize	interests
0.4	**0.5**	**0.0**	**0.1**
0.0	0.1	0.5	0.4

BiTAM: A Generative Process

- Sample topic weights θ from a Dirichlet(α)
- Sample a topic z from multinomial (θ)
- For each word f in the sentence $\vec{f}$
  - Sample an alignment a from **an alignment model**
  - Generate f with word e_a from a **topic-specific lexicon**

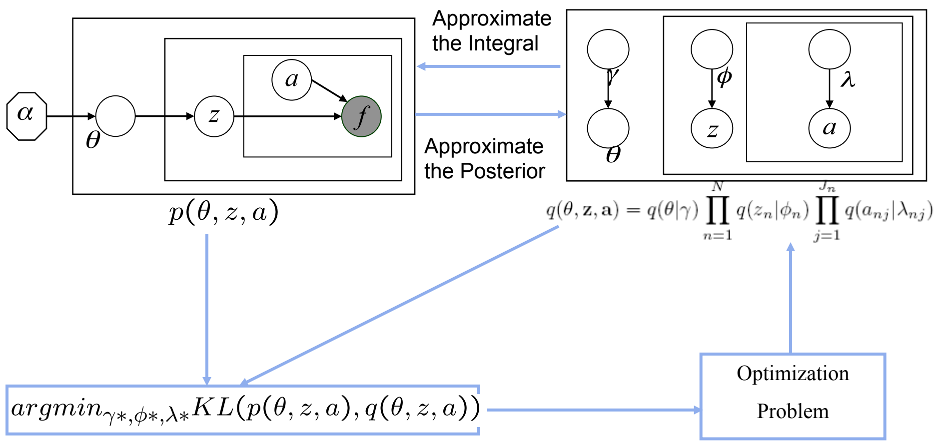
BiTAM Model-1

- Graphical Model (a language to encode dependencies)

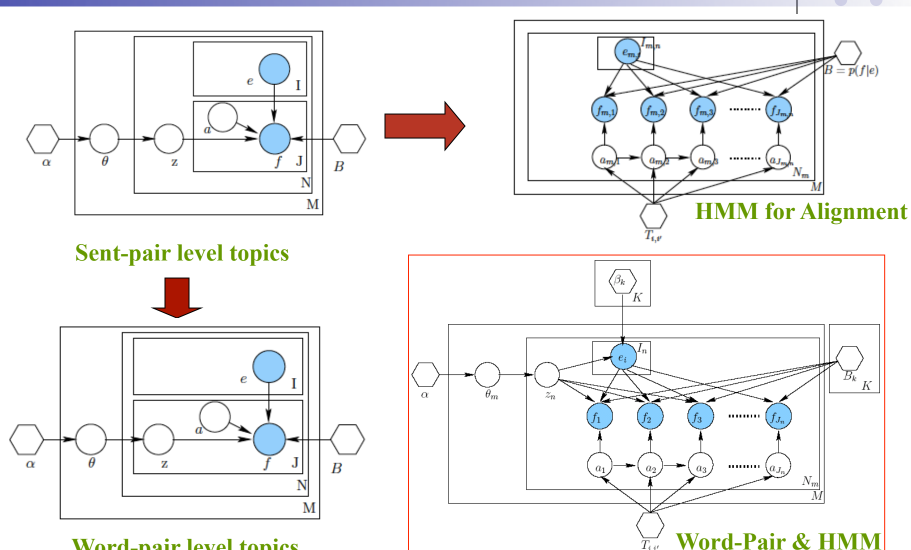


$$p(F \mid A, E, \alpha, B) = \int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(f_n \mid a_n, e_n, B_{z_n}) d\theta$$

