Machine Learning

How to put things together ?

A case-study of model design, inference, learning, evaluation in text analysis

Eric Xing

Lecture 20, August 16, 2010

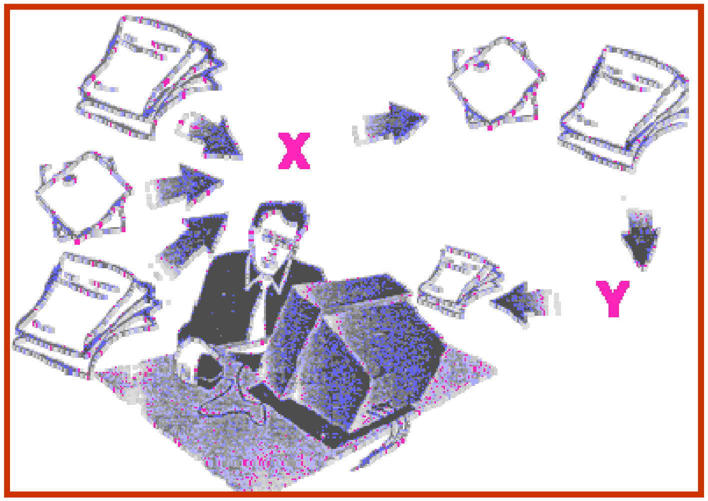
Reading:

Need computers to help us…

(from images.google.cn)

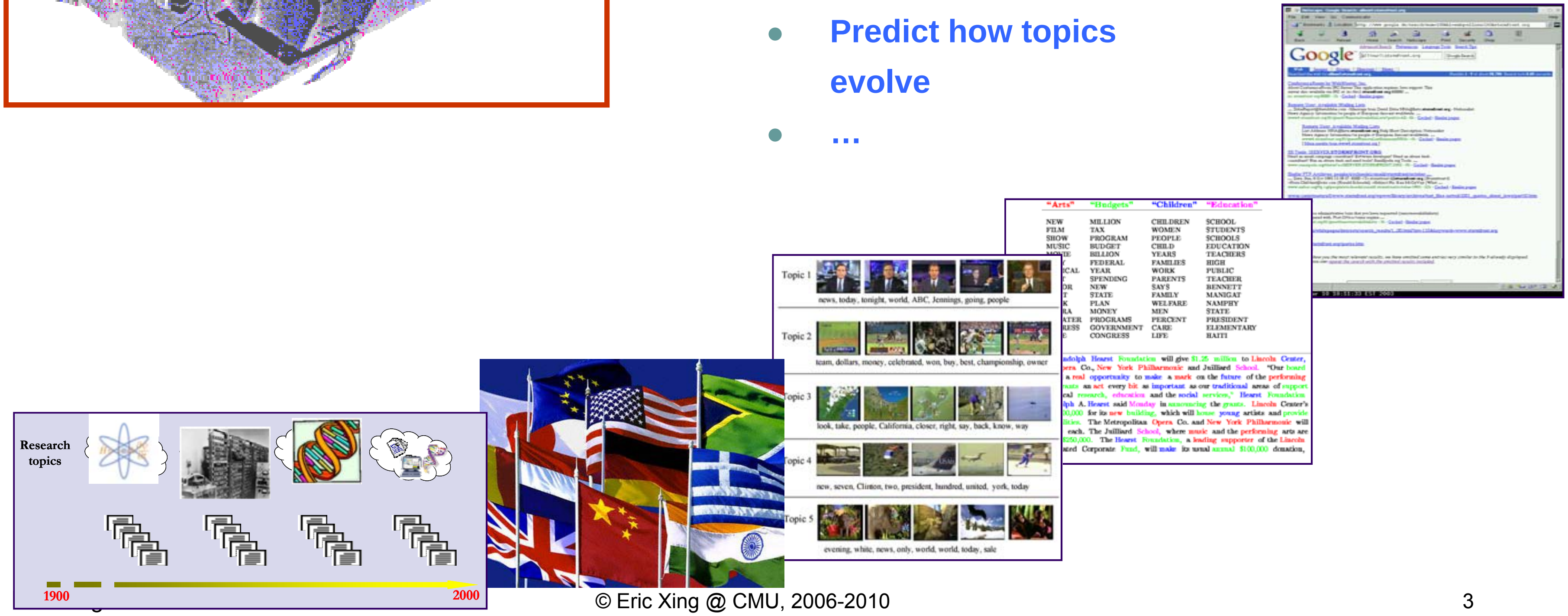
- Humans cannot afford to deal with (e.g., search, browse, or measure similarity) a huge number of text documents
- We need computers to help out …

NLP and Data Mining



We want:

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



Topic 1 news, today, tonight, world, ABC, Jennings, going, people

Topic 2 team, dollars, money, celebrated, won, buy, best, championship, owner

Topic 3 look, take, people, California, closer, right, say, back, know, way

Topic 4 new, seven, Clinton, two, president, hundred, united, york, today

Topic 5 evening, white, news, only, world, world, today, sale

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
Y	FEDERAL	FAMILIES	HIGH
ICAL	YEAR	WORK	PUBLIC
T	SPENDING	PARENTS	TEACHER
OR	NEW	SAYS	BENNETT
T	STATE	FAMILY	MANIGAT
K	PLAN	WELFARE	NAMPHY
RA	MONEY	MEN	STATE
ATER	PROGRAMS	PERCENT	PRESIDENT
RESS	GOVERNMENT	CARE	ELEMENTARY
E	CONGRESS	LIFE	HAITI

adolph Hearst Foundation will give $1.25 million to Lincoln Center,
pera Co., New York Philharmonic and Juilliard School. "Our board
a real opportunity to make a mark on the future of the performing
rants an act every bit as important as our traditional areas of support
cal research, education and the social services," Hearst Foundation
lph A. Hearst said Monday in announcing the grants. Lincoln Center's
00,000 for its new building, which will house young artists and provide
lities. The Metropolitan Opera Co. and New York Philharmonic will
each. The Juilliard School, where music and the performing arts are
$250,000. The Hearst Foundation, a leading supporter of the Lincoln
ated Corporate Fund, will make its usual annual $100,000 donation,

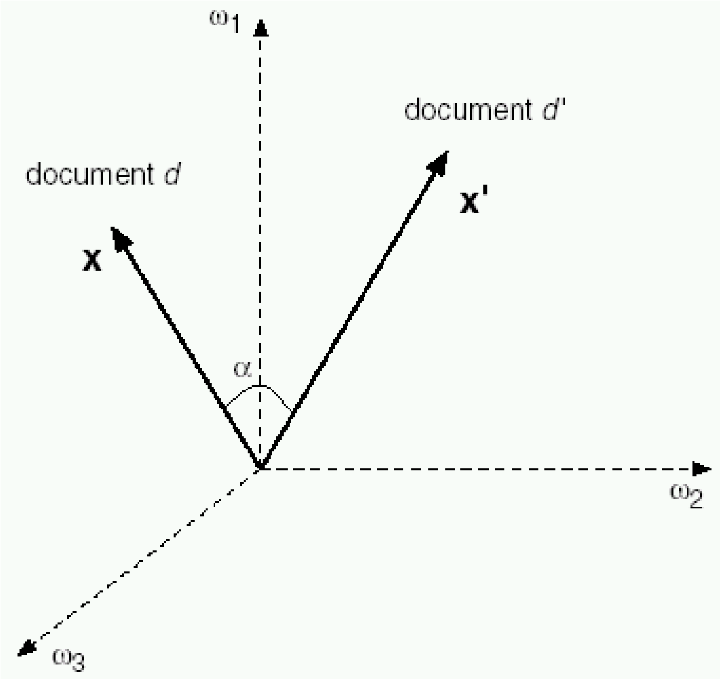
How to get started?

- **Here are some important elements to consider before you start:**
  - Task:
    - Embedding? Classification? Clustering? Topic extraction? …
  - Data representation:
    - Input and output (e.g., continuous, binary, counts, …)
  - Model:
    - BN? MRF? Regression? SVM?
  - Inference:
    - Exact inference? MCMC? Variational?
  - Learning:
    - MLE? MCLE? Max margin?
  - 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/categorizing
  - Distill semantics and perspectives
  - ..

Modeling document collections

- A document collection is a dataset where each data point is itself a collection of simpler data.
  - Text documents are collections of words.
  - Segmented images are collections of regions.
  - User histories are collections of purchased items.

- Many modern problems ask questions of such data.
  - Is this text document relevant to my query?
  - Which category is this image in?
  - What movies would I probably like?
  - Create a caption for this image.
  - Modeling document collections

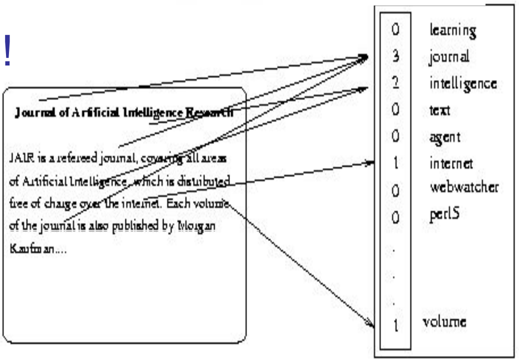
Representation:

- 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 $(|V| \gg D)$
  - Not efficient text processing tasks, e.g., search, document classification, or similarity measure
  - Not effective for browsing



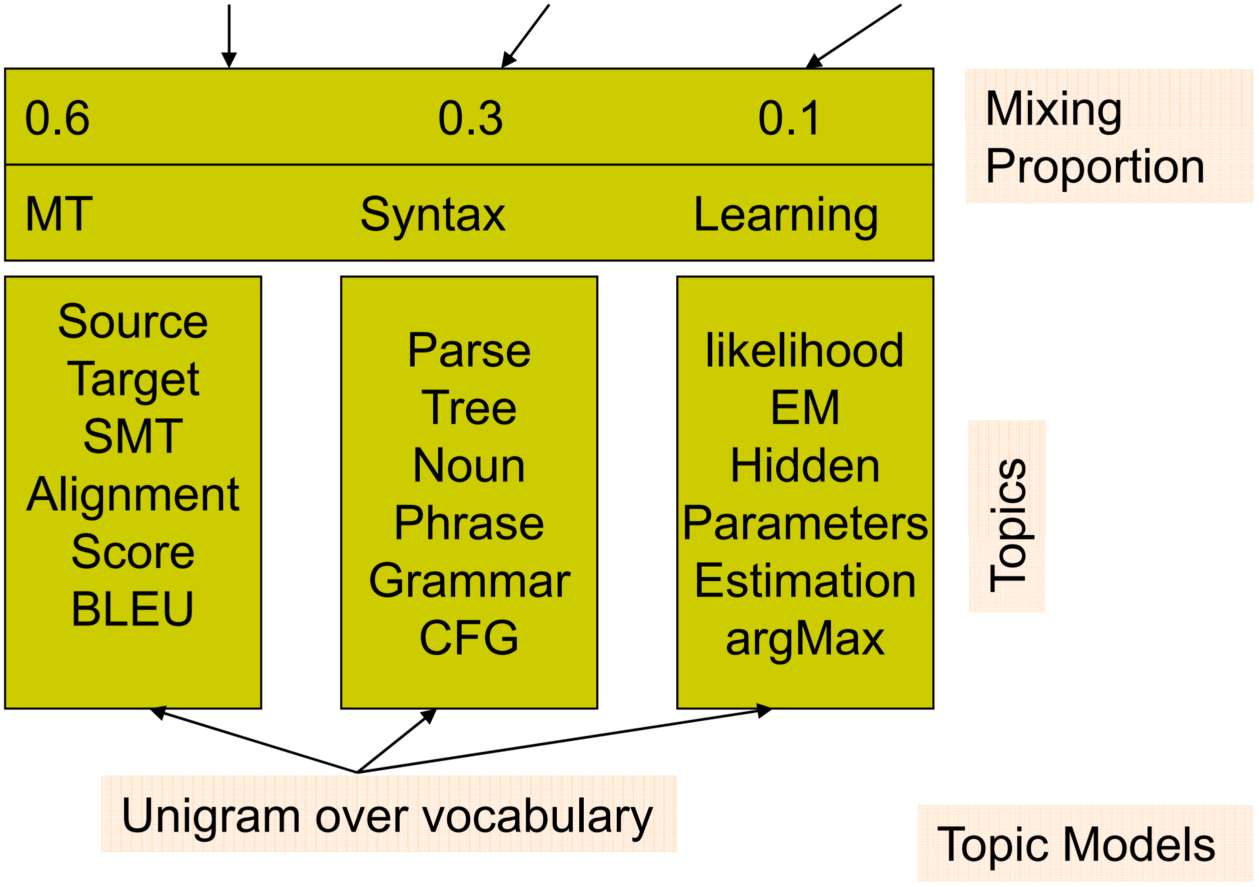
Journal of Artificial Intelligence Research

JAIR is a refereed journal, covering all areas of Artificial Intelligence, which is distributed free of charge over the internet. Each volume of the journal is also published by Morgan Kaufman....

