Unsupervised Learning Of Finite Mixture Models With Deterministic Annealing For Large-scale Data Analysis

Thesis Defense, January 12, 2012

Student: **Jong Youl Choi**

Advisor: **Geoffrey Fox**

School of Informatics and Computing

Pervasive Technology Institute

Indiana University

INDIANA UNIVERSITY

PERVASIVE TECHNOLOGY INSTITUTE

SALSA HPC

SALSA project

http://salsahpc.indiana.edu

Outline

I. Finite Mixture Model (FMM)

II. Model Fitting with Deterministic Annealing (DA)

III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)

IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)

V. Conclusion And Future Work

I. Finite Mixture Model (FMM)

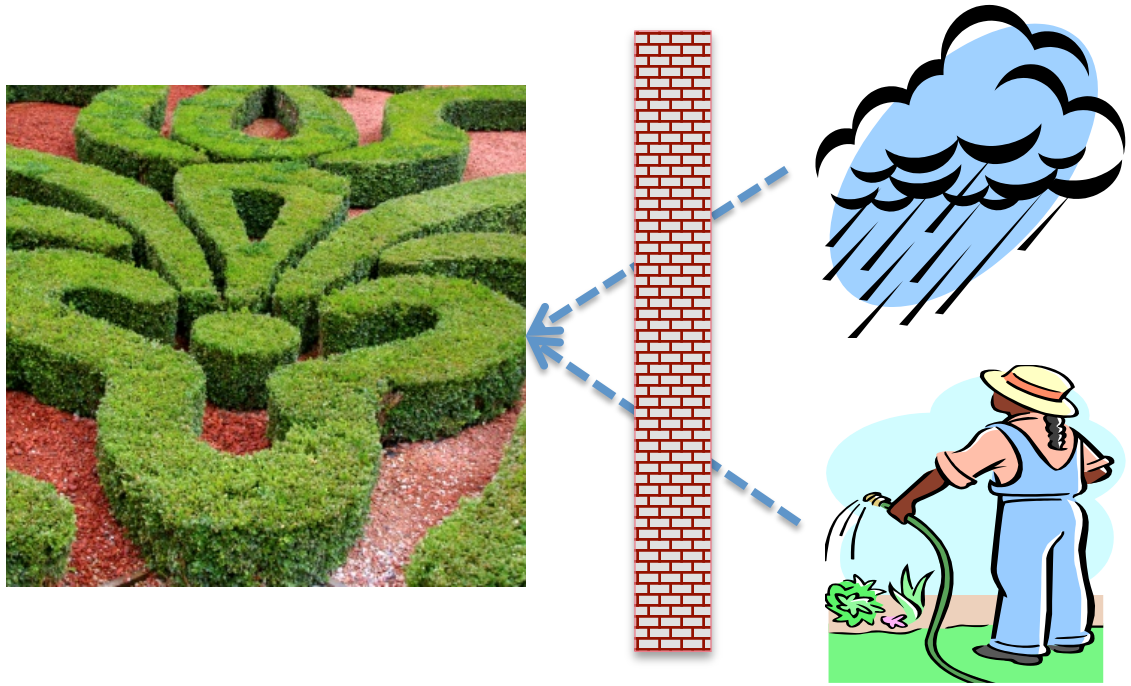
II. Model Fitting with Deterministic Annealing (DA)

III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)

IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)

V. Conclusion And Future Work

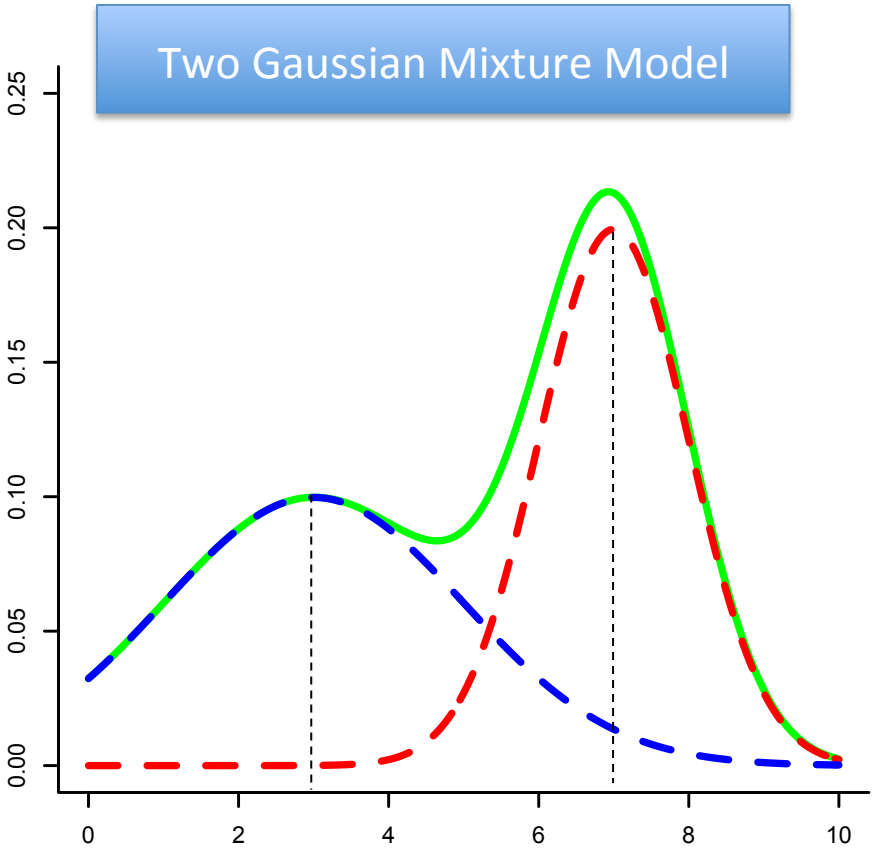
Machine Learning Problem



Observations Hidden Components

- Learning from observed data
- Inferring a data generating process
- Essential structure, abstract, summary, …

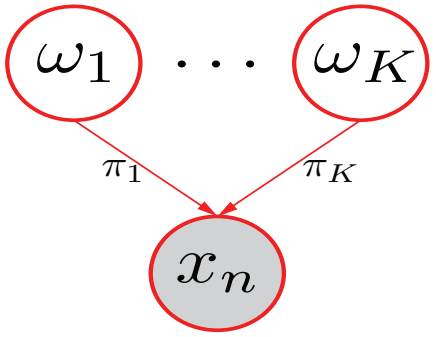
Finite Mixture Model (FMM)



- Component Model
  - A mixture of simple distributions (components)
  - Hidden or latent components
  - Serve as abstract or summary of data
- Generative Model
  - Simulate observed random sample
- Convenient and flexible
- Model fitting is hard
  - Too many parameters

Two Finite Mixture Models

FMM Type-1

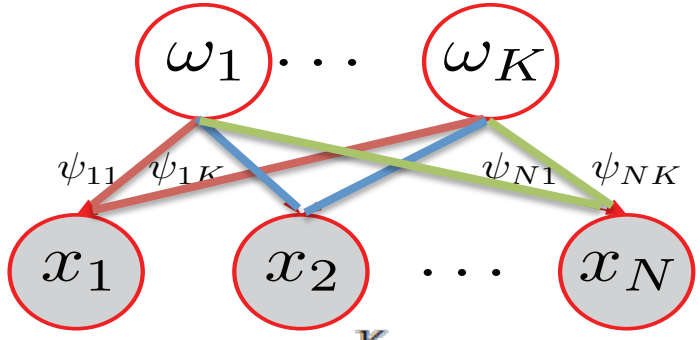


$$P(x_i \mid \Omega, \pi) = \sum_{k=1}^{K} \pi_k P(x_i \mid \omega_k)$$

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

- Traditional model
- Component ≈ Cluster
- Clustering, Gaussian Mixture (GM), GTM, …

FMM Type-2



$$P(x_i \mid \Omega, \Psi) = \sum_{k=1}^{K} \psi_{ik} P(x_i \mid \omega_k)$$

$$\sum_{k=1}^{K} \psi_{ik} = 1$$

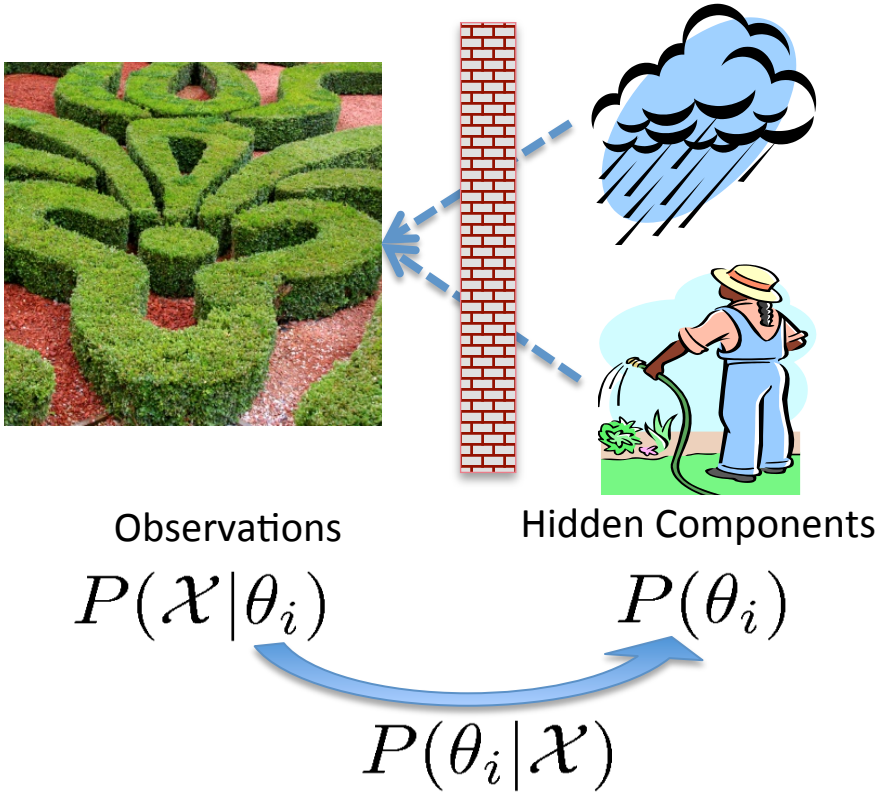
- Factor model
- Component ≈ Generator
- PLSA