GMF Inference



An upgrade path for BiTAMs



Experiments

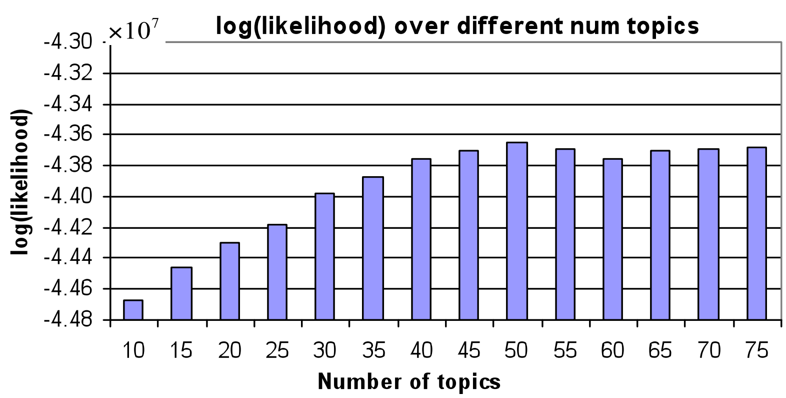
- Training data
  - Small: Treebank 316 doc-pairs (133K English words)
  - Large: FBIS-Beijing, Sinorama, XinHuaNews, (15M English words).

Train	#Doc.	#Sent.	#Tokens English	#Tokens Chinese
Treebank	316	4172	133K	105K
FBIS.BJ	6,111	105K	4.18M	3.54M
Sinorama	2,373	103K	3.81M	3.60M
XinHua	19,140	115K	3.85M	3.93M
FOUO	15,478	368K	13.14M	11.93M
Test	95	627	25,500	19,726

- Word Alignment Accuracy & Translation Quality
  - F-measure
  - BLEU

Model Selection

- Choosing num-topics K
  - 10-fold cross-validation
  - Number of topics is set to be 50 for 23 million words corpus

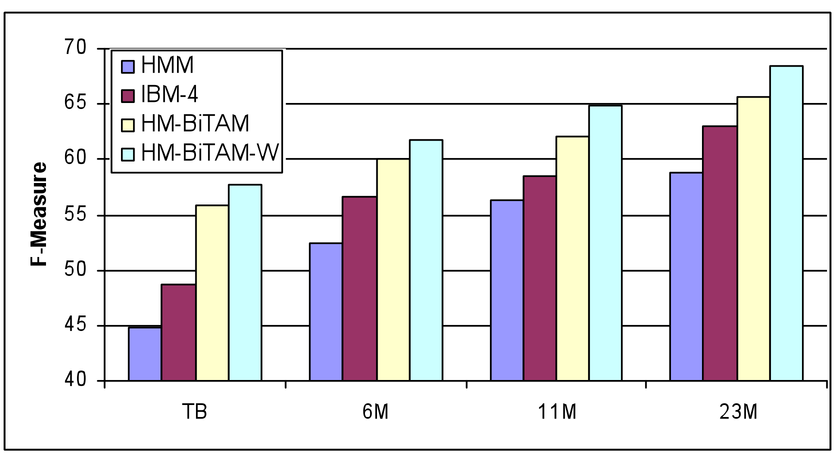


Topics

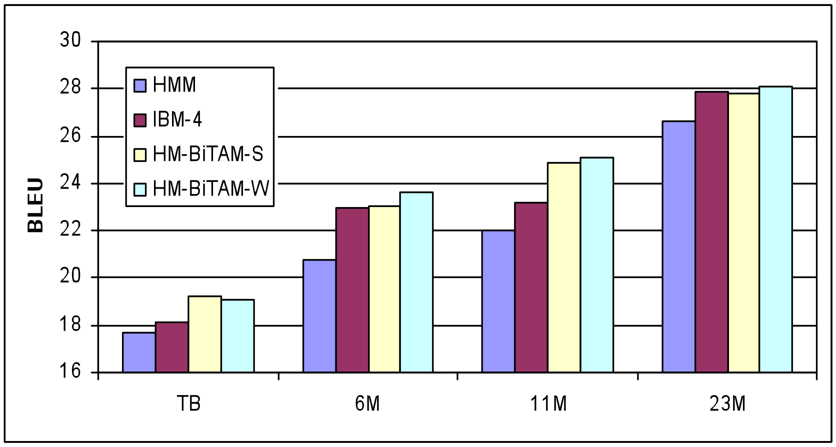
T1	Teams, sports, disabled, games members, people, cause, water, national, handicapped
T2	Shenzhen, singapore, hongkong, stock, national, investment, yuan, options, million, dollar
T3	Chongqing, company, takeover, shenzhen, tianjin, city, national, government, project, companies
T4	Hongkong, trade, export, import, foreign, tech., high, 1998, year, technology
T5	House, construction, government, employee, living, provinces, macau, anhui, yuan
T6	Gas, company, energy, usa, russia, france, chongqing, resource, china, economy, oil