0	learning
3	journal
2	intelligence
0	text
0	agent
1	internet
0	webwatcher
0	perl5
.	
.	
.	
1	volume

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

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

Topics

Unigram over vocabulary

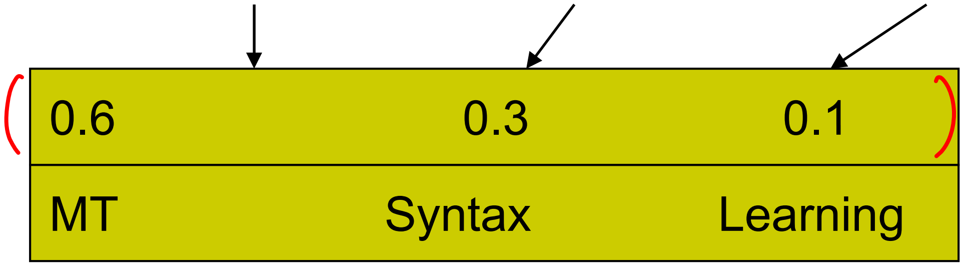
Topic Models

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.

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

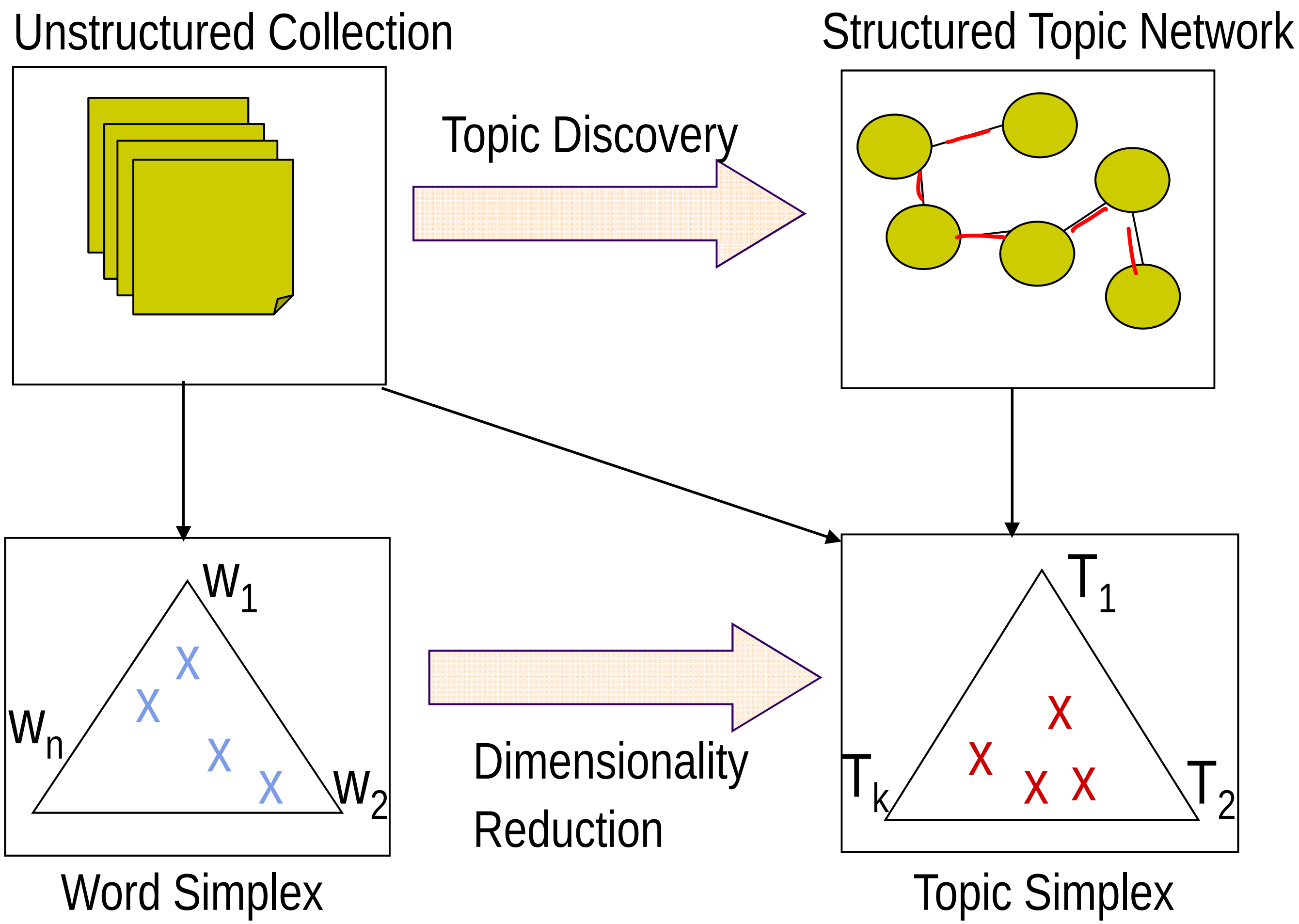
Mixing Proportion

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 …

Topic Models: The Big Picture



Topic Models

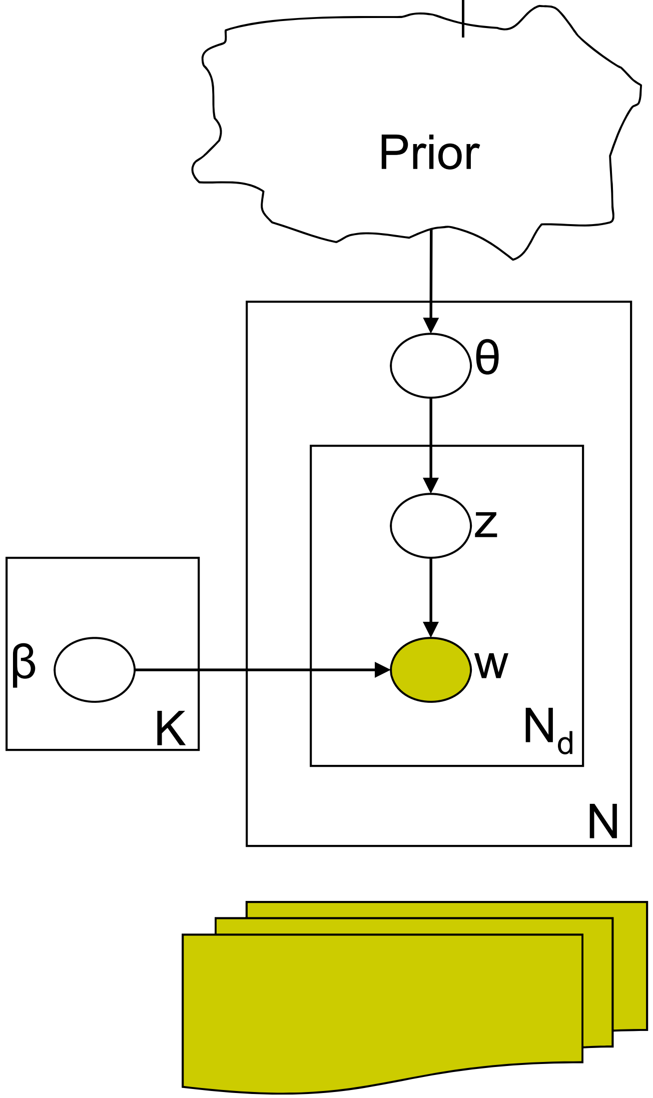
Generating a document

– *Draw* θ from the prior

For each word n

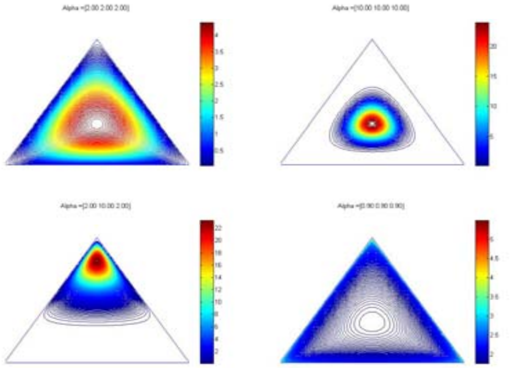
- Draw z_n from $multinomial(\theta)$
- Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$

Which prior to use?

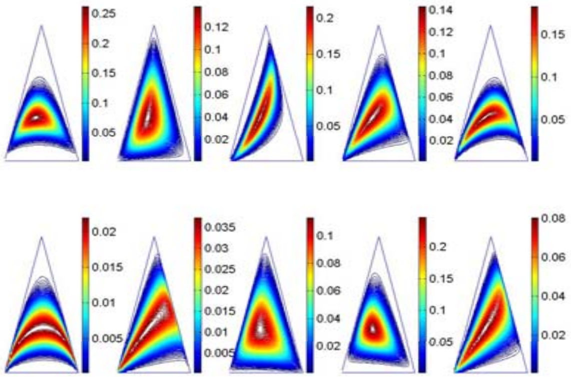


Choices of Priors

- 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



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 $multinomial(\beta_{z_n})$

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

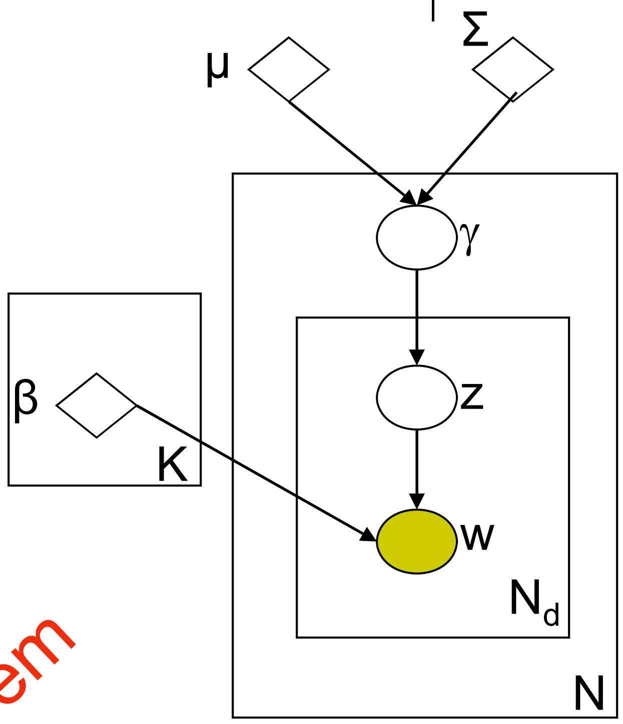
$$\gamma \sim N_{K-1}(\mu, \Sigma) \qquad \gamma_K = 0$$

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

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

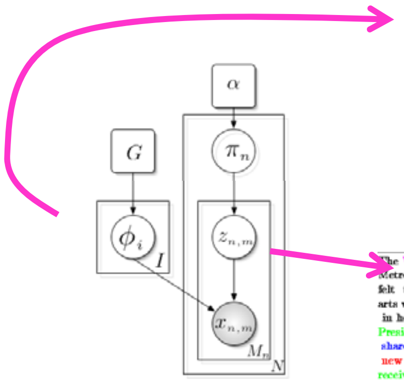
- Log Partition Function
- Normalization Constant

Problem



Using the Model

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



"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.

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

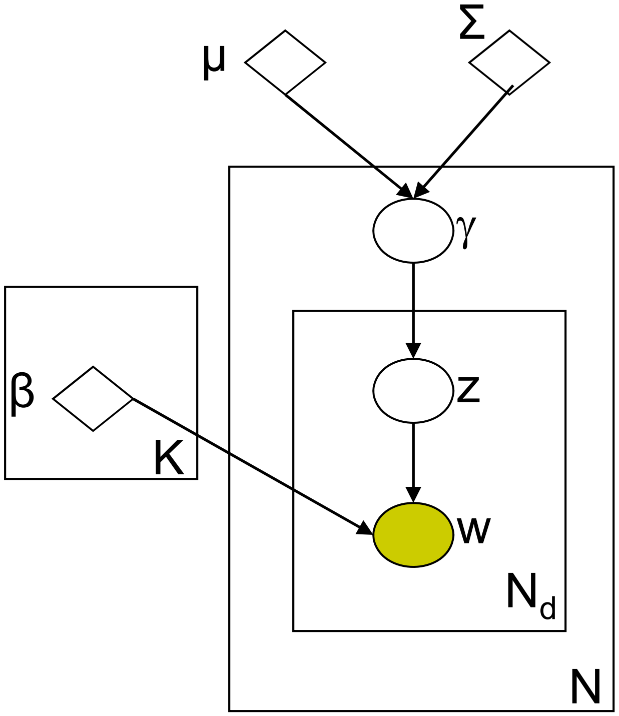
$$\arg\max_{(\mu,\Sigma,\beta)} \sum \log\left(P\left(D_i \middle| \mu,\Sigma,\beta\right)\right)$$