Model Fitting

- Bayes' Theorem

$$P(\theta_i|\mathcal{X}) = \frac{P(\mathcal{X}|\theta_i)P(\theta_i)}{P(\mathcal{X})}$$

  - θ_i: Parameters
  - X : Observations
  - $P(\theta_i)$: Prior
  - $P(X \mid \theta_i)$: Likelihood



- Maximum Likelihood Estimator (MLE)
  - Used to find the most plausible θ, given X
  - Maximize likelihood or log-likelihood

    ➔ Optimization problem

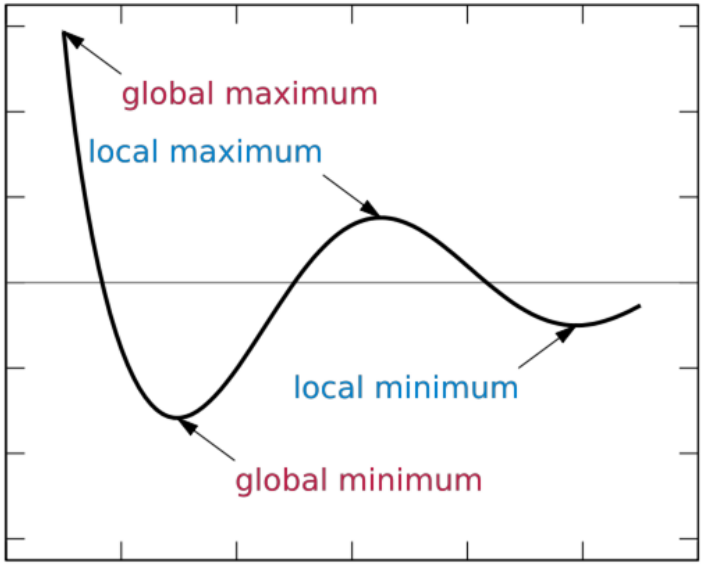
Expectation-Maximization (EM) Algorithm

- Problems in MLE
  - Observation X is often not complete
  - Latent (hidden) variable Z exists
  - Hard to explore whole parameter space

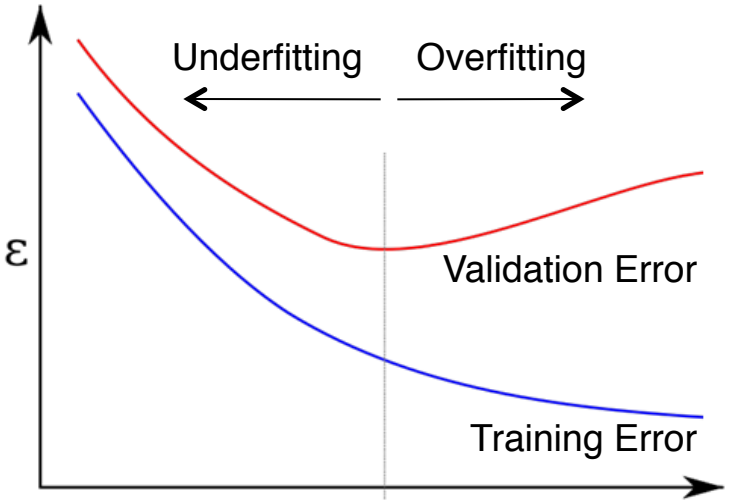
- EM algorithm
  - Random initialization θ^{old}
  - E-step : Expectation P(Z I X, θ^{old})
  - M-step : Maximize (log-)likelihood
  - Repeat E-,M-step until converge.

Motivation

- Local optimum problem
  - Easily trap in the local optimum
  - Sensitive to initial conditions or parameters
  - High-variance solution
  - SA, GA, …
- Overfitting problem
  - Poor generalization quality
  - Directly related with predicting power
  - Early stopping, cross validation, …



(Maxima and Minima, Wikipedia)



(Overfitting, Wikipedia)

Contributions

- Solve FMM with DA
  - Avoid local optimum problem
  - Avoid overfitting problem
- Present DA applications
  - GTM with DA (DA-GTM)
  - PLSA with DA (DA-PLSA)
- Experimental results
  - Data visualization
  - Text mining

I. Finite Mixture Model (FMM)

II. Model Fitting with Deterministic Annealing (DA)

III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)

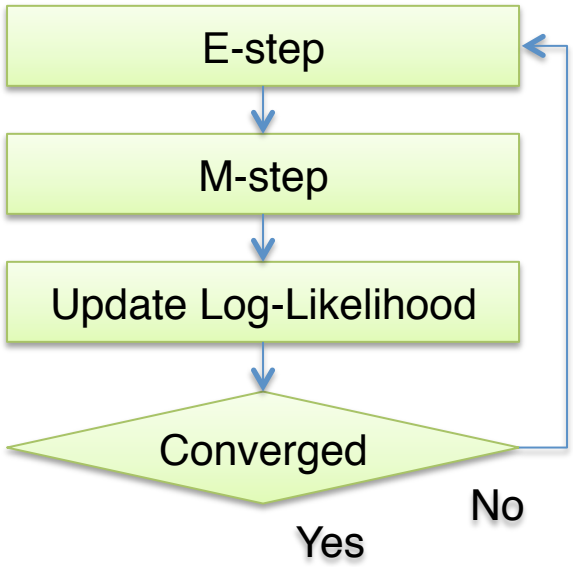
IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)

V. Conclusion And Future Work

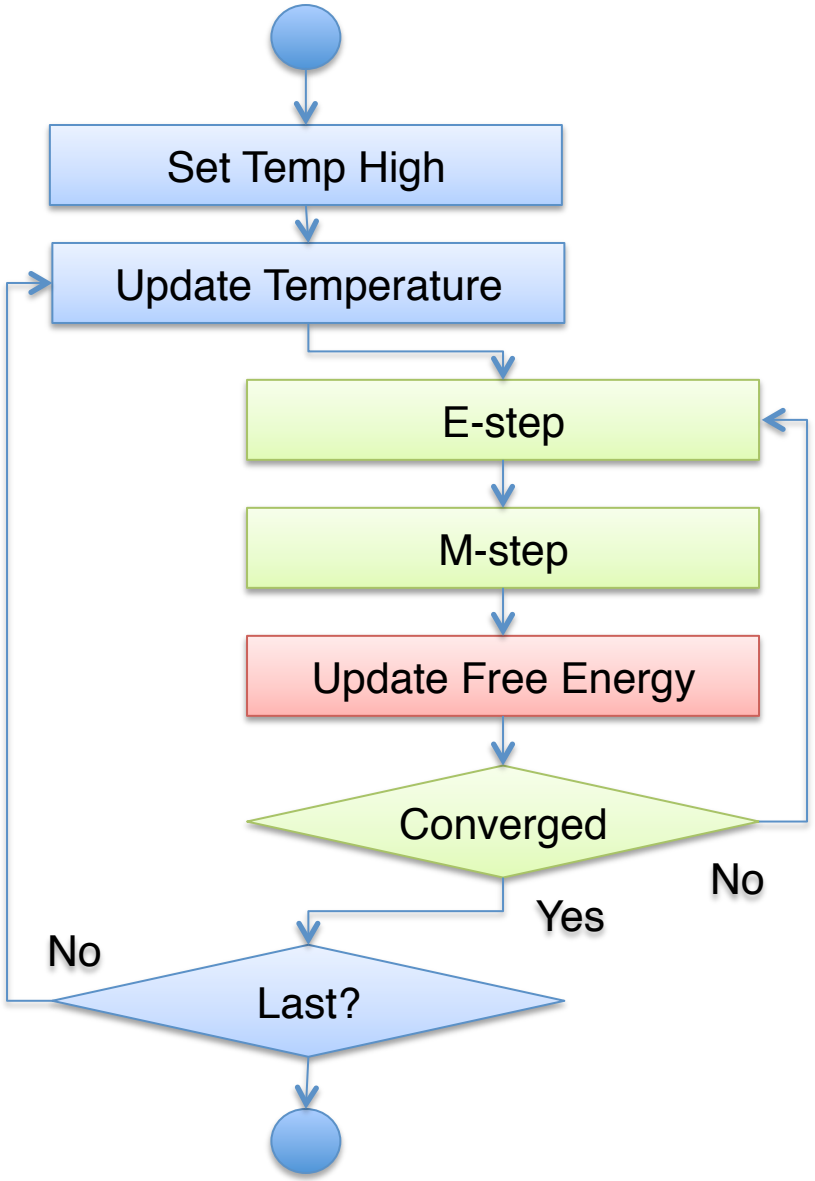
Deterministic Annealing (DA)

- Optimization
  - Gradually lowering numeric temperature
  - No stochastic process
- Local optimum avoidance
  - Tracing the global solution by changing level of smoothness
  - Smoothed → bumpy
- Principle of Maximum Entropy
  - A solution with maximum entropy
  - Minimize the free energy *F* of log-likelihood
  - Eventually, we will have maximized log-likelihood

Finite Mixture Model with EM



Finite Mixture Model with DA



- Annealing (High → Low)
- High temperature
  - Soft (or fuzzy) association
  - Smooth cost function
- Low temperature
  - Hard association
  - Bumpy cost function
  - Revealing full complexity

- Minimize free energy
- Free energy F = *f*(Temp, Entropy of Log-Likelihood)

Free Energy for Finite Mixture Model

- Free Energy

$$F = D - TS$$
$$= -T \sum_{n=1}^{N} \ln Z_n$$

  - D : expected cost <d_{nk}>
  - S : Shannon entropy
  - T : computational temperature
  - Z_n : partition function

$$Z_n = \sum_{k=1}^{K} \exp\left(\frac{-d_{nk}}{T}\right)$$

- General form for Finite Mixture Model

$$F_{FMM} = -T \sum_{n=1}^{N} \log \sum_{k=1}^{K} \{c(n,k)\ p(x_n|y_k)\}^{\frac{1}{T}}$$