T1	人, 残疾, 体育, 事业, 水, 世界, 区, 新华社, 队员, 记者
T2	深圳, 深, 新, 元, 有, 股, 香港, 国有, 外资, 新华社
T3	国家, 重庆, 市, 区, 厂 , 天津, 政府, 项目, 国, 深圳
T4	香港, 贸易, 出口, 外资, 合作, 今年, 项目, 利用, 新, 技术
T5	住房, 房, 九江, 建设, 澳门, 元, 职工, 目前, 国家, 占, 省
T6	公司, 天然气, 两, 国, 美国, 记者, 关系, 俄, 法, 重庆

HM-BiTAM versus others



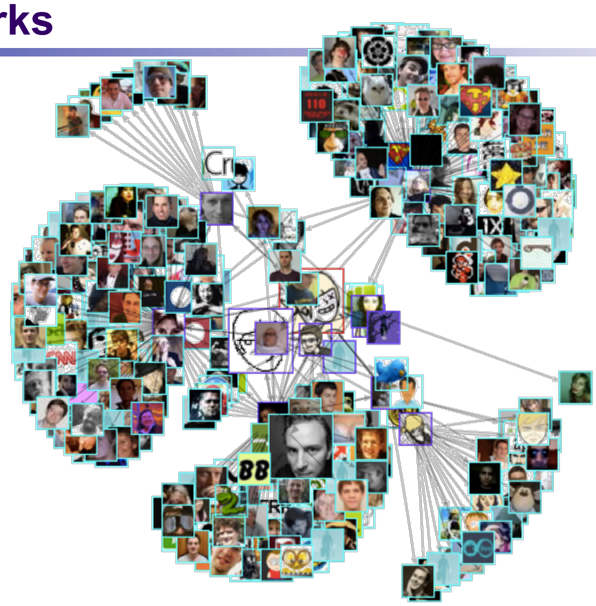
Translation Evaluations



Translation Evaluations

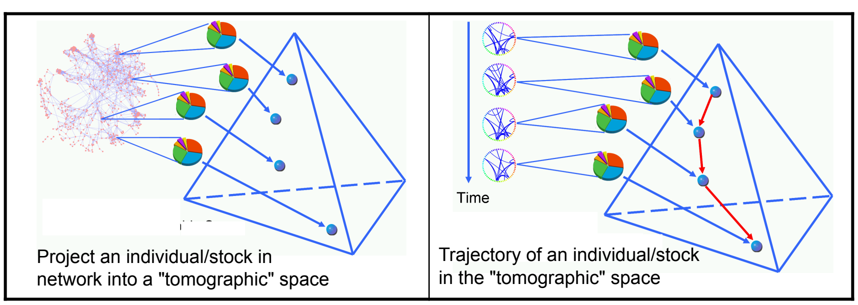
Systems	1-gram	2-gram	3-gram	4-gram	BLEUr4
Hiero Sys.	73.92	40.57	23.21	13.84	30.70
Gale Sys.	75.63	42.71	25.00	14.30	**32.78**
HM-BiTAM	**76.77**	**42.99**	**25.42**	**14.04**	**33.19**
Ground Truth	**76.10**	**43.85**	**26.70**	**15.73**	**34.17**