Inference

$$P(D|\mu,\Sigma,\beta)=\prod_{n=1}^{N_d}P(w_n|\mu,\Sigma,\beta)$$

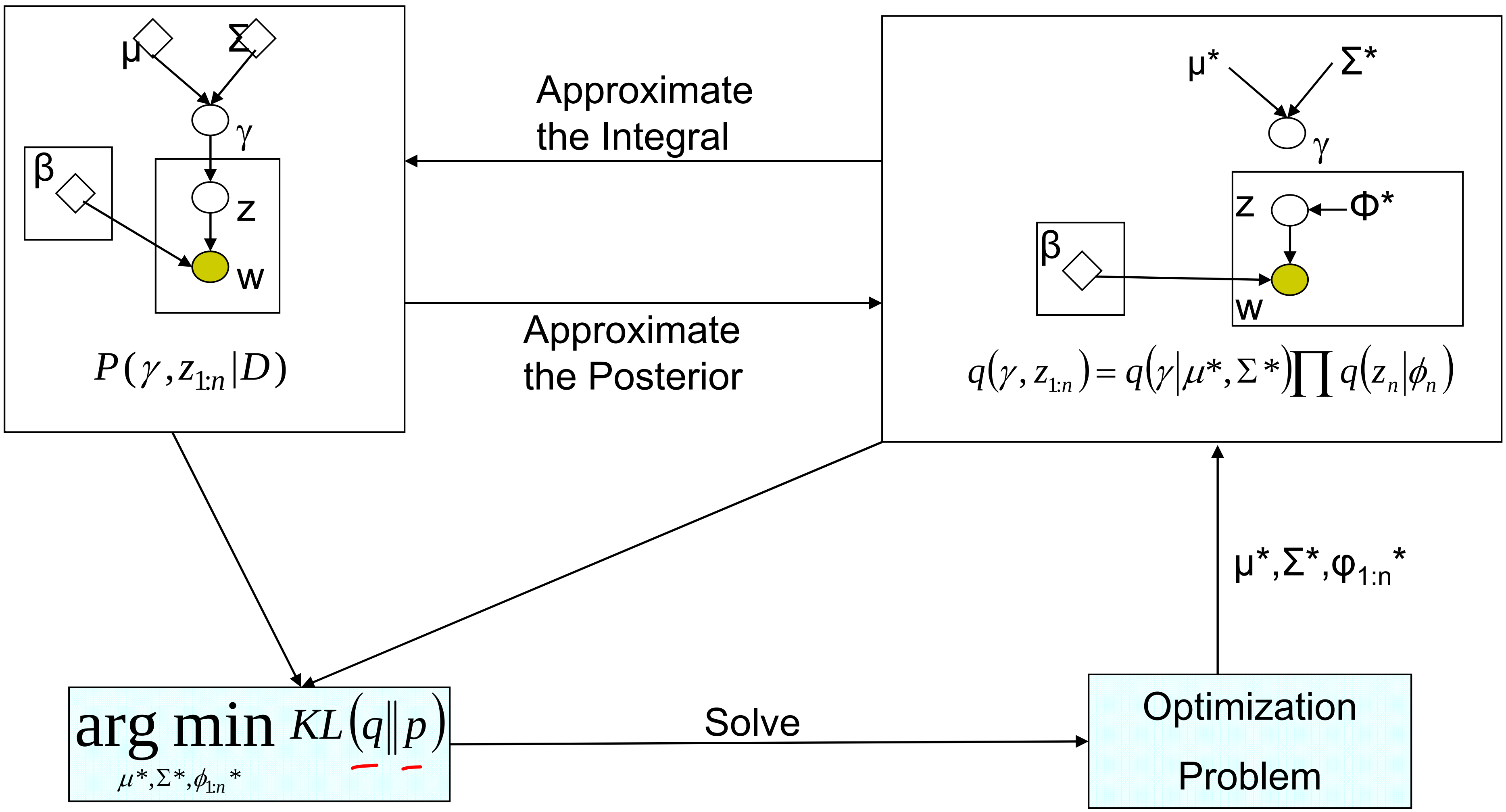
$$=\int_{\gamma}\prod_{n=1}^{N_d}\sum_{z_n=1}^{K}P(w_n,z_n,\gamma,|\mu,\Sigma,\beta)d\gamma$$

$$=\int_{\gamma}\left(\prod_{n=1}^{N_d}\sum_{z_n=1}^{K}P(w_n|z_n,\beta)P(z_n|\gamma)\right)P(\gamma|\mu,\Sigma)d\gamma$$



Intractable → approximate inference

Variational Inference



$P(\gamma, z_{1:n} | D)$

Approximate the Integral

Approximate the Posterior

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

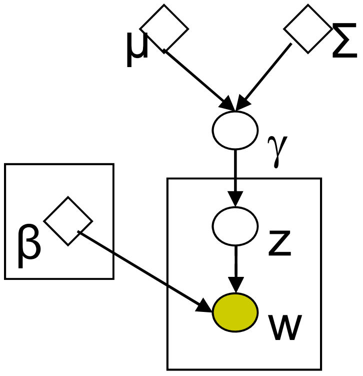
$\mu^*, \Sigma^*, \varphi_{1:n}{}^*$

$\arg\min_{\mu^*, \Sigma^*, \phi_{1:n}{}^*} KL(\underline{q} \| \underline{p})$

Solve

Optimization Problem

Variational Inference With no Tears



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

Iterate until Convergence

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

- 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

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

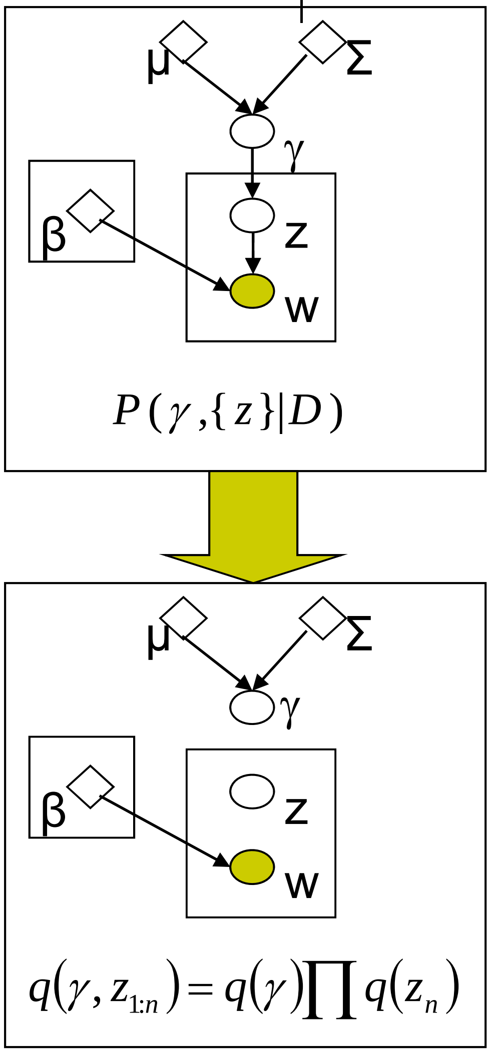
- Two clusters: λ and $Z_{1:n}$

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

- 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 $\langle S_z \rangle_{q_z}$?

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

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

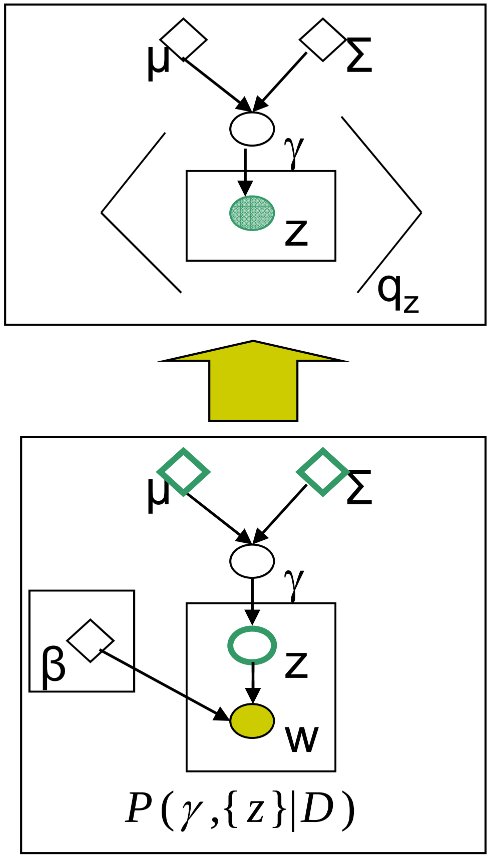
$$\propto \exp\left\{\underbrace{-\frac{1}{2}\gamma'\Sigma^{-1}\gamma + \gamma\Sigma^{-1}\mu + \langle m \rangle_{q_z} \gamma} - N \times \underbrace{\boxed{C(\gamma)}}\right\}$$

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

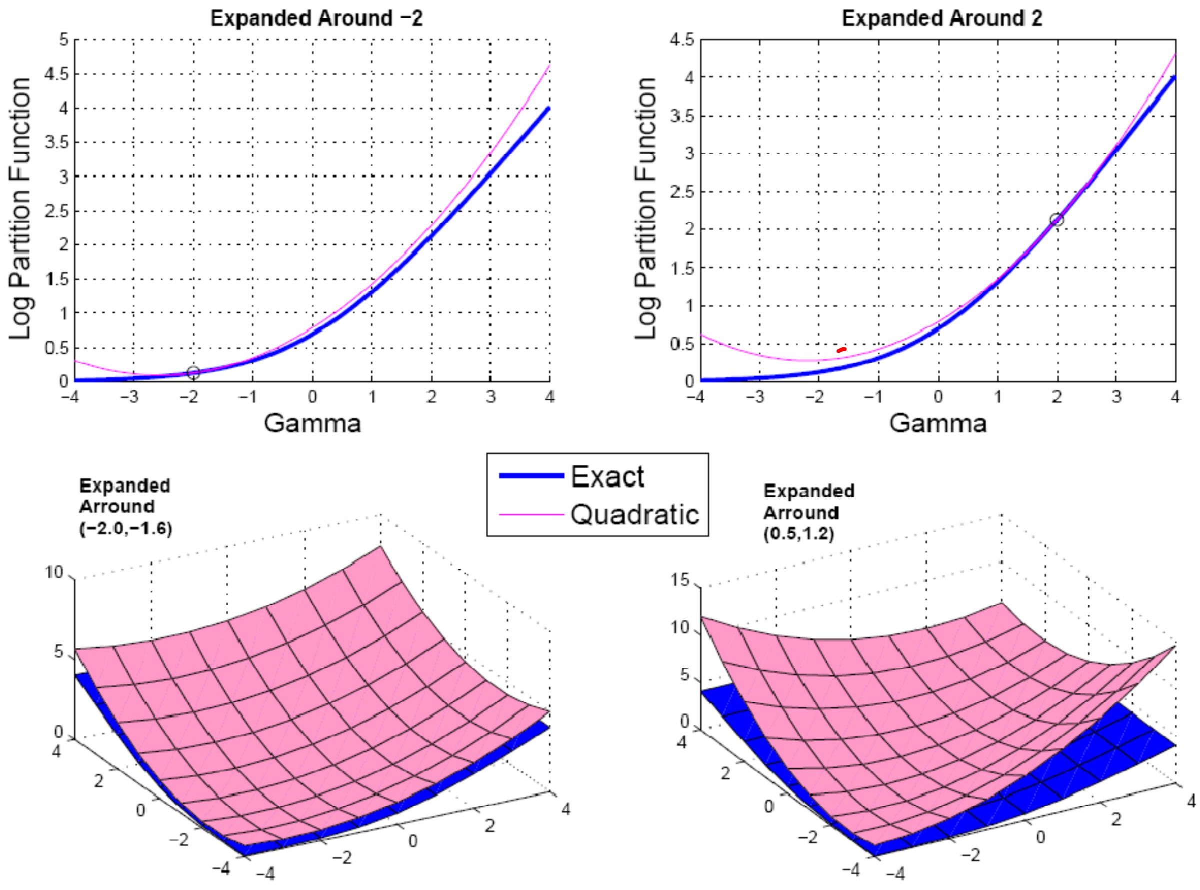
$$q_\lambda^*(\gamma) = N(\mu_\gamma, \Sigma_\gamma)$$