  - Cost function: $d_{nk} = -\log p(x_n|y_k)$

I. Finite Mixture Model (FMM)

II. Model Fitting with Deterministic Annealing (DA)

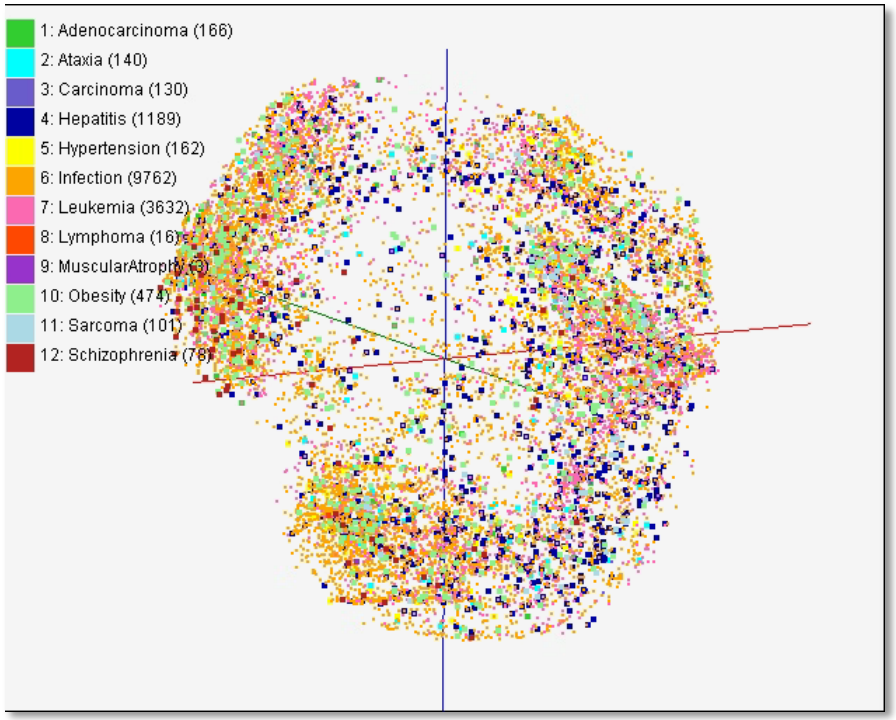
III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)

IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)

V. Conclusion And Future Work

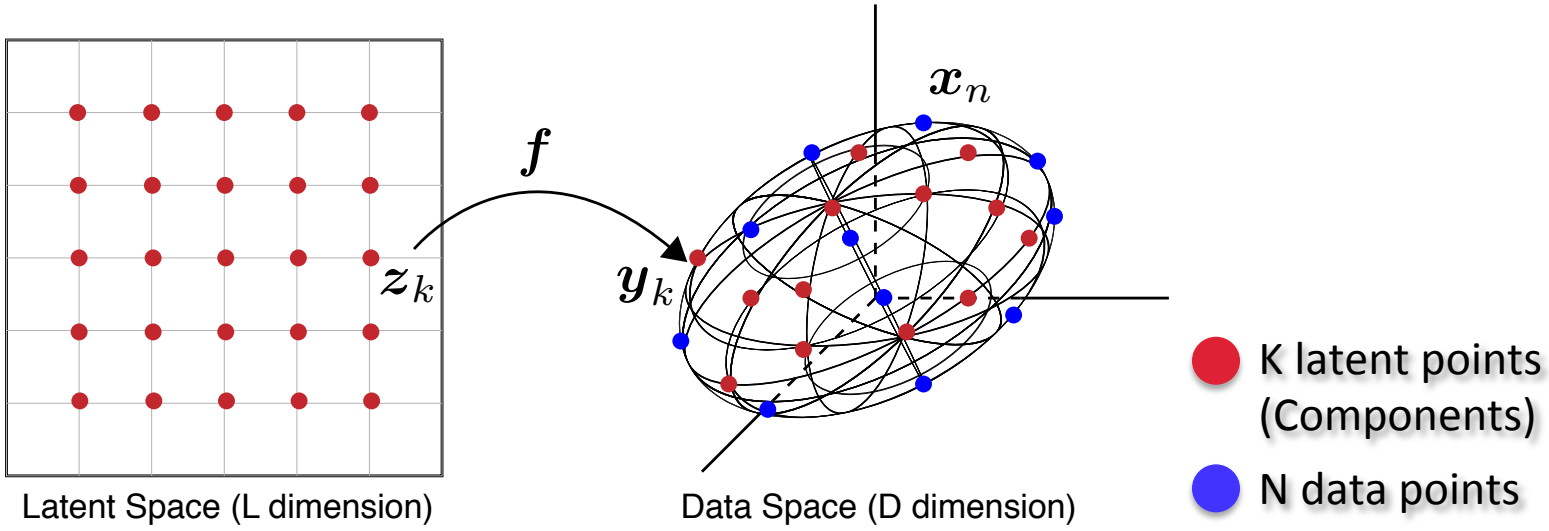
Dimension Reduction

PubChem Data
(166 dimensions)



- Simplification, feature selection/extraction, visualization, etc.
- Preserve the original data's information as much as possible in lower dimension

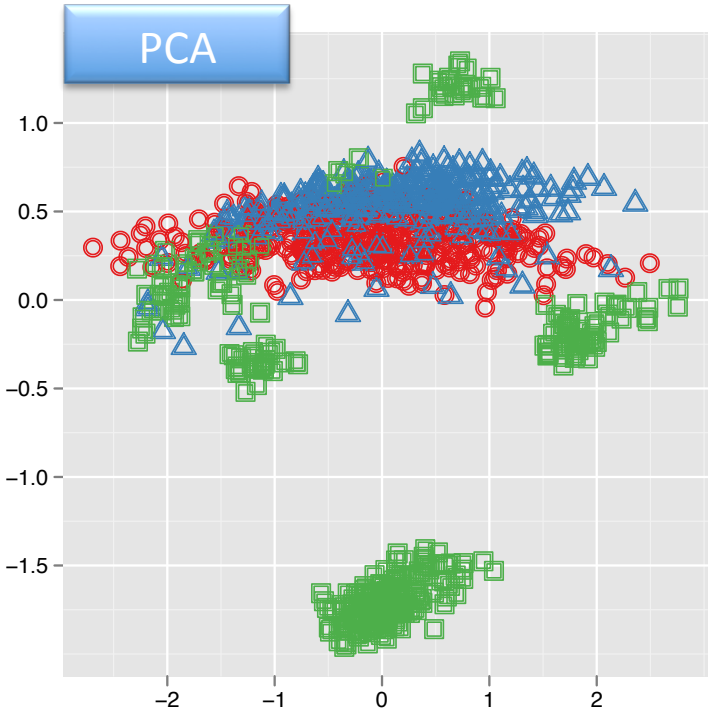
Generative Topographic Mapping

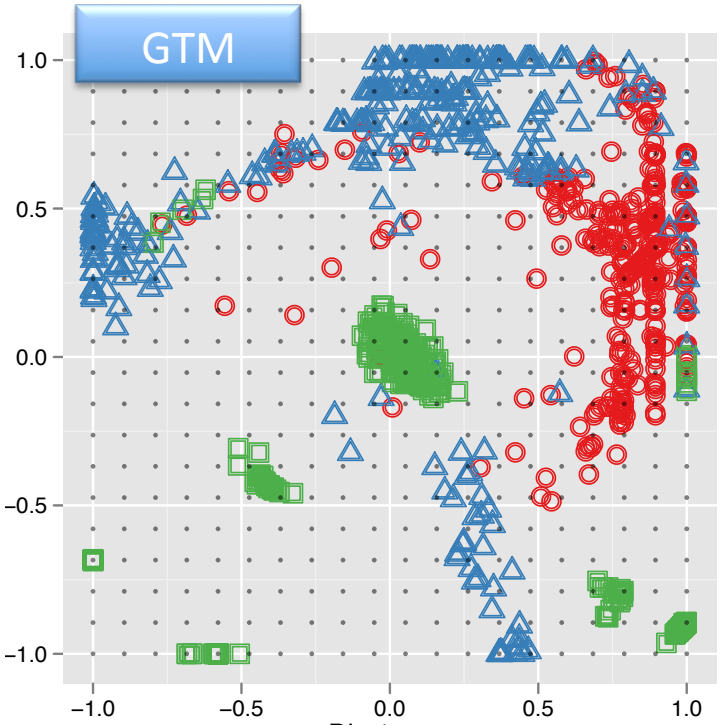


- An algorithm for dimension reduction
  - Find an optimal K latent variables in a latent space
  - *f* is a non-linear mappings
  - Strict Gaussian mixture model
  - EM model fitting
- DA optimization can improve the fitting process

Advantages of GTM

- Computational complexity is *O*(KN), where
  - N is the number of data points
  - K is the number of latent variables or *clusters*. K << N
- Efficient, compared with MDS which is *O*(N^2)
- Produce more separable map (right) than PCA (left)





Oil flow data
-. 1000 points
-. 12 Dimensions
-. 3 Clusters

Free Energy for GTM

$$F_{FMM} = -T \sum_{n=1}^{N} \log \sum_{k=1}^{K} \{c(n,k)\ p(x_n|y_k)\}^{\frac{1}{T}}$$