II. Exploring and deciphering social networks

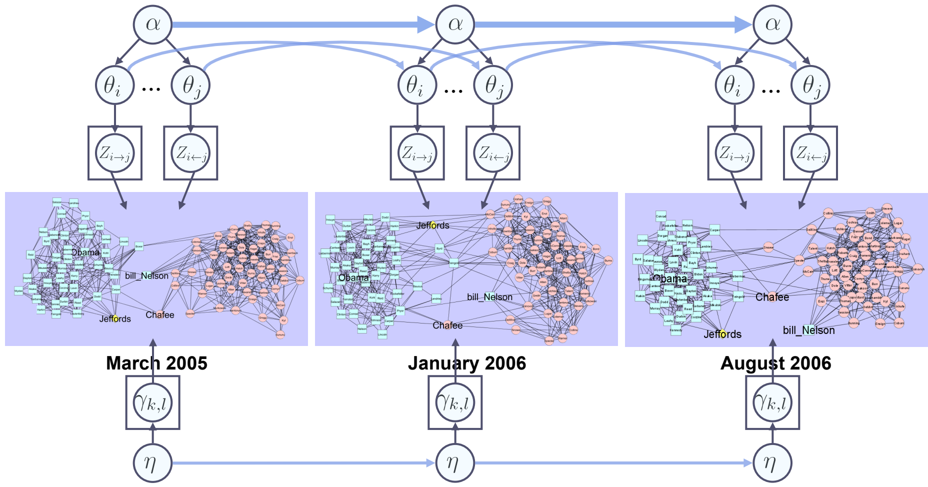


Dynamic network tomography

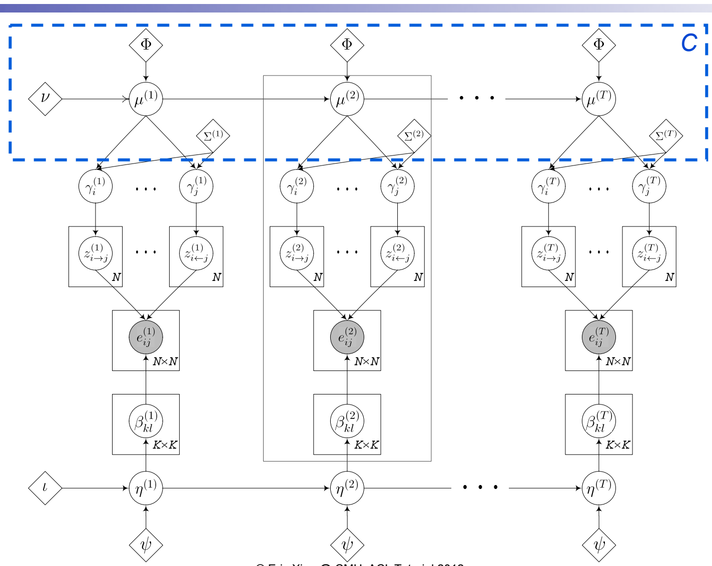
- How to model dynamics in a simplex?



Evolving networks

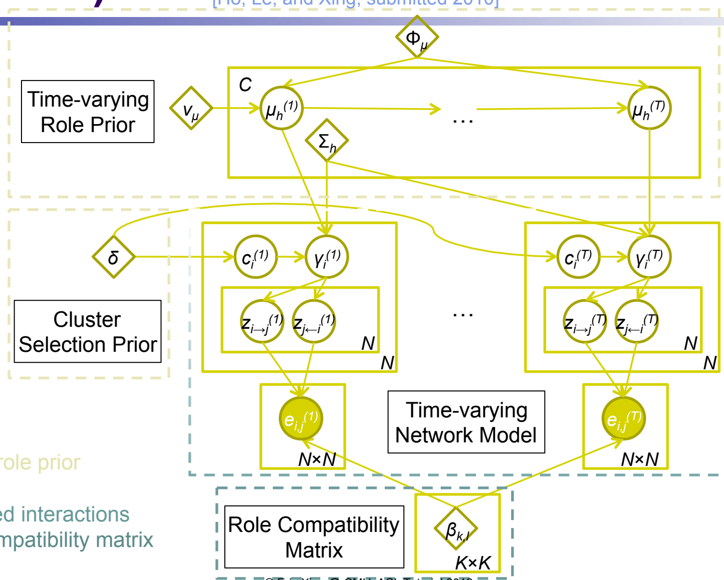


Dynamic MMSB (dMMSB) [Xing, Fu, and Song, AOAS 2009]



Dynamic Mixture of MMSB (dM³SB)

[Ho, Le, and Xing, submitted 2010]