$$\Sigma_\gamma = inv\left(\Sigma^{-1} + NH\right)$$

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



Approximation Quality



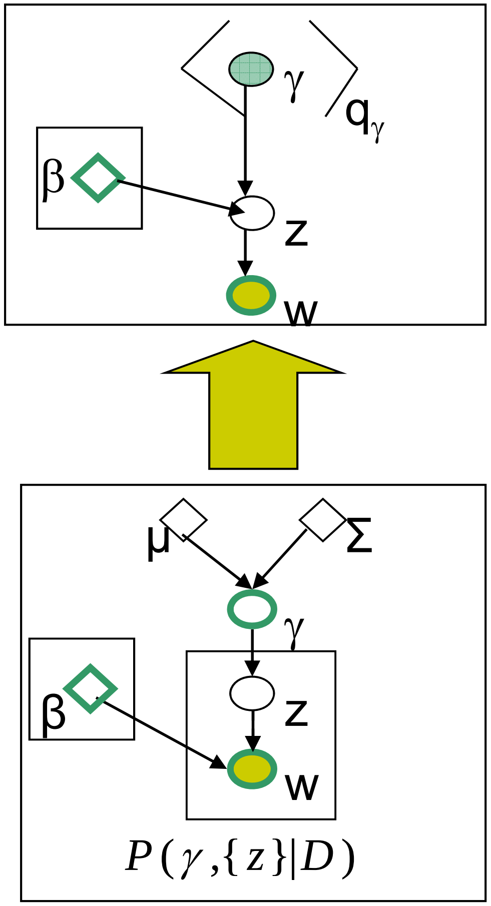
Variational Z

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

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

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

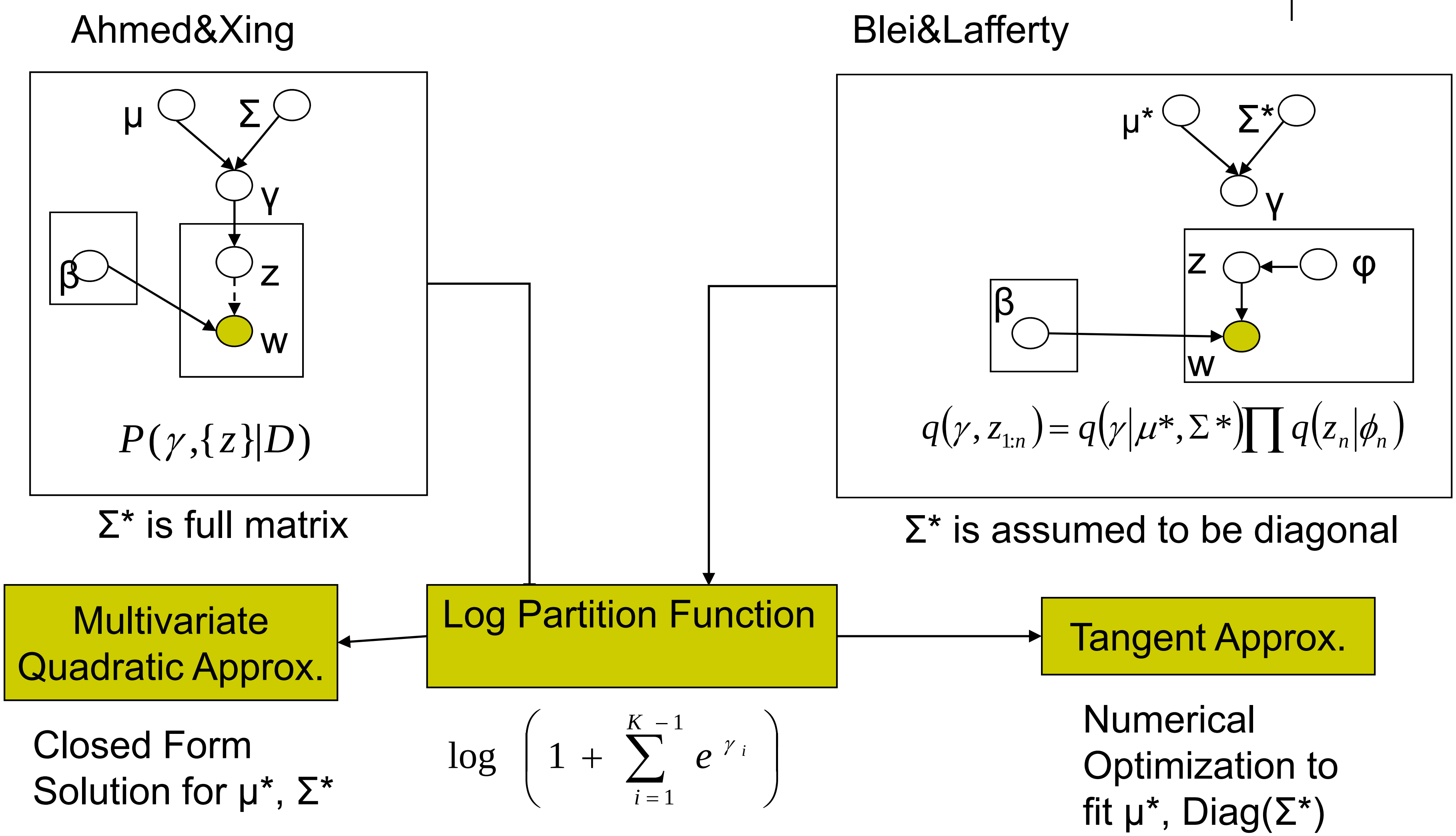
$$\propto \exp\left\{\mu_{\gamma,k}\right\} \beta_{kj}$$



Variational Inference: Recap

- Run one document at a time
- Message Passing Scheme
  - $\mathrm{GMF}_{\{z\} \rightarrow \gamma} = \langle m \rangle$
  - $\mathrm{GMF}_{\gamma \rightarrow z} = \langle \gamma \rangle$
- Iterate until Convergence
- Posterior over γ is a MVN with full covariance

Now you’ve got an algorithm, don’t forget to compare with related work



Ahmed&Xing

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

Σ* is full matrix

Blei&Lafferty

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

Σ* is assumed to be diagonal

Multivariate Quadratic Approx.

Closed Form Solution for μ*, Σ*

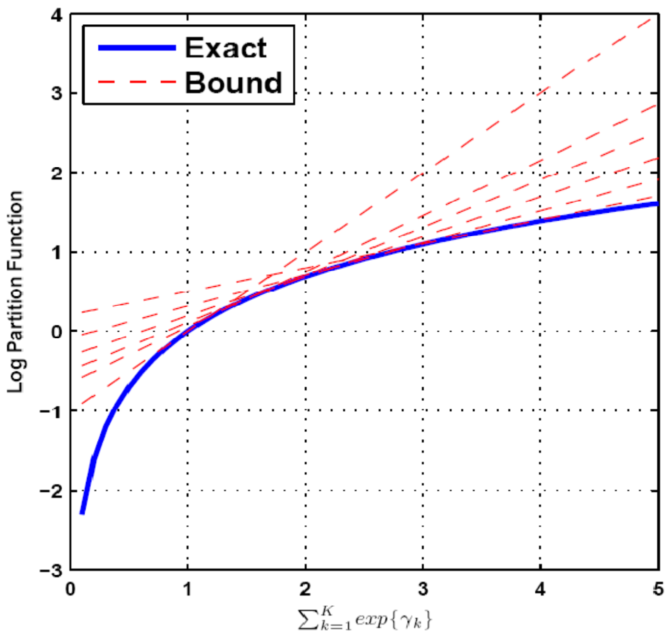
Log Partition Function

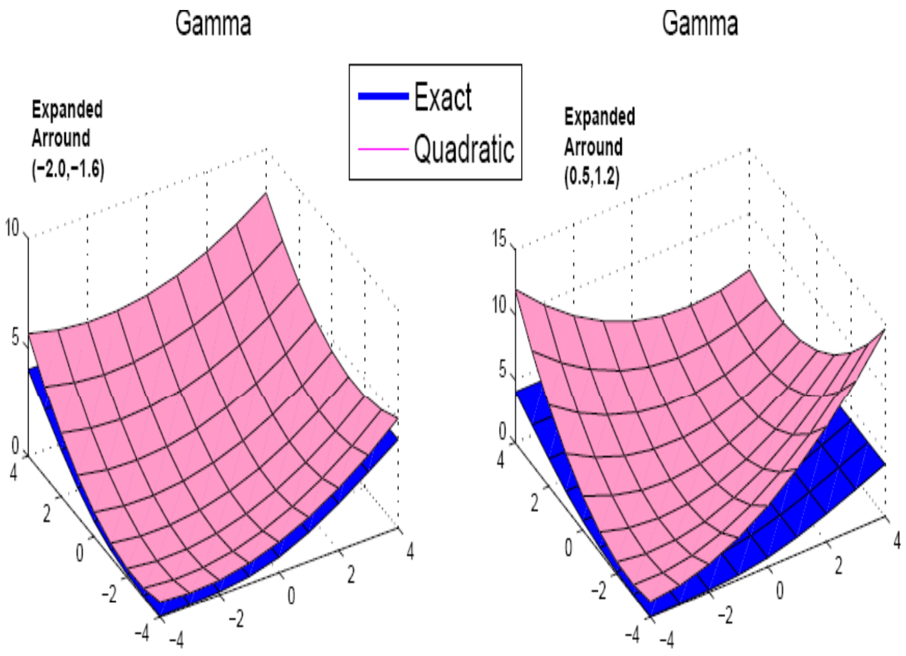
$$\log\left(1 + \sum_{i=1}^{K-1} e^{\gamma_i}\right)$$

Tangent Approx.

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

Tangent Approximation





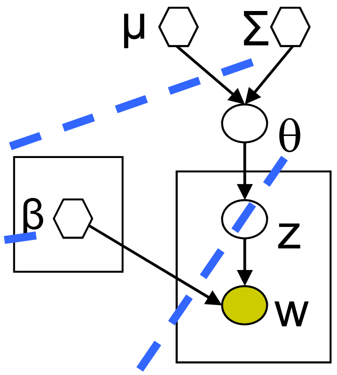
Evaluation

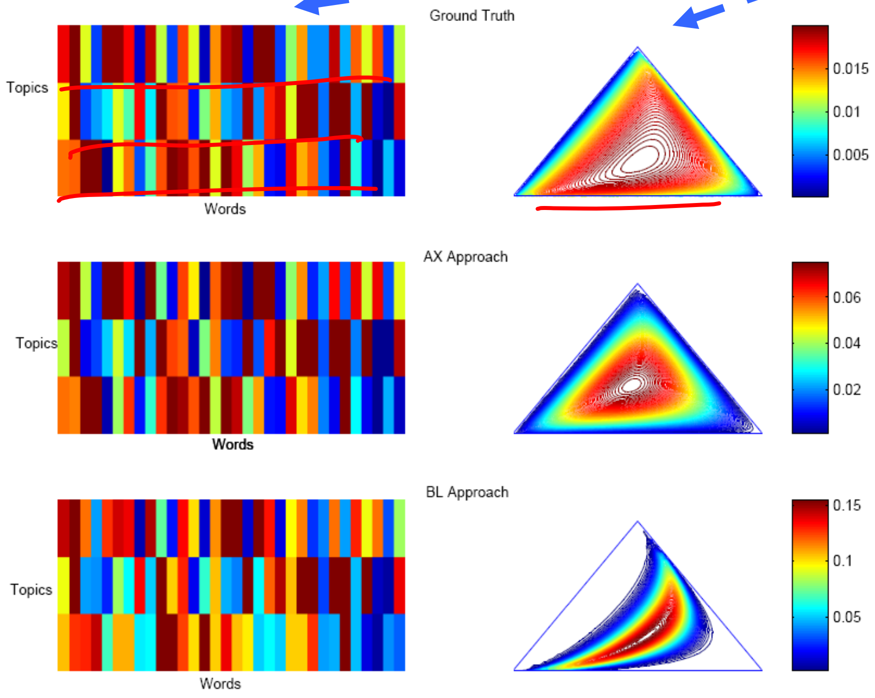
- A common (but not so right) practice
  - Try models on real data for some empirical task, say classifications or topic extraction; two reactions
    - Hmm! The results “make sense to me”, so the model is good!
      - Objective?
    - Gee! The results are terrible! Model is bad!
      - Where does the error come from?

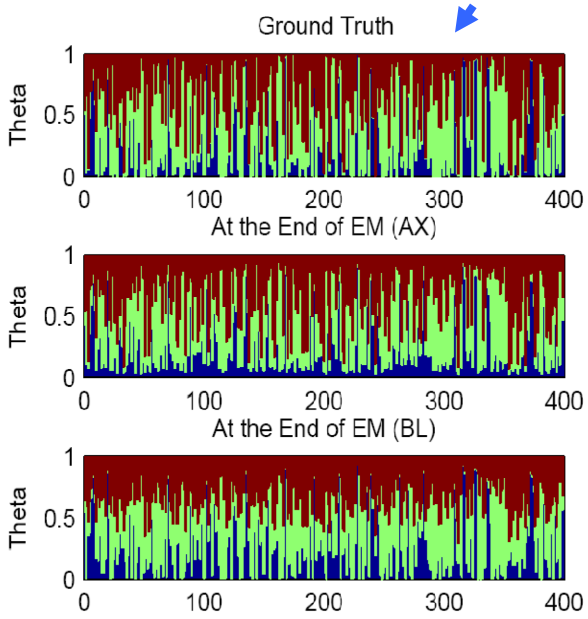
Evaluation: testing inference

- Simulated Data
  - **We know the ground truth** for Θ ,
    - This is a crucial step because it can discern performance loss due to modeling insufficiency from inference inaccuracy
  - Vary model dimensions
    - K= Number of topics
    - M= vocabulary size
    - Nd= number of words per document
- Test
  - Inference
    - Accuracy of the recovered Θ
    - Number of Iteration to converge (1e-6, default setting)
  - Parameter Estimation
    - Goal: $\arg\max_{(\mu,\Sigma,\beta)} \sum \log\left(P\left(D_i \middle| \mu, \Sigma, \beta\right)\right)$
    - Standard VEM + Deterministic Annealing