- GTM Model Setting
  - Strict Gaussian assumption

$$p(x_n|y_k) = \mathcal{N}(x_n|\mu_k, \sigma_k)$$

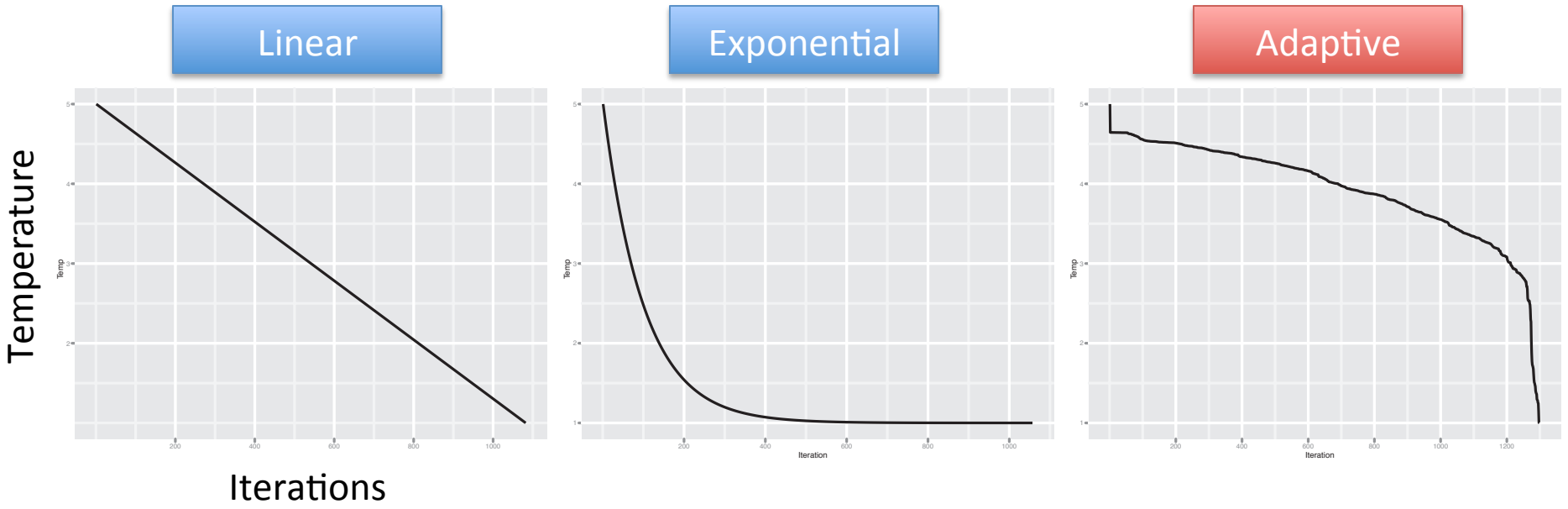
  - Constant mixing weight

$$c(n,k) = \frac{1}{K}$$

GTM with Deterministic Annealing

	EM-GTM (Traditional method)	DA-GTM (New algorithm)
Optimization	Maximize log-likelihood *L*	Minimize free energy *F*
Objective Function	$\sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{k=1}^{K} p(\boldsymbol{x}_n \mid \boldsymbol{y}_k) \right\}$	$-T \sum_{n=1}^{N} \ln \left\{ \left(\frac{1}{K} \right)^{\frac{1}{T}} \sum_{k=1}^{K} p(\boldsymbol{x}_n \mid \boldsymbol{y}_k)^{\frac{1}{T}} \right\}$
	When $T = 1$, $L = -F$.	
Pros & Cons	▪ Very sensitive to parameters ▪ Trapped in local optima ▪ Faster	▪ Less sensitive to poor parameters ▪ Avoid local optimum ▪ Require more computational time

Cooling Schedules



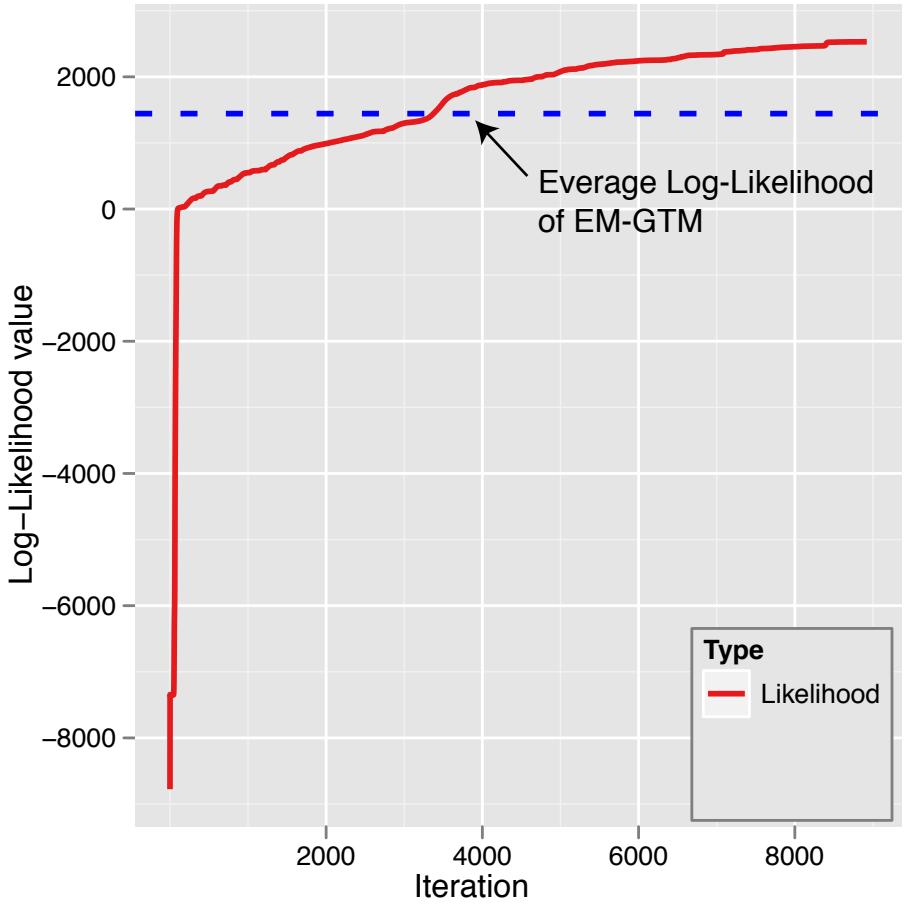
- Traditional method : static cooling schedule
- Adaptive cooling, a dynamic cooling schedule
  - Able to adjust the problem on the fly
  - Move to a temperature at which *F* may change

Phase Transition

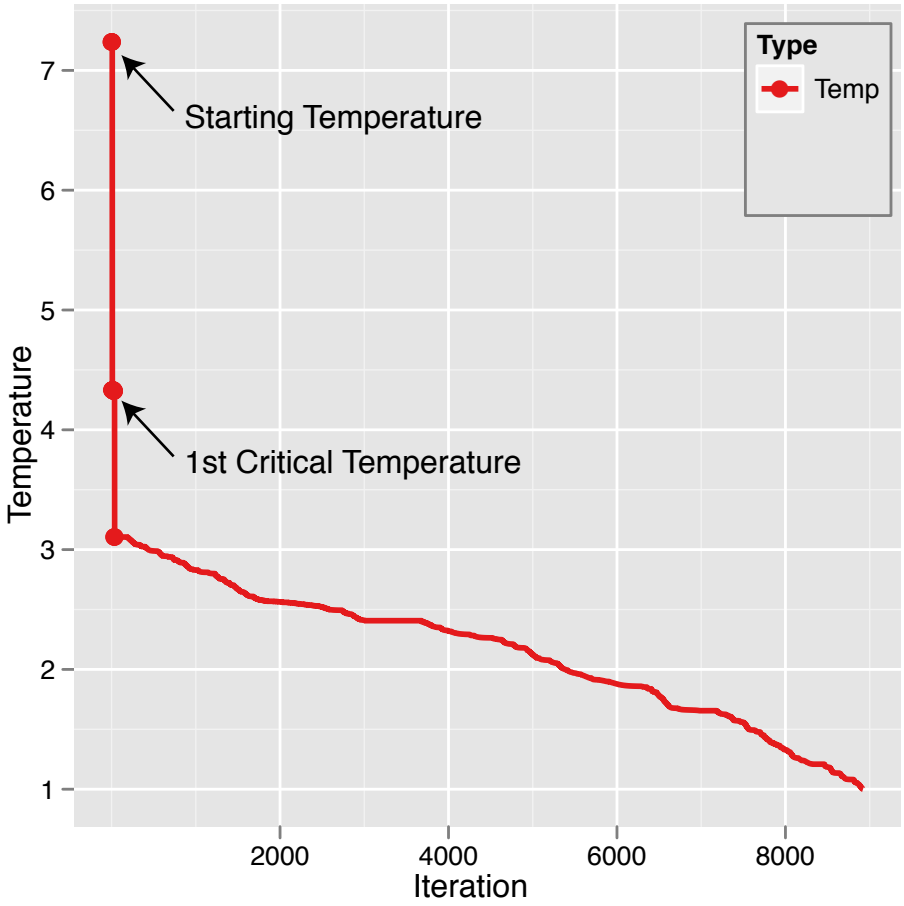
- Discrete behavior of DA
  - In some temperatures, the free energy is stable
  - At a specific temperature, start to explode, which is known as critical temperature T_c
- Critical temperature T_c
  - Free energy F is drastically changing at T_c
  - Second derivative test : Hessian matrix loose its positive definiteness at T_c
  - det (H) = 0 at T_c, where

$$H = \begin{bmatrix} H_{11} & \cdots & H_{1K} \\ \vdots & & \vdots \\ H_{K1} & \cdots & H_{KK} \end{bmatrix} \qquad H_{kk} = \frac{\partial^2 F}{\partial \boldsymbol{y}_k \partial \boldsymbol{y}_k^{\mathsf{Tr}}} \qquad H_{kk'} = \frac{\partial^2 F}{\partial \boldsymbol{y}_k \partial \boldsymbol{y}_{k'}^{\mathsf{Tr}}}$$