Legend

Hidden role prior

Observed interactions

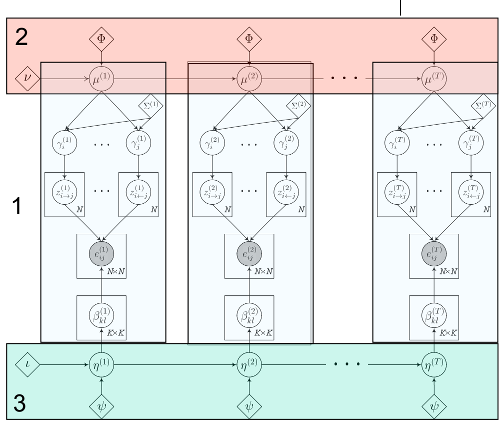
Role compatibility matrix

Algorithm: Generalized Mean Field

(xing et al. 2004)

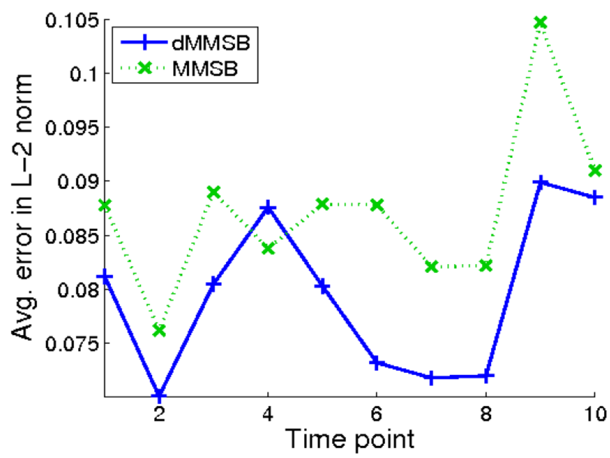
Approximate the joint posterior $p\left(\{\vec{\mathbf{z}}^{(t)}, \vec{\pi}^{(t)}, \vec{\mu}^{(t)}, B^{(t)}\}_{t=1}^{T} | \Theta, \{G^{(t)}\}_{t=1}^{T}\right)$ where Θ denotes the model parameters, by a factored approximate distribution:

$$q\left(\{\vec{\mathbf{z}}^{(t)}, \vec{\pi}^{(t)}, \vec{\mu}^{(t)}, B^{(t)}\}_{t=1}^{T}\right) = q_1\left(\{\vec{\mathbf{z}}^{(t)}, \vec{\pi}^{(t)}\}_{t=1}^{T}\right) \times q_2\left(\{\vec{\mu}^{(t)}\}_{t=1}^{T}\right) \times q_3\left(\{B^{(t)}\}_{t=1}^{T}\right),$$

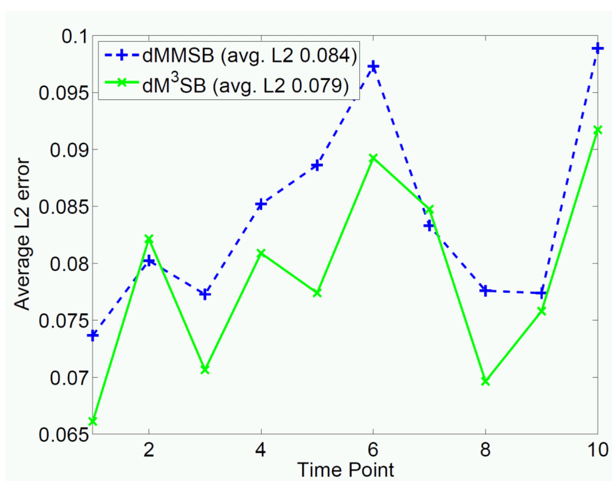


- Inference via variational EM
  - Generalized mean field
  - Laplace approximation
  - Kalman filter & RTS smoother