Test on Synthetic Text



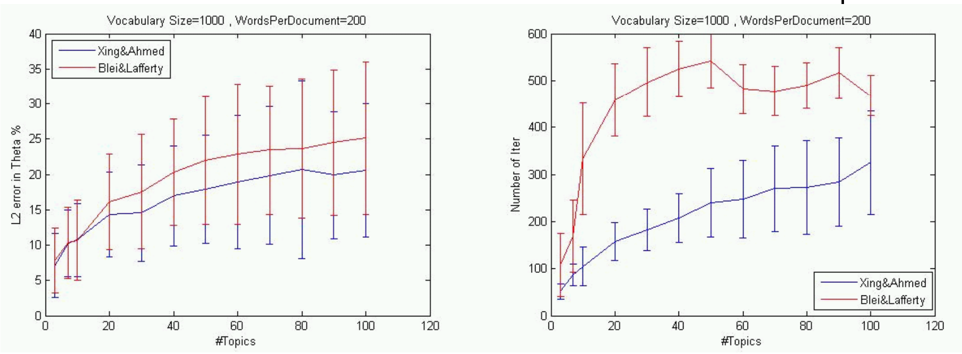




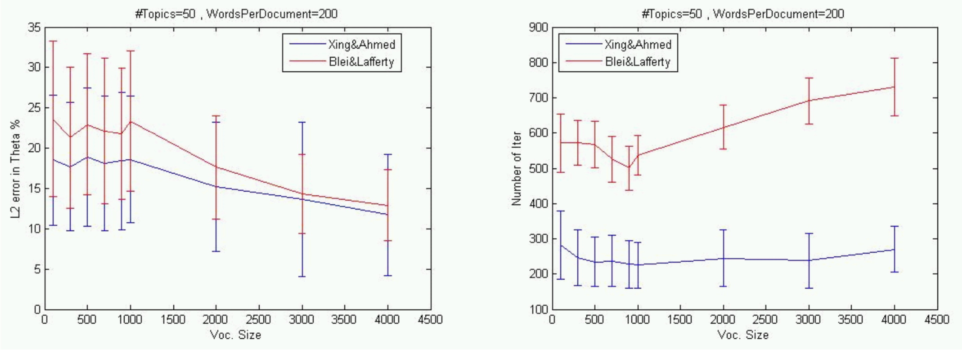
Comparison: accuracy and speed

L2 error in topic vector est. and # of iterations

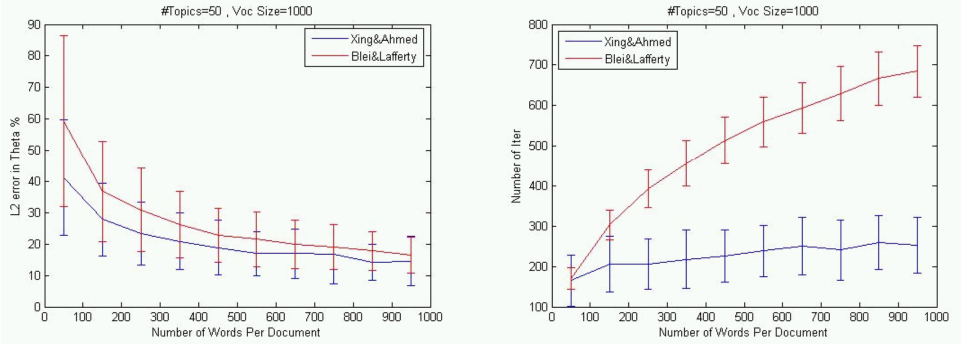
- Varying Num. of Topics



- Varying Voc. Size



- Varying Num. Words Per Document

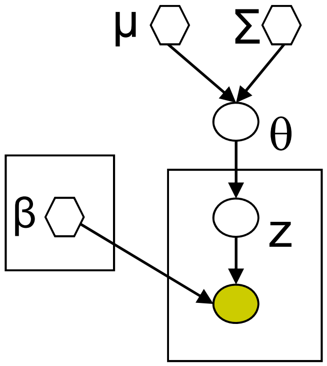


Parameter Estimation

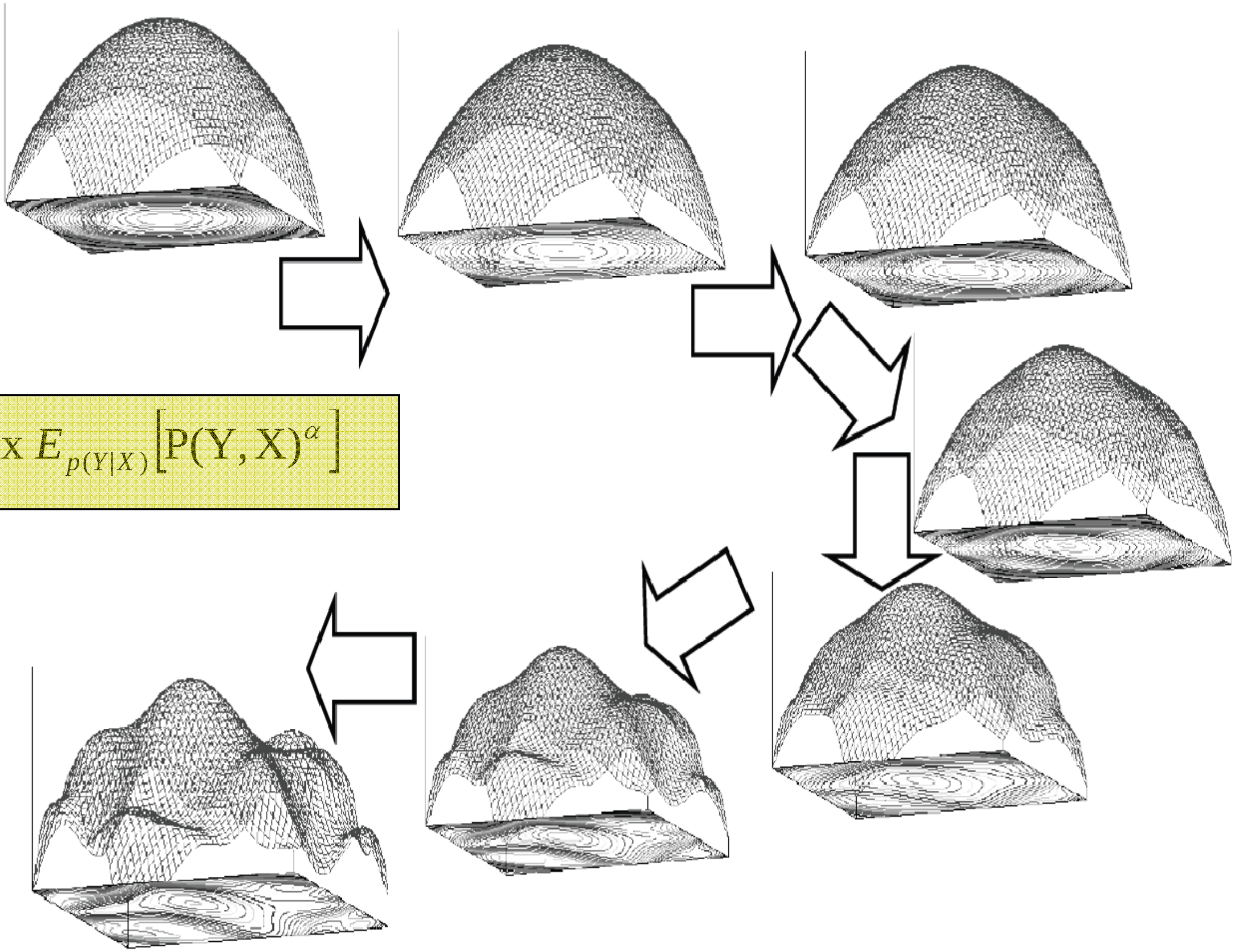
- Goal:

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

- Standard VEM
  - fitting μ*, Σ* for each document
  - get model parameter using their expected sufficient Statistics
  - Problems
    - Having full covariance for the posterior traps our model in local maxima
    - Solution:
      - Deterministic Annealing



Deterministic Annealing: Big Picture



$$(\mu,\Sigma,\beta)^* = \arg\max_{\mu,\Sigma,\beta} E_{p(Y|X)}\left[\mathrm{P}(\mathrm{Y},\mathrm{X})^{\alpha}\right]$$

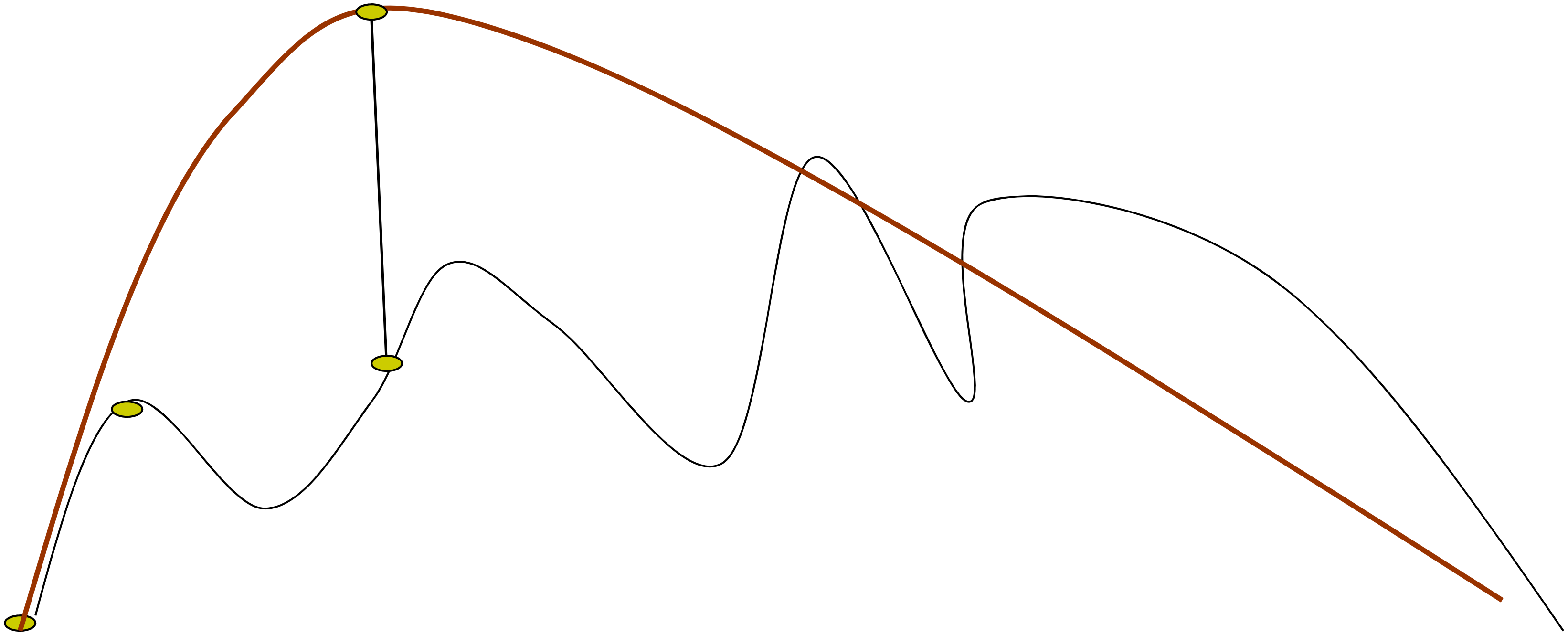
Deterministic Annealing

- EM

$$(\mu,\Sigma,\beta)^* = \arg\max_{\mu,\Sigma,\beta} E_{p(Y|X)}\left[\mathrm{P}(\mathrm{Y},\mathrm{X})\right]$$

- DA-EM

$$(\mu,\Sigma,\beta)^* = \arg\max_{\mu,\Sigma,\beta} E_{p(Y|X)}\left[\mathrm{P}(\mathrm{Y},\mathrm{X})^\alpha\right]$$



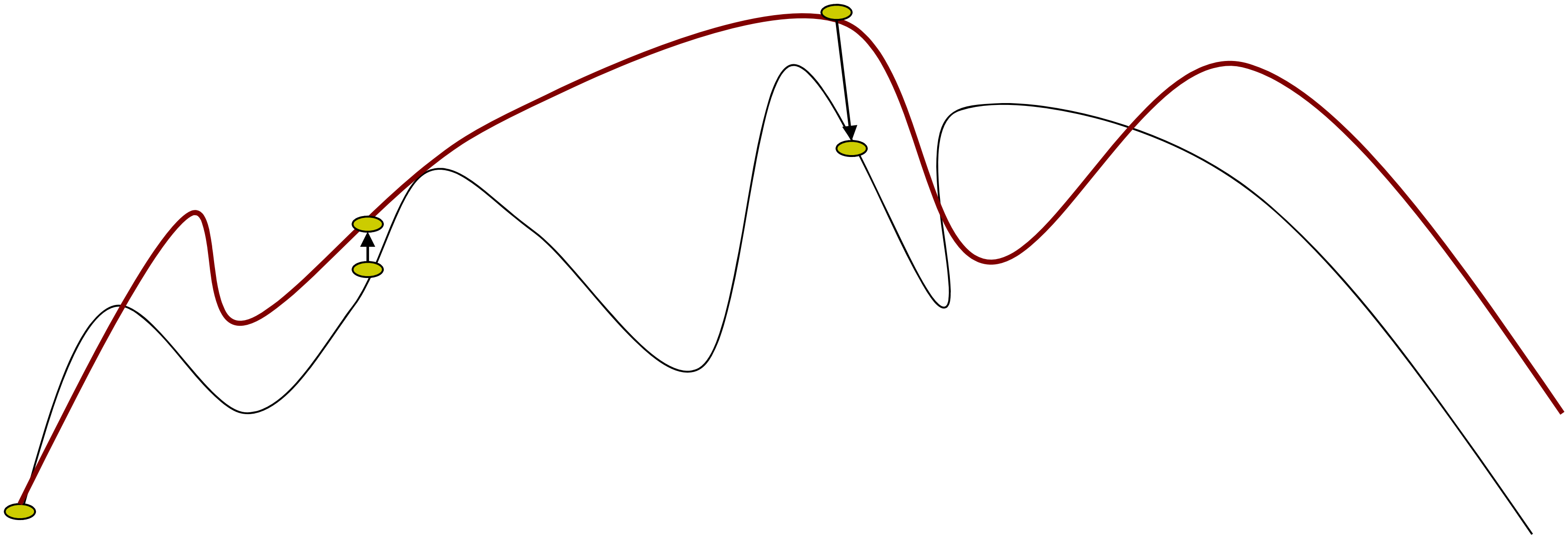
Deterministic Annealing

- EM

$$(\mu, \Sigma, \beta)^* = \arg\max_{\mu,\Sigma,\beta} E_{p(Y|X)}[\mathrm{P}(\mathrm{Y}, \mathrm{X})]$$