DA-GTM with Adaptive Cooling



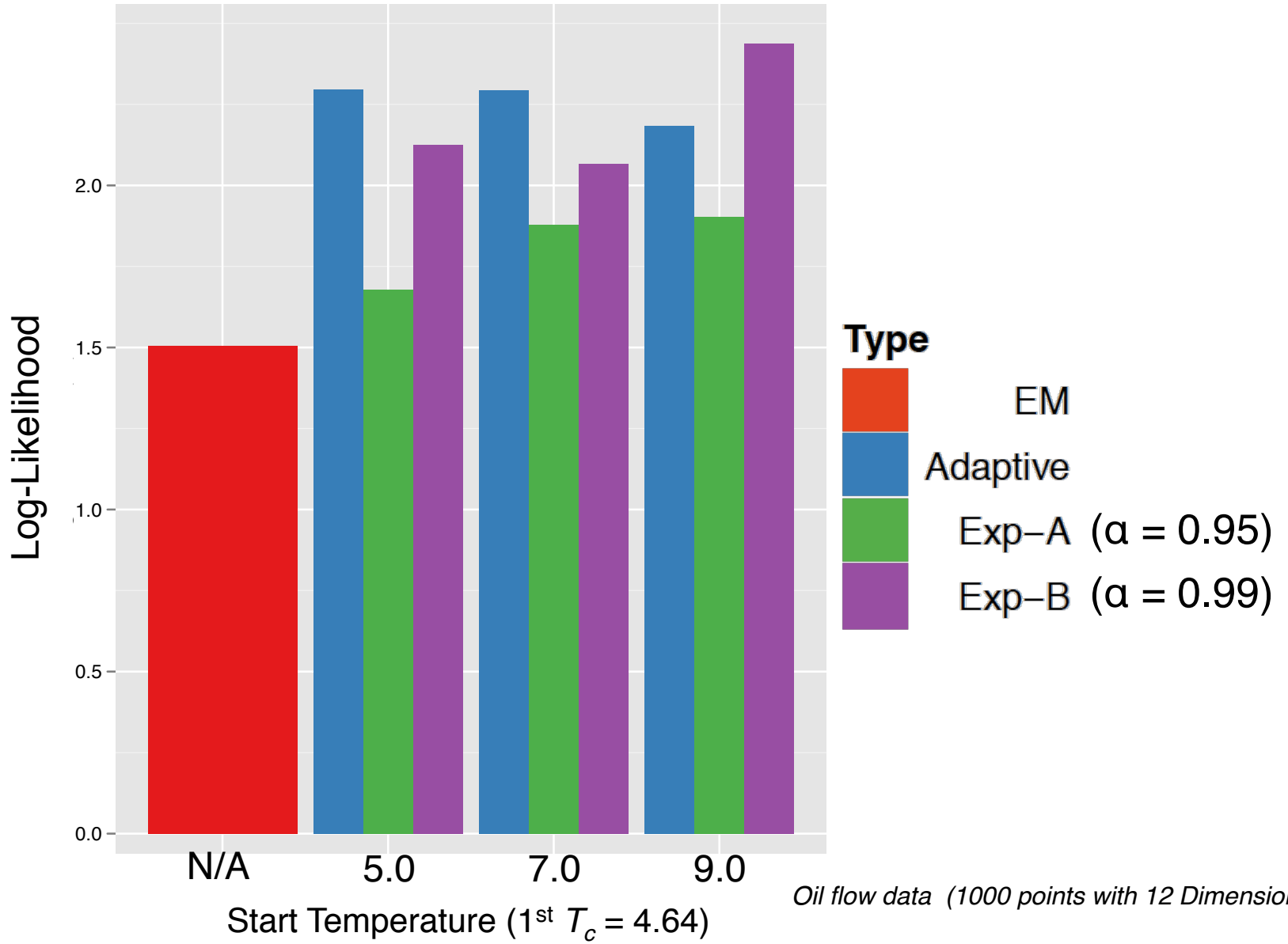
Progress of log-likelihood



Adaptive changes in cooling schedule

Oil flow data (1000 points with 12 Dimensions)

DA-GTM Result



Oil flow data (1000 points with 12 Dimensions)

Conclusion

- GTM with Deterministic Annealing (DA-GTM)
  - Overcome short-comes of traditional EM method
  - Avoid local optimum
  - Robust against poor initial parameters
- Phase-transitions in DA-GTM
  - Use Hessian matrix for detection
  - Eigenvalue computation
- Adaptive cooling schedule
  - New convergence approach
  - Dynamically determine next convergence point

I. Finite Mixture Model (FMM)

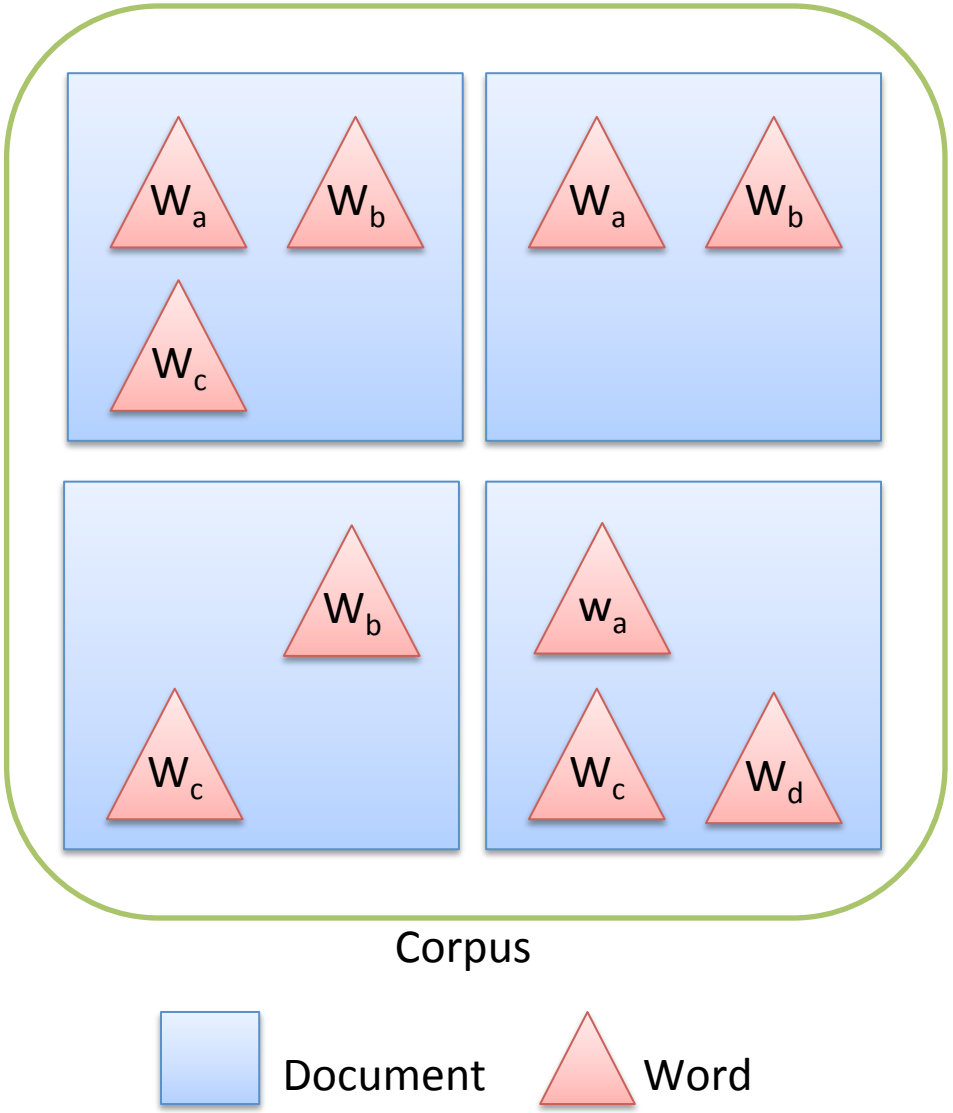
II. Model Fitting with Deterministic Annealing (DA)

III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)

IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)

V. Conclusion And Future Work

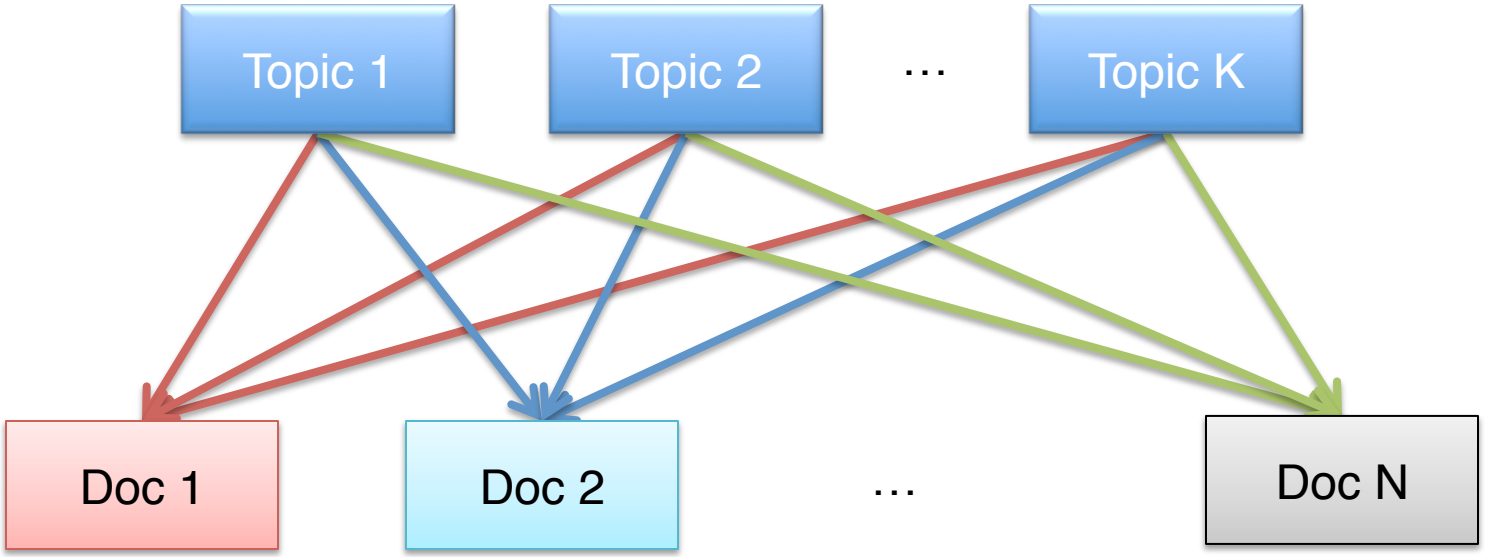
Corpus Analysis



- Polysems
  - A word with multiple meanings
  - E.g., ‘thread’
- Synonyms
  - Different words that have similar meaning, a topic
  - E.g., ‘car’ and ‘automotive’

Probabilistic Latent Semantic Analysis (PLSA)

- Topic model
  - Assume latent K topics generating words
  - Each document is a mixture of K topics
- FMM Type-2
  - The original proposal used EM for model fitting