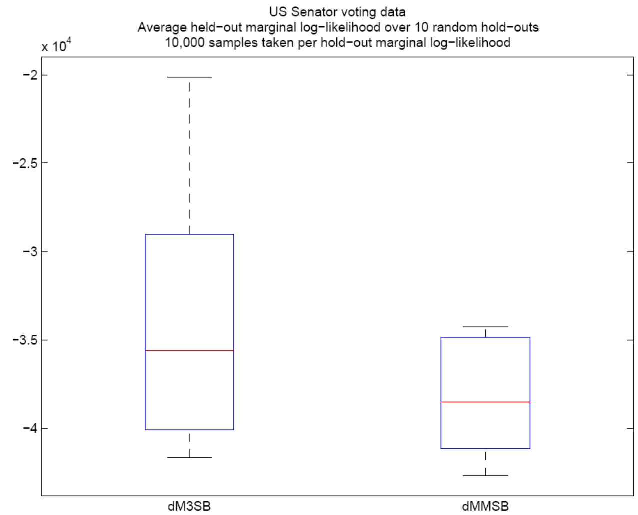
dMMSB vs. MMSB



dM³SB vs. dMMSB

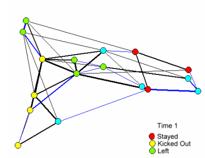


Goodness of fit

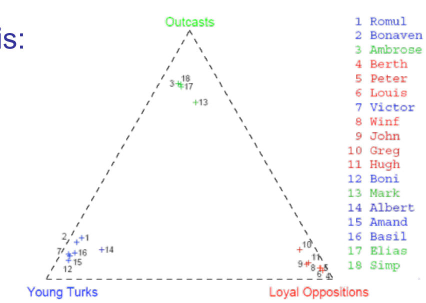


Case Study 1: Sampson's Monk Network

- Dataset Description
  - 18 monks (junior members in a monastery)
  - Liking relations recorded
  - 3 time-points in one year period
  - Timing: before a major conflict outbreak

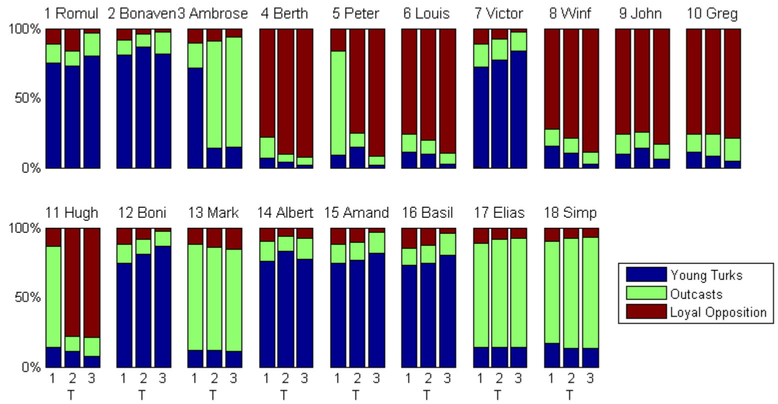


- Recall static analysis:

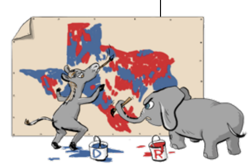


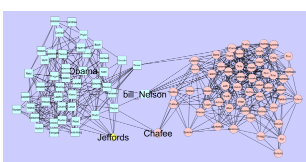
Sampson's Monk Network: role trajectories

- The trajectories of the varying role-vectors over time

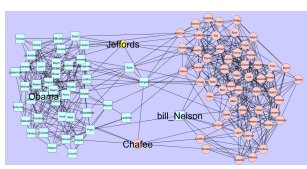


Case Study 2: The 109th congress

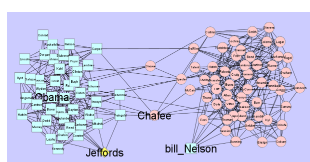




March 2005



January 2006



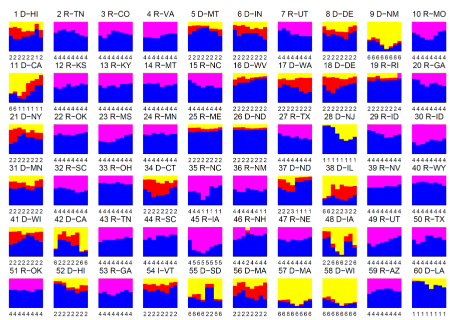
August 2006

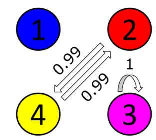
US senator voting records

100 senators, 109th Congress (Jan 2005 – Dec 2006) in 8 epochs

Senate Network: role trajectories

Voting data preprocessed into a network graph using (Kolar *et al.*, 2008)





Colored bars: Estimated latent space vector
Numbers under bars: Estimated cluster
Letters beside actor index: Political party and State

Role Compatibility Matrix *B*
Role 1 = Passive, 2/4 = Democratic clique,
3 = Republican clique

Senate Network: role trajectories

Cluster legend

Jon Corzine's seat (#28, Democrat, New Jersey) was taken over by Bob Menendez from *t*=5 onwards.