- DA-EM

$$(\mu, \Sigma, \beta)^* = \arg\max_{\mu,\Sigma,\beta} E_{p(Y|X)}[\mathrm{P}(\mathrm{Y}, \mathrm{X})^\alpha]$$



Deterministic Annealing

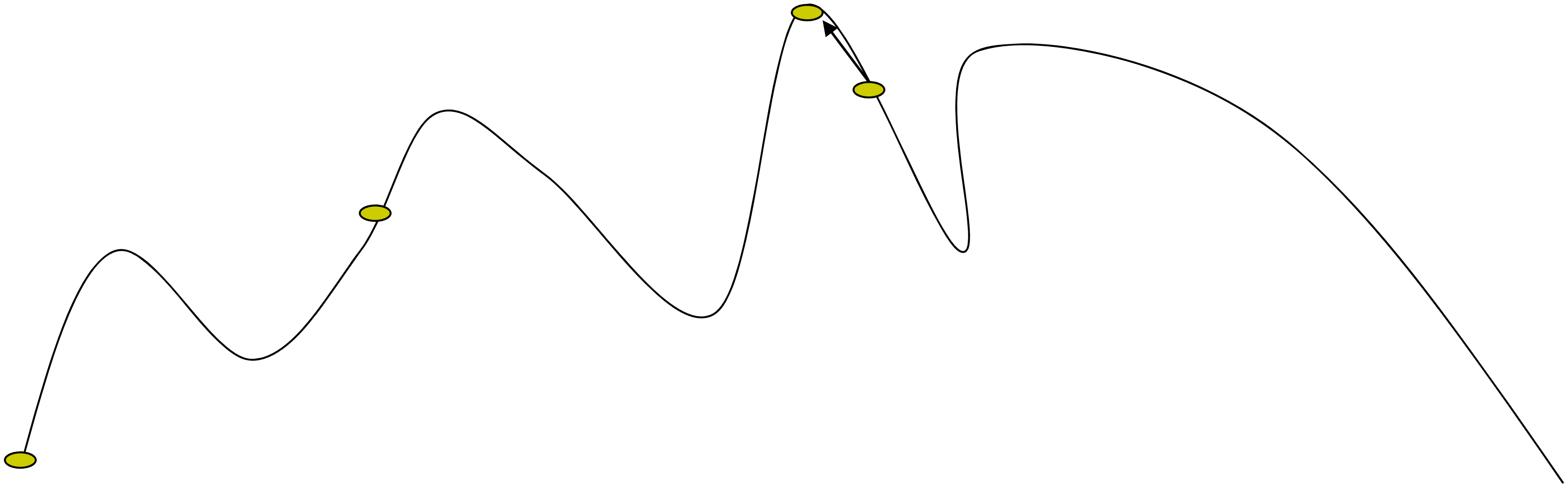
- EM

$$(\mu, \Sigma, \beta)^* = \arg\max_{\mu,\Sigma,\beta} E_{p(Y|X)}\left[\mathrm{P}(\mathrm{Y}, \mathrm{X})\right]$$

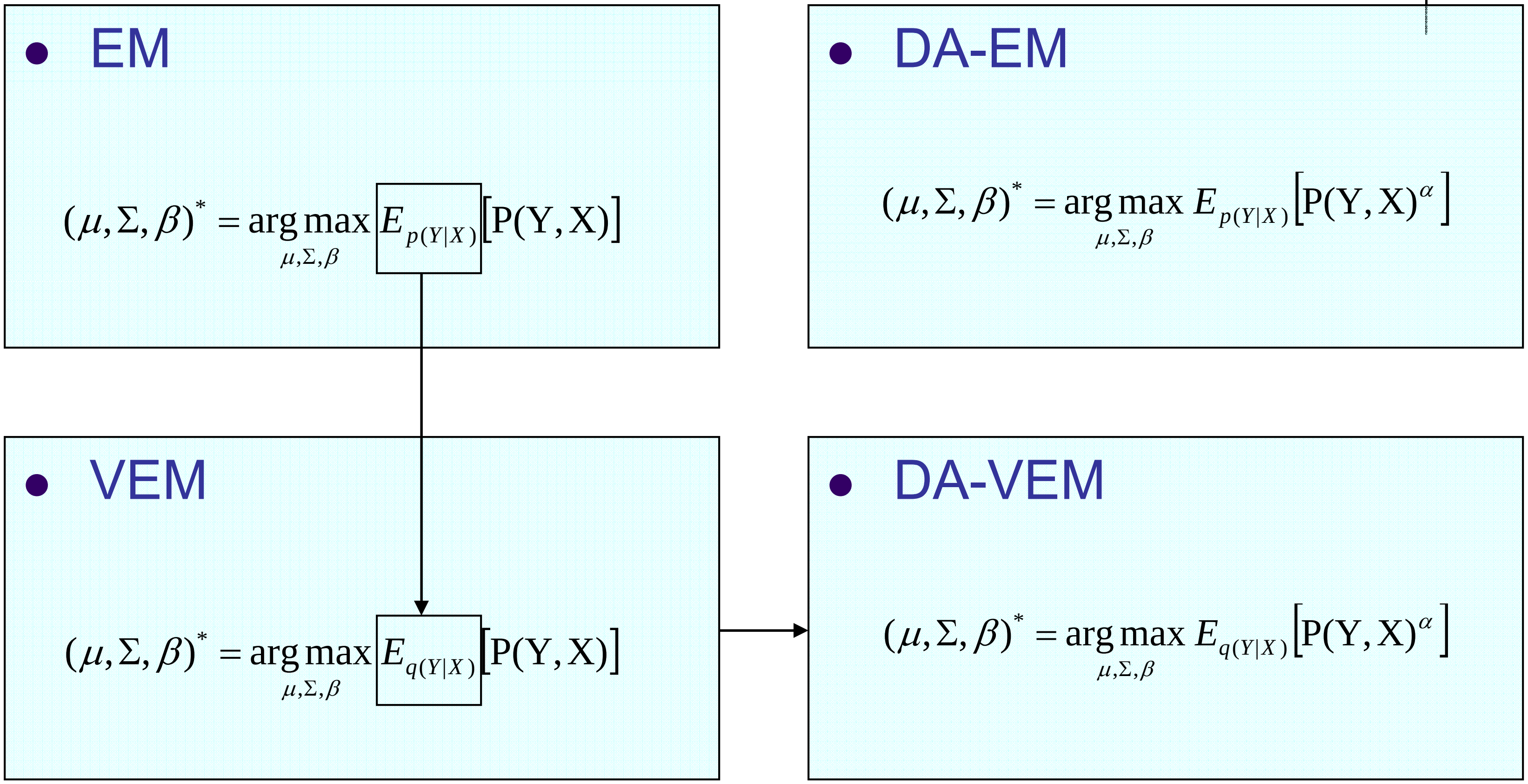
- DA-EM

$$(\mu, \Sigma, \beta)^* = \arg\max_{\mu,\Sigma,\beta} E_{p(Y|X)}\left[\mathrm{P}(\mathrm{Y}, \mathrm{X})^{\alpha}\right]$$

Life is not always that Good



Deterministic Annealing



- EM

$$(\mu,\Sigma,\beta)^* = \arg\max_{\mu,\Sigma,\beta} \boxed{E_{p(Y|X)}}\left[\mathrm{P}(\mathrm{Y},\mathrm{X})\right]$$

- DA-EM

$$(\mu,\Sigma,\beta)^* = \arg\max_{\mu,\Sigma,\beta} E_{p(Y|X)}\left[\mathrm{P}(\mathrm{Y},\mathrm{X})^{\alpha}\right]$$

- VEM

$$(\mu,\Sigma,\beta)^* = \arg\max_{\mu,\Sigma,\beta} \boxed{E_{q(Y|X)}}\left[\mathrm{P}(\mathrm{Y},\mathrm{X})\right]$$

- DA-VEM

$$(\mu,\Sigma,\beta)^* = \arg\max_{\mu,\Sigma,\beta} E_{q(Y|X)}\left[\mathrm{P}(\mathrm{Y},\mathrm{X})^{\alpha}\right]$$

-For exponential Families, this requires two line change to standard (V)EM

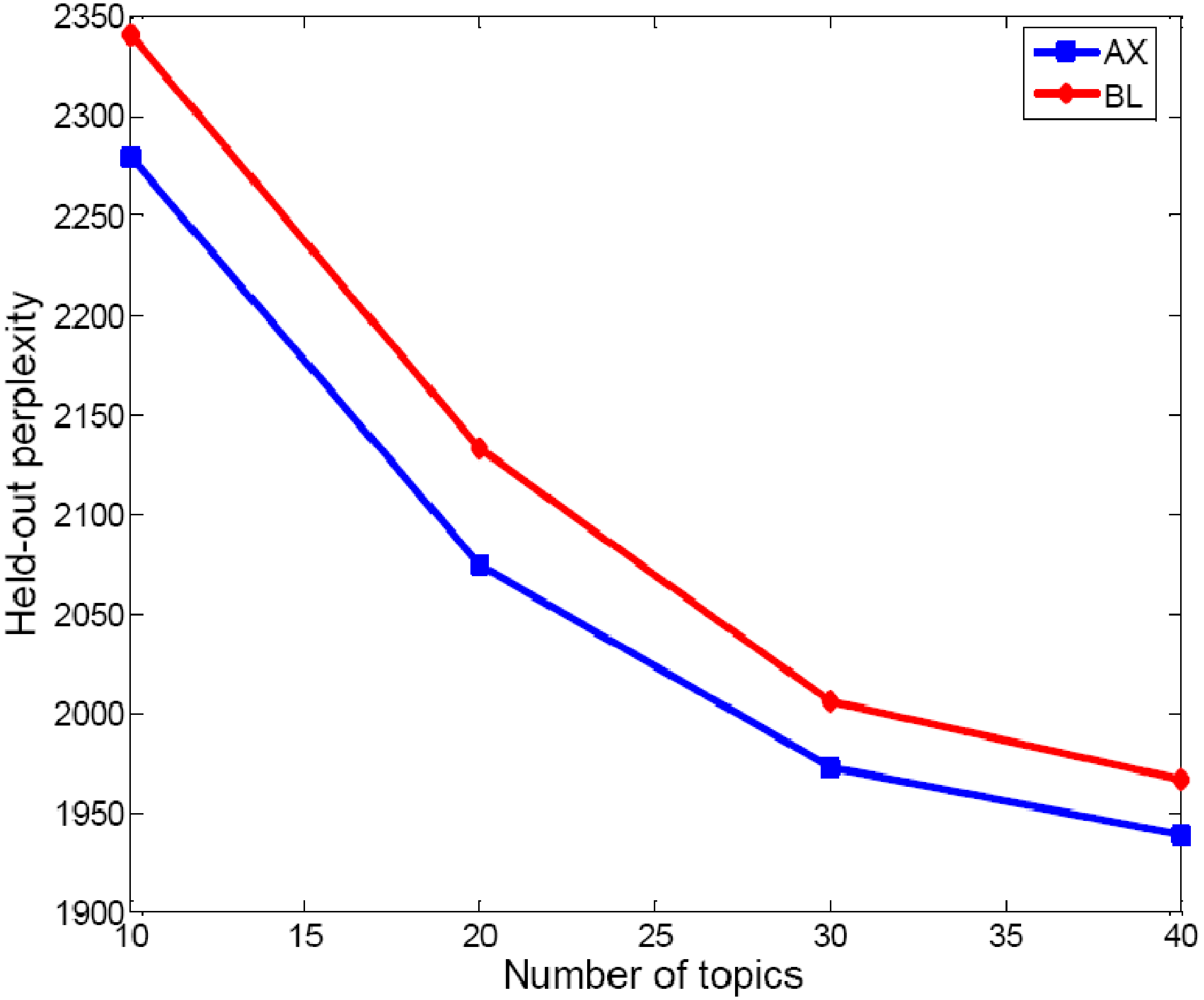
-Read more on that (Noah Smith & Jason Eisner ACL2004, COLLING-ACL2006)