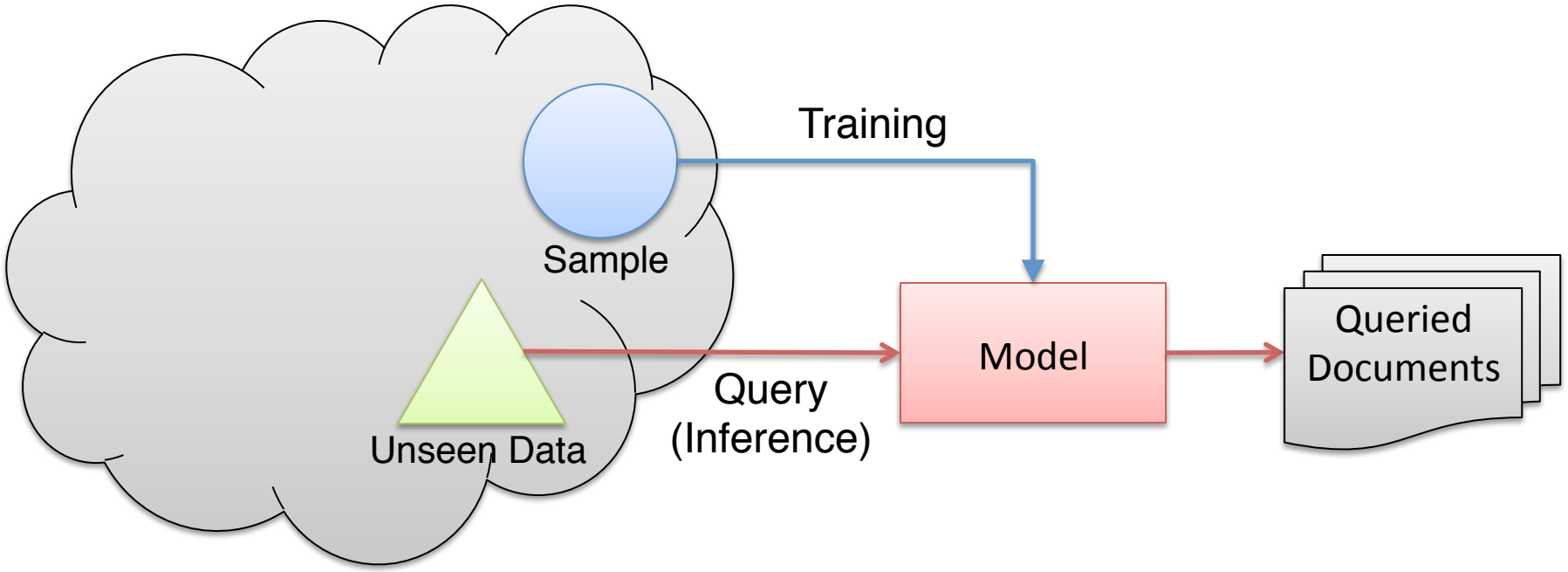
An Example of DA-PLSA

(AP: 2,246 documents & 10,473 words)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
percent	stock	soviet	bush	percent
million	market	gorbachev	dukakis	computer
year	index	party	percent	aids
sales	million	i	i	year
billion	percent	president	jackson	new
new	stocks	union	campaign	drug
company	trading	gorbachevs	poll	virus
last	shares	government	president	futures
corp	new	new	new	people
share	exchange	news	israel	two

Top 10 list of the best words of the AP news dataset for 30 topics.
Processed by DA-PLSA and shown only 5 topics among 30 topics

Overfitting Problem



- Predictive power
  - Maintain good performance on unseen data
  - A generalized model is preferable

Free Energy for PLSA

$$F_{FMM} = -T \sum_{n=1}^{N} \log \sum_{k=1}^{K} \{c(n,k)\ p(x_n|y_k)\}^{\frac{1}{T}}$$

- PLSA Model Setting
  - Use Multinomial distribution

$$p(x_n|y_k) = \mathrm{Multi}(x_n|\theta_k)$$

  - Flexible mixing weight

$$c(n,k) = \psi_{nk}$$

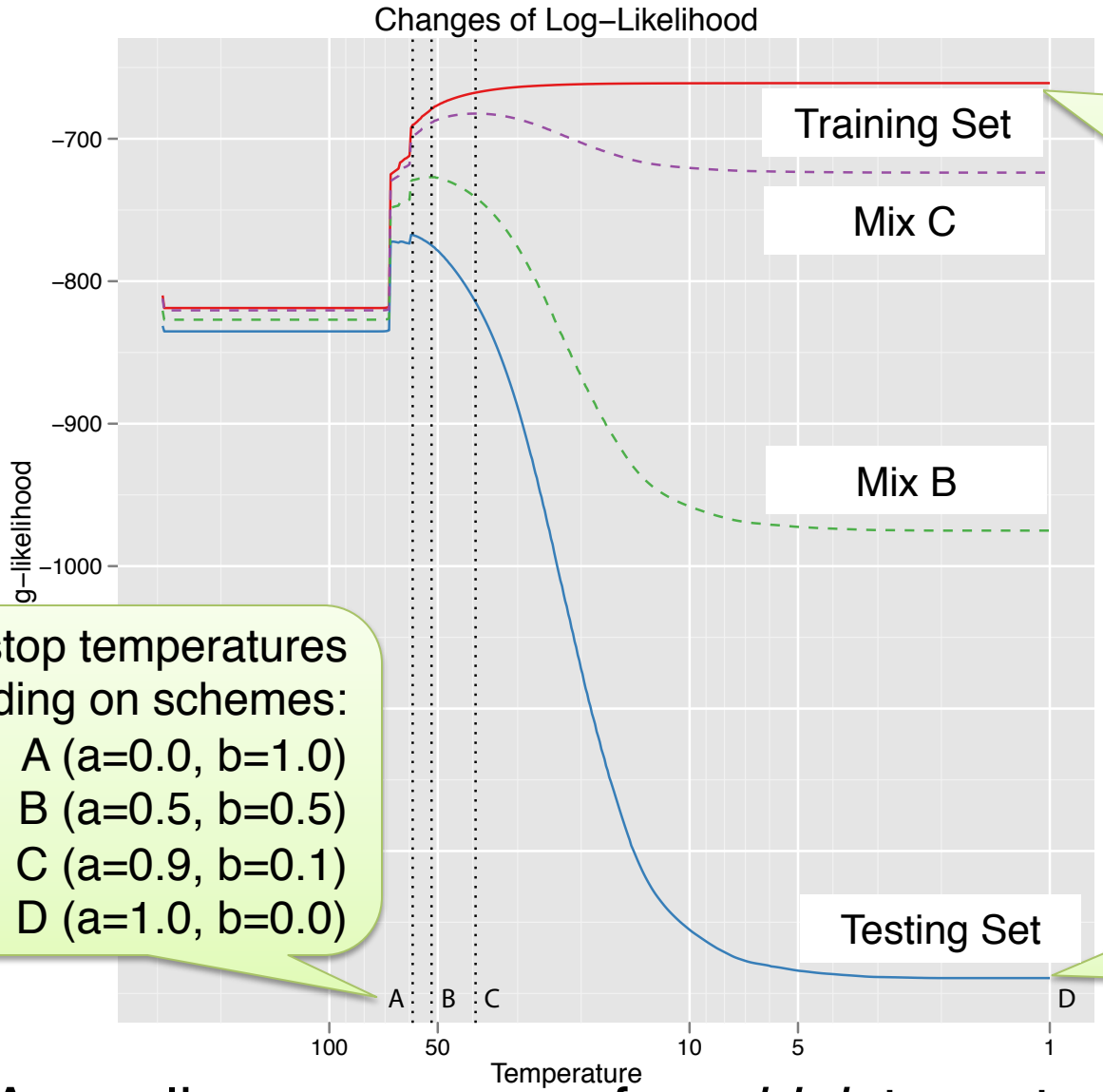
Overfitting Avoidance in DA

- DA can control smoothness
  - Smoothed solution at high temperature
  - Getting specific as annealing
  - Early stopping to get a smoothed (general) model
- Stop condition
  - Use a V-fold cross validation method
  - Measure total perplexity, sum of log-likelihood of both training set and testing set

$$Total\ Perplexity = a \cdot \mathcal{L}_{PLSA}(X_{training}, \Theta, \Psi) + b \cdot \mathcal{L}_{PLSA}(X_{testing}, \Theta, \Psi)$$

- Tempered-EM, proposed by Hofmann (the original author of PLSA), but annealing is done in a reversed way