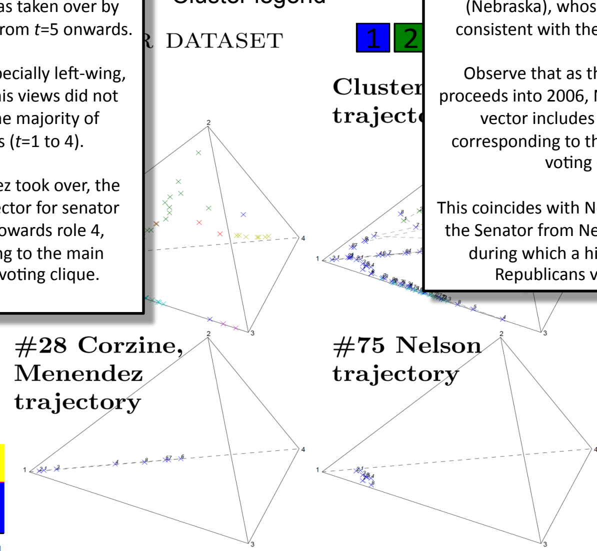
Corzine was especially left-wing, so much that his views did not align with the majority of Democrats (*t*=1 to 4).

Once Menendez took over, the latent space vector for senator #28 shifted towards role 4, corresponding to the main Democratic voting clique.

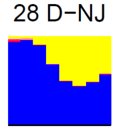
Ben Nelson (#75) is a right-wing Democrat (Nebraska), whose views are more consistent with the Republican party.

Observe that as the 109th Congress proceeds into 2006, Nelson's latent space vector includes more of role 3, corresponding to the main Republican voting clique.

This coincides with Nelson's re-election as the Senator from Nebraska in late 2006, during which a high proportion of Republicans voted for him.

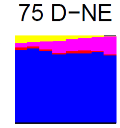


28 D−NJ



1 1 1 1 1 1 1 1

75 D−NE



1 1 1 1 1 1 1 1

Summary of this tutorial

- 1. Overview of basic topic models
- 2. Computational challenges and two classical algorithmic paths
- 3. Scenario I: Multimodal data
- 4. Scenario II: When supervision is available
- 5. Scenario III: What if I don't know the total number of topics
- 6. Scenario IV: Topic evolution in streaming corpus.
- 7: Advanced subject I: Sparsity in topic modeling (see EMNLP talk)
- 8: Advanced subject II: Scalability, complexity, and fast algorithms (optional)
- 9: Other applications (optional)

Conclusion

- GM-based topic models are cool
  - Flexible
  - Modular
  - Interactive
- There are many ways of implementing topic models
  - unsupervised
  - supervised
- Efficient Inference/learning algorithms
  - GMF, with Laplace approx. for non-conjugate dist.
  - MCMC
- Many applications
  - …
  - Word-sense disambiguation
  - Image understanding
  - Network inference

More research questions we ask:

- Event detection/prediction
  - Emergence/disappearance/evolution of perspective, bias, object, theme, etc.
  - Is there going to be a war? When? Can we predict the economy or stock from traditional or internet news?
- Automated summary
  - Describe a scene or arbitrary image
  - From keyword or class-label to story
- Semantic-based browsing and search
  - Ranking/matching based on topic/perspective
  - Video retrieval based on story
- Theoretical properties
  - Does TM have an invariant, unique solution, under what condition it is attainable?
  - How fast we converge to such solution? What requirement data must satisfy?
- Scalable computing
  - Easy, converging, fault tolerant, distributed, and online topical inference/learning
  - **Doing all these with Facebook or Twitter or Flickr**