Result on NIPS collection

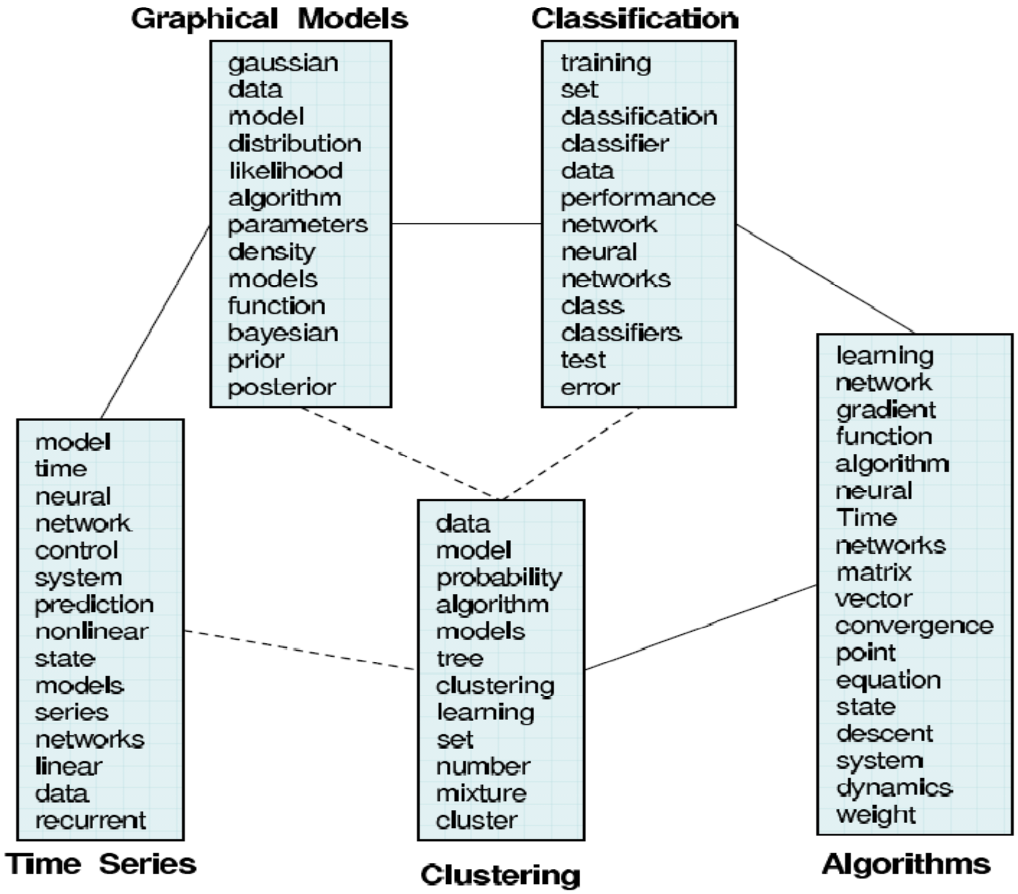
- NIPS proceeding from 1988-2003
- 14036 words
- 2484 docs
- 80% for training and 20% for testing
- Fit both models with 10,20,30,40 topics
- Compare **perplexity** on held out data
  - The perplexity of a language model with respect to text x is the reciprocal of the geometric average of the probabilities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

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

Comparison: perplexity



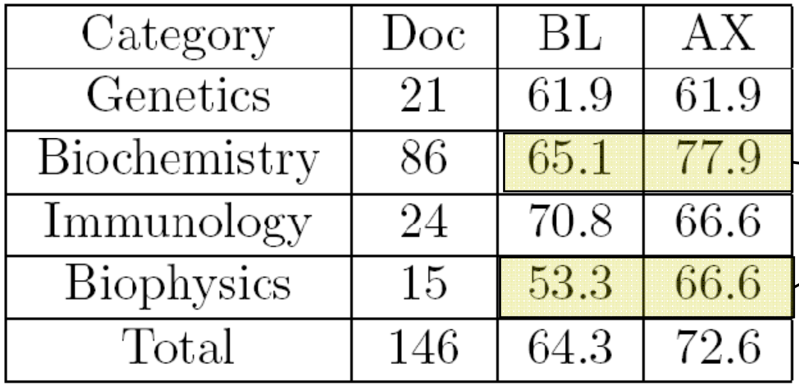
Topics and topic graphs



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

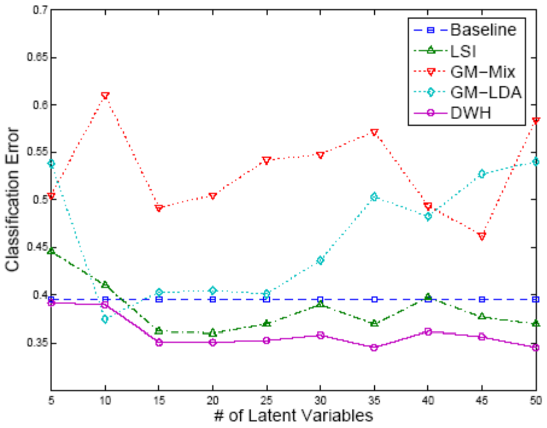
-Notable Difference

-Examine the low dimensional representations below

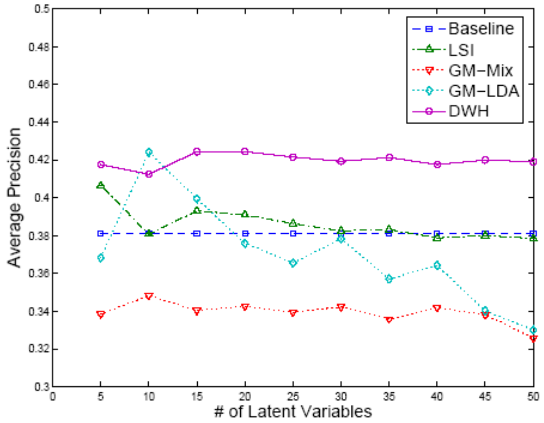
Are we done?

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

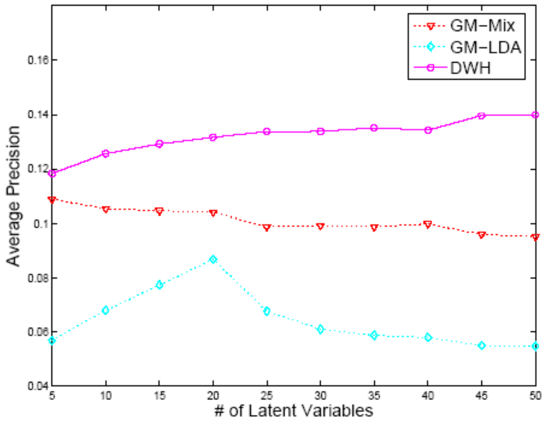
Some shocking results on LDA



Classification



Retrieval

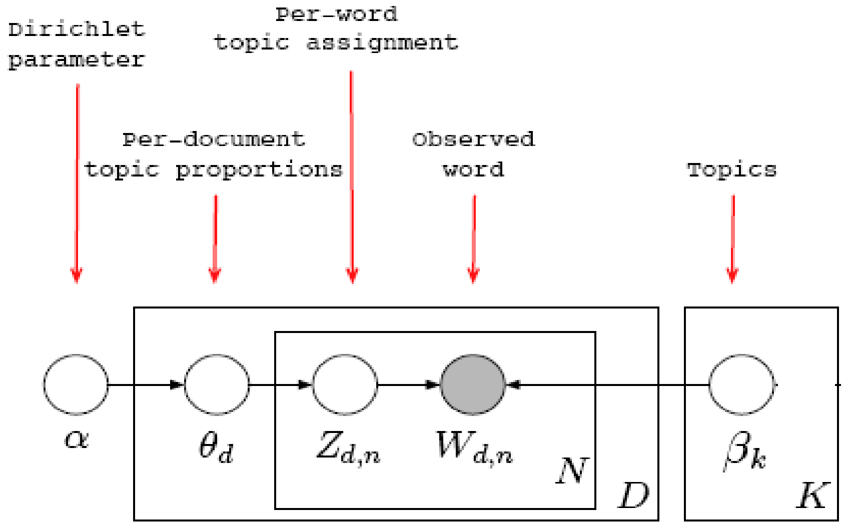


Annotation

- 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



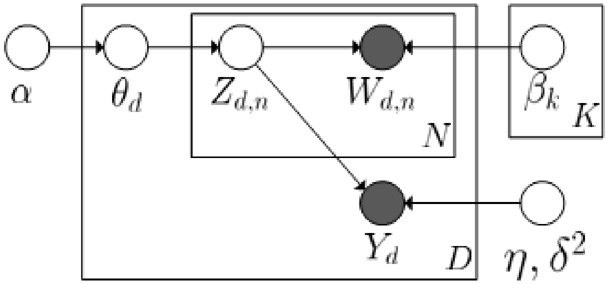
$$\overbrace{\begin{matrix}\theta_1 & \ldots & \theta_D \\ 0.8 & \ldots & 0.3 \\ 0.2 & \ldots & 0.7\end{matrix}}^{\theta}$$

$$\beta : \begin{pmatrix} 0.70 & 0.05 & 0.03 & \ldots \\ 0.12 & 0.52 & 0.05 & \ldots \end{pmatrix}$$

Supervised Topic Model (sLDA)

- 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)



(Blei & McAuliffe, 2007)

- 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 $\begin{cases} y_d \sim \mathcal{N}(\eta^\top \bar{Z}_d, \delta^2) \\ y_d \sim \mathrm{GLM}(\bar{Z}_d, \eta^\top, \delta^2) \end{cases}$

Continuous (regression)

Discrete (classification)

- Joint distribution:

$$p(\theta, \mathbf{z}, \mathbf{y}, \mathbf{W} | \alpha, \beta, \eta, \delta^2) = \prod_{d=1}^{D} p(\theta_d|\alpha)(\prod_{n=1}^{N} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta)) p(y_d|\eta^\top \bar{z}_d, \delta^2)$$

- Variational inference:

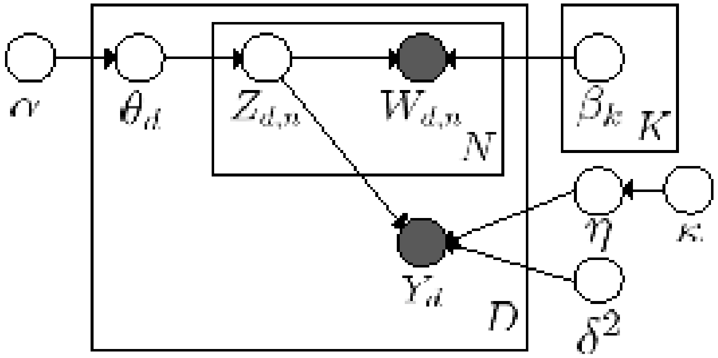
$$\mathcal{L}(q) \triangleq -E_q[\log p(\theta, \mathbf{z}, \mathbf{y}, \mathbf{W}|\alpha, \beta, \eta, \delta^2)] - \mathcal{H}(q(\mathbf{z}, \theta)) \geq -\log p(\mathbf{y}, \mathbf{W}|\alpha, \beta, \eta, \delta^2)$$

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

- Generative Procedure (Bayesian sLDA):
  - Sample a parameter $\eta \sim p_0(\eta)$
  - 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 : $y_d \sim \mathcal{N}(\eta^\top \bar{Z}_d, \delta^2)$



- Def:

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

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

predictive accuracy

model fitting

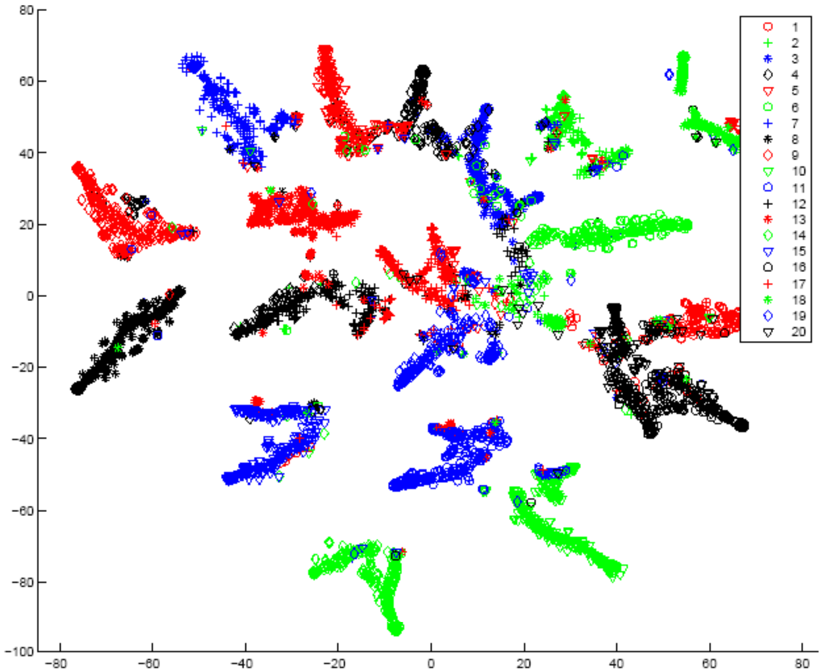
Experiments

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

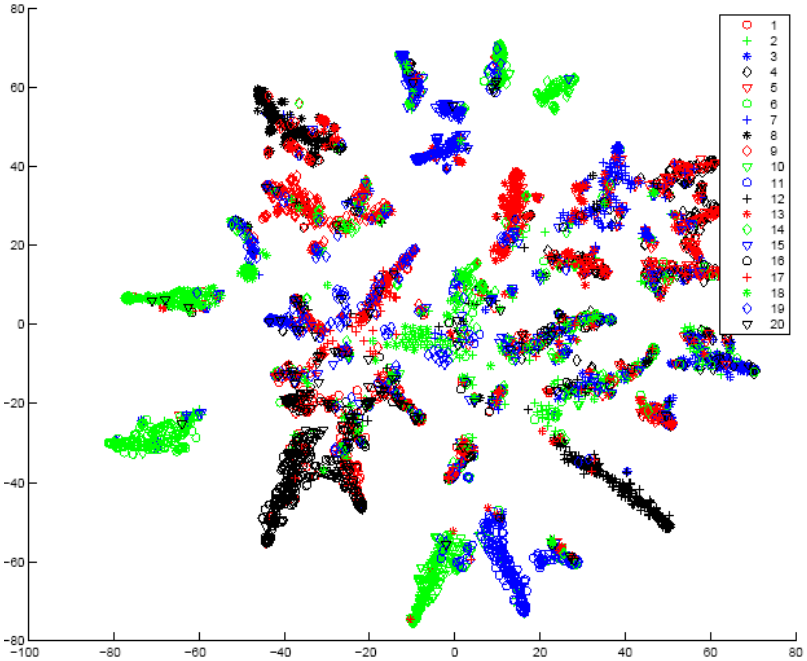
- Data Sets：
  - 20 Newsgroups
    - Documents from 20 categories
    - ~ 20,000 documents in each group
    - Remove stop word as listed in
  - Movie Review
    - 5006 documents, and 1.6M words
    - Dictionary: 5000 terms selected by tf-idf
    - Preprocessing to make the response approximately normal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