Annealing in DA-PLSA



Improved fitting quality with training set during annealing

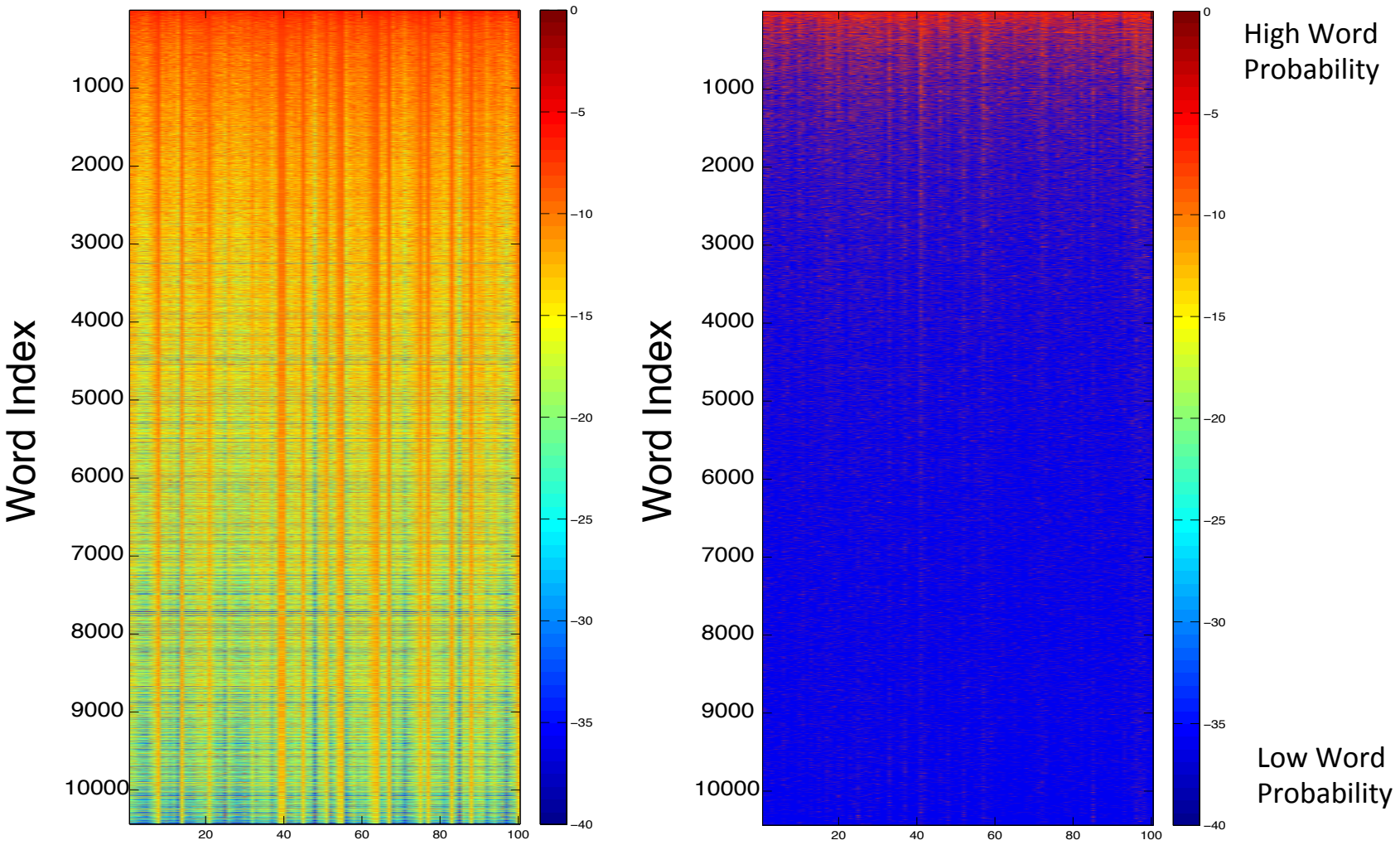
Early-stop temperatures depending on schemes:
A (a=0.0, b=1.0)
B (a=0.5, b=0.5)
C (a=0.9, b=0.1)
D (a=1.0, b=0.0)

Over-fitting at Temp=1

Annealing progresses from *high* temp to *low* temp

Predicting Power in DA-PLSA

Log of word probabilities of AP data (100 topics for 10,473 words)



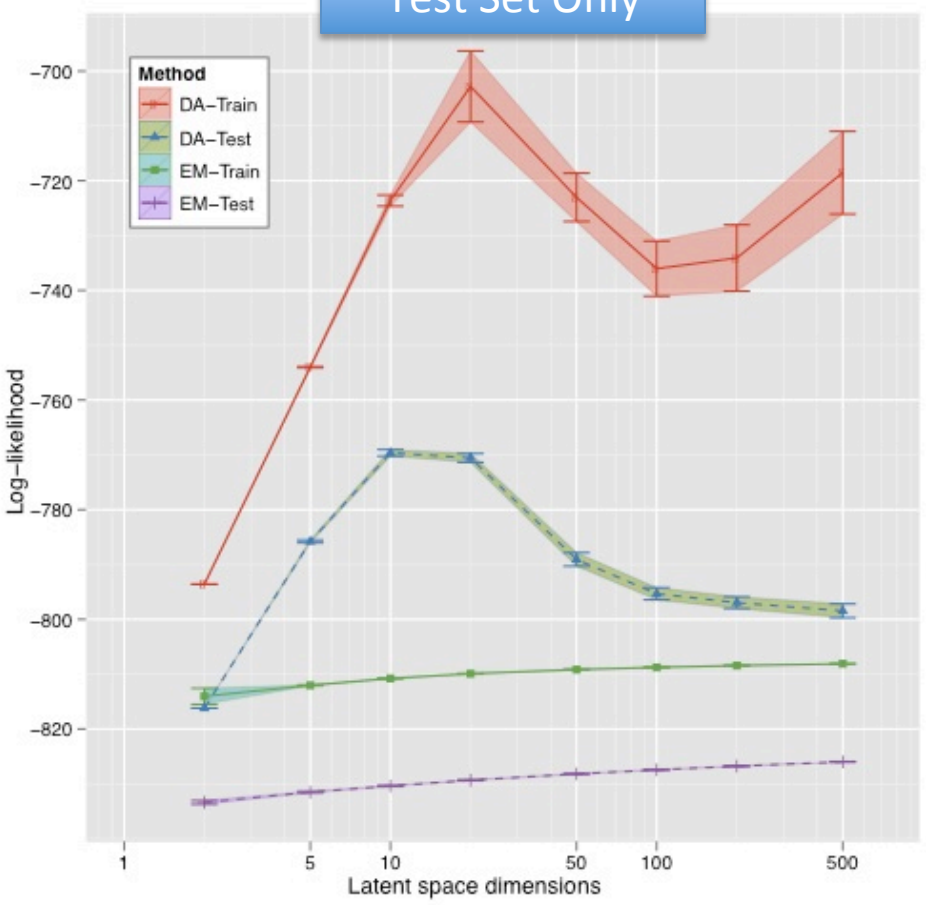
Early stop
(Temp = 49.98)

Over-fitting
(Temp = 1.0)

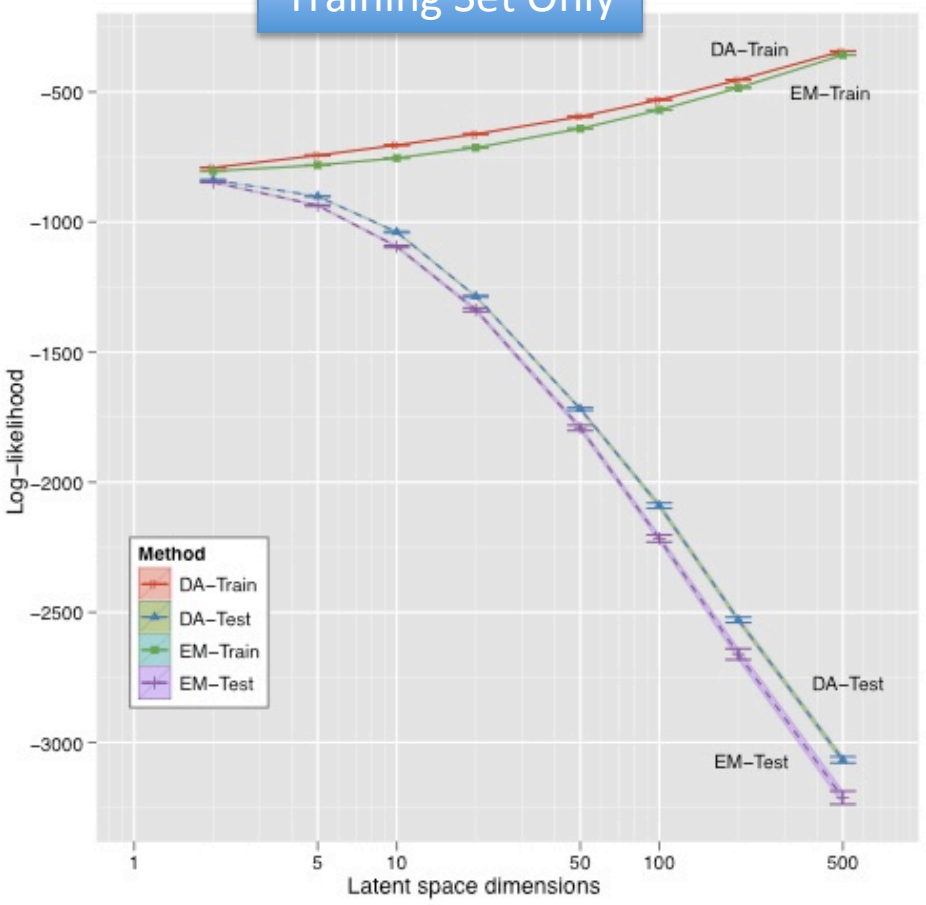
AP data with DA-PLSA

(AP: 2,246 documents & 10,473 words)