Document Modeling

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



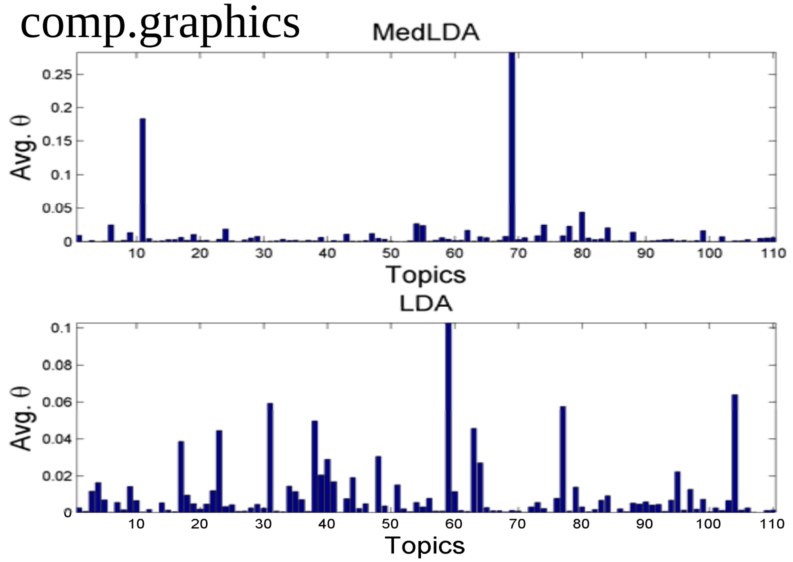
MedLDA



LDA

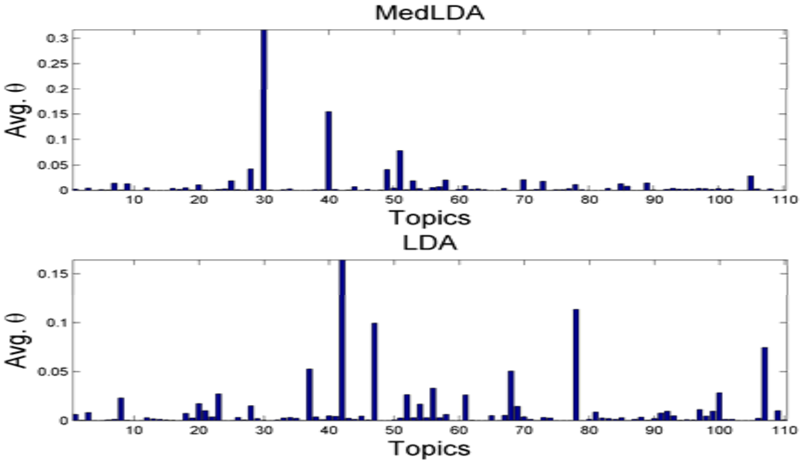
Document Modeling (cont')

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

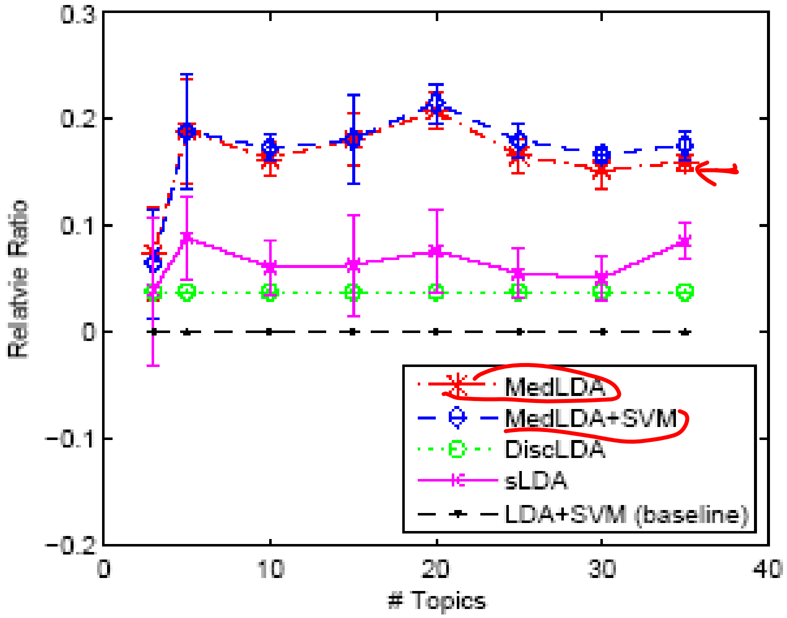


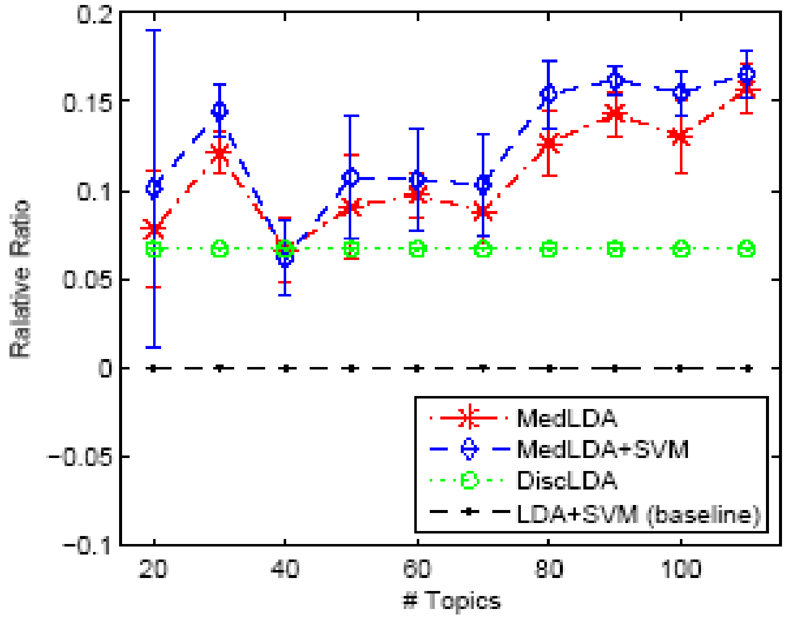
T 30	T 40	T 51	T 42	T 78	T 47
israel	turkish	israel	israel	jews	armenian
israeli	armenian	lebanese	israeli	jewish	turkish
jews	armenians	israeli	peace	israel	armenians
arab	armenia	lebanon	writes	israeli	armenia
writes	people	people	article	arab	turks
people	turks	attacks	arab	people	genocide
article	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

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

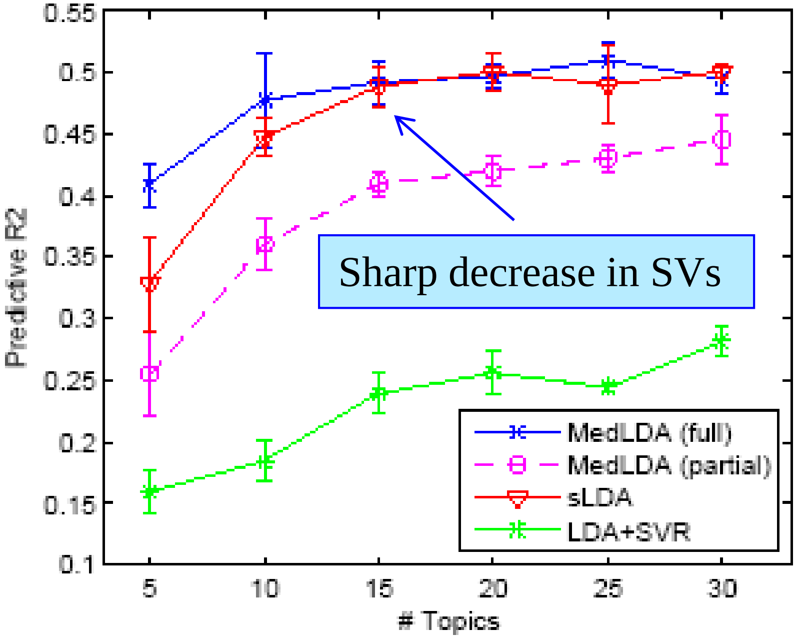




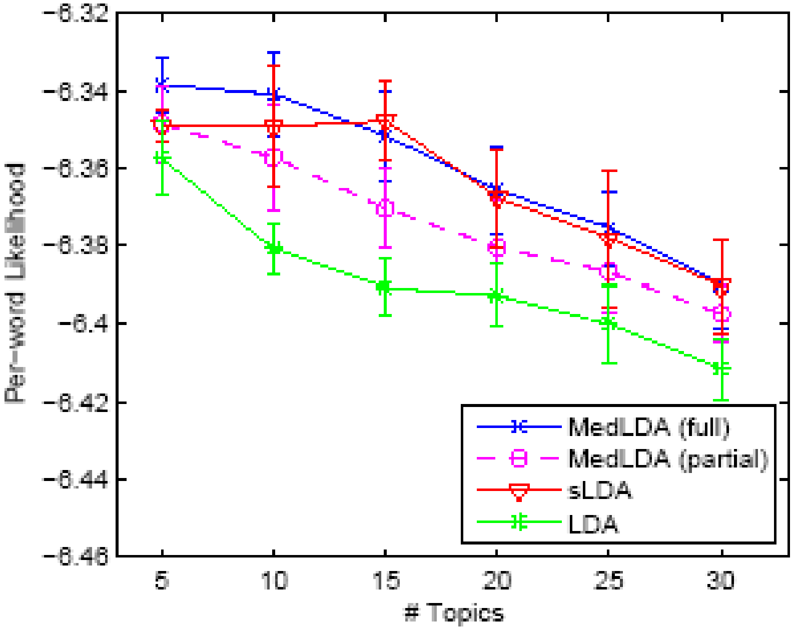
Regression

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

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

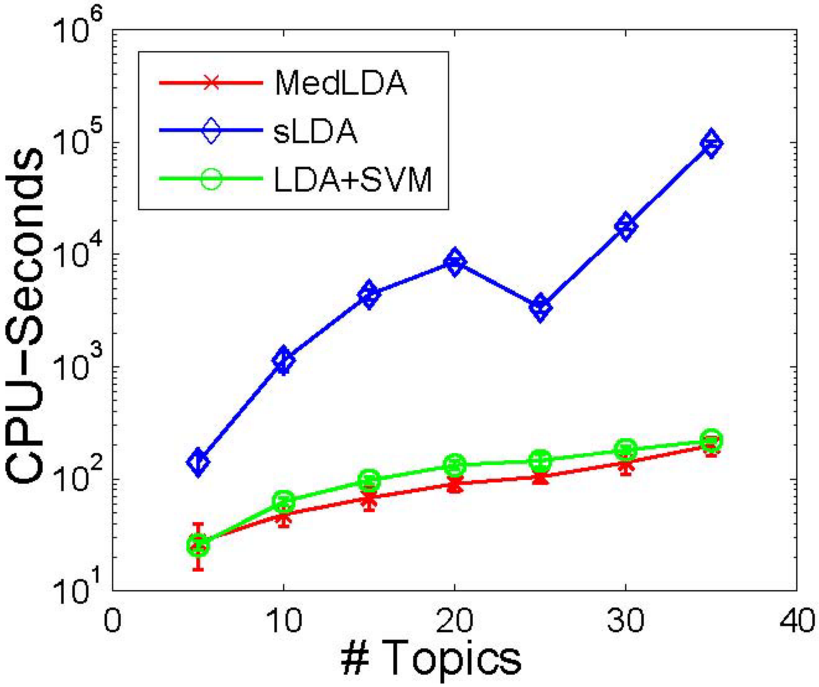


Sharp decrease in SVs



Time Efficiency

- Binary Classific



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

Finally, think about 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):

$$\mathrm{P(MedTM)}: \min_{q(H),q(\Upsilon),\Psi,\xi} \mathcal{L}(q(H)) + KL(q(\Upsilon)\|p_0(\Upsilon)) + U(\xi)$$

$$\text{s.t. } \textit{expected} \text{ margin constraints}$$

- 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

Summary

- **A 6-dimensional space of working with graphical models**
  - Task:
    - Embedding? Classification? Clustering? Topic extraction? …
  - Data representation:
    - Input and output (e.g., continuous, binary, counts, …)
  - Model:
    - BN? MRF? Regression? SVM?
  - Inference:
    - Exact inference? MCMC? Variational?
  - Learning:
    - MLE? MCLE? Max margin?
  - Evaluation:
    - Visualization? Human interpretability? Perperlexity? Predictive accuracy?
- **It is better to consider one element at a time!**