Test Set Only

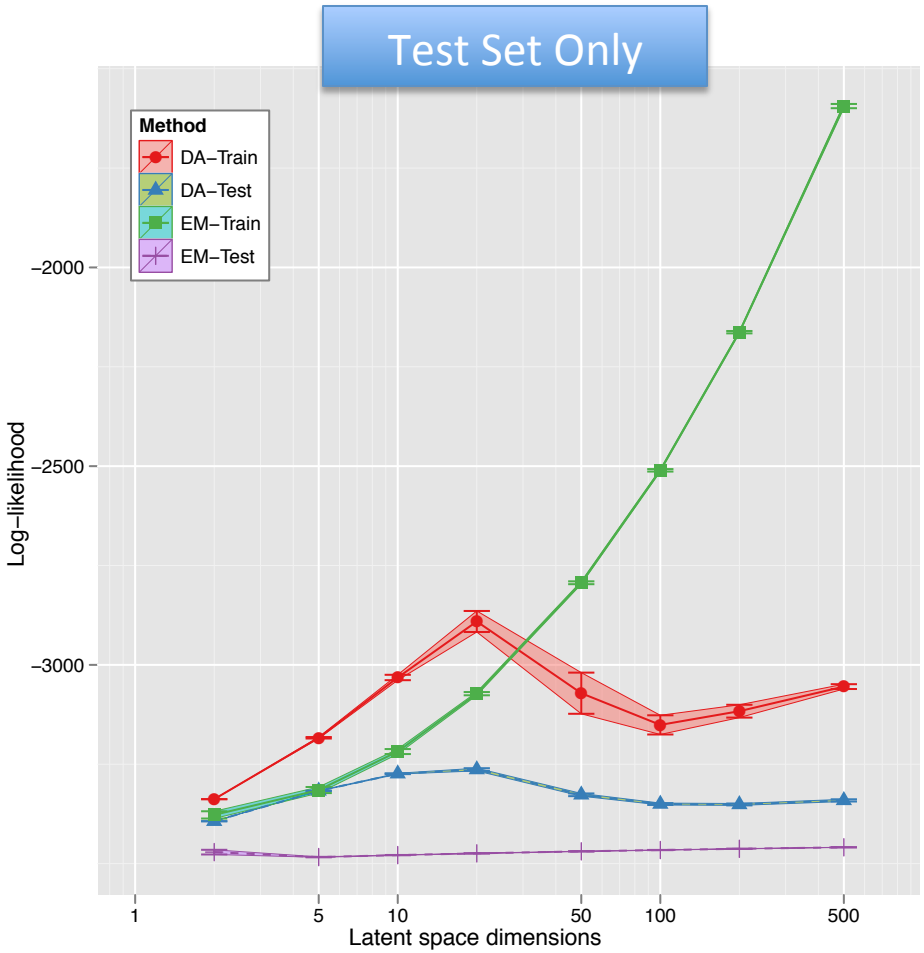


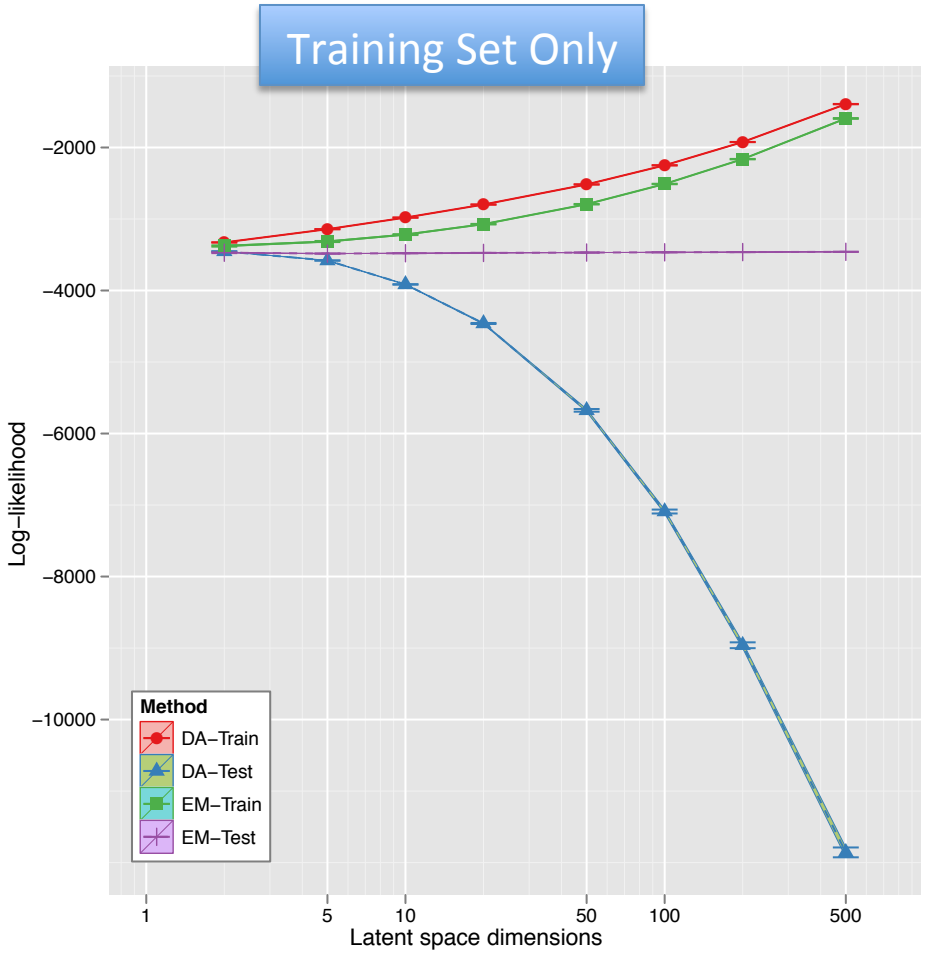
Training Set Only



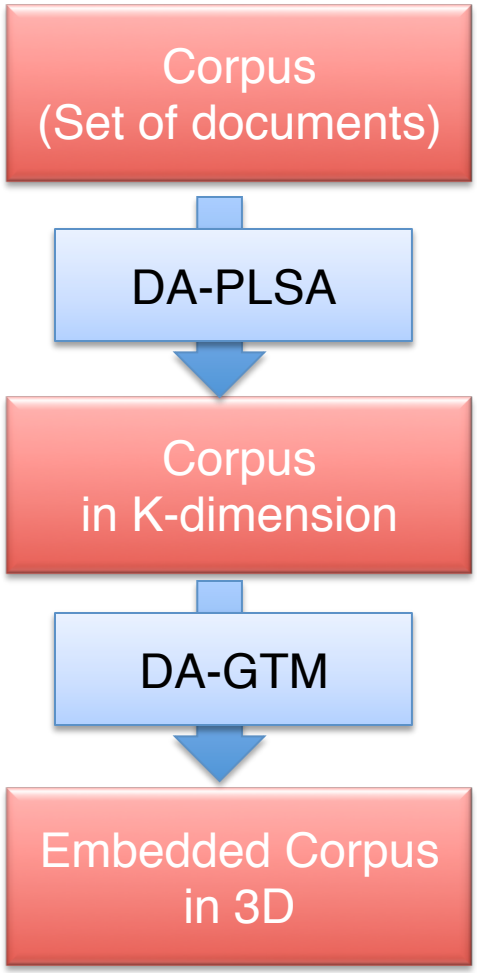
NIPS data with DA-PLSA

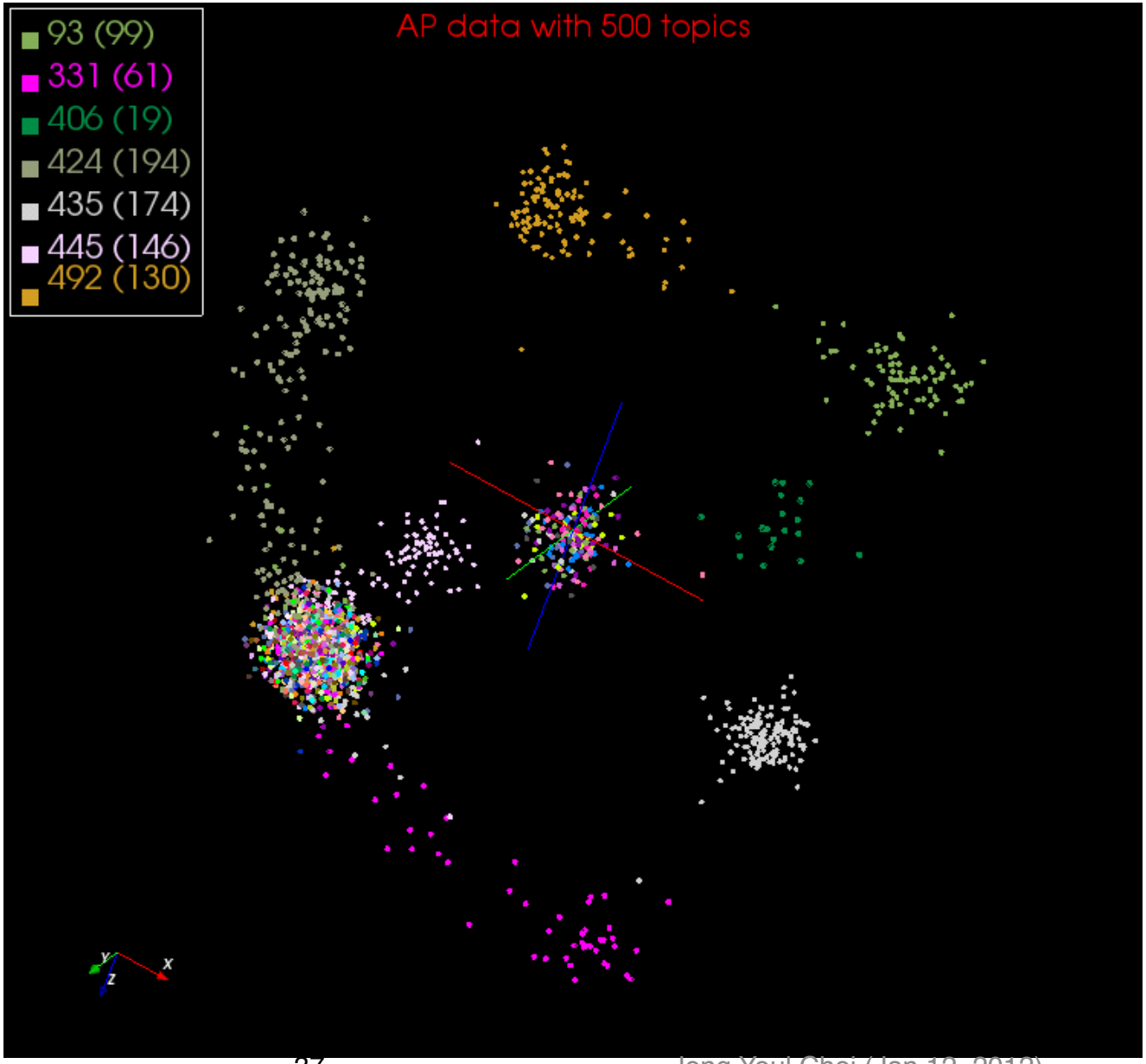
(NIPS: 1,500 doc & 12,419 words)





DA-PLSA with DA-GTM





AP Data Top Topic Words

- In the previous picture, we found among 500 topics:

Topic 331	Topic 435	Topic 424	Topic 492	Topic 445	Topic 406
lately	lately	mandate	mandate	mandate	plunging
oferrell	oferrell	kuwaits	kuwaits	lately	referred
mandate	ACK	cardboard	cardboard	ACK	informal
ACK	fcc	commuter	commuter	cardboard	Anticommu.
fcc	mandate	ACK	lately	fcc	origin
cardboard	cardboard	fcc	ACK	commuter	details
commuter	exam	lately	exam	oferrell	relieve
exam	commuter	exam	fcc	exam	psychologist
kuwaits	fabrics	fabrics	oferrell	kuwaits	lately
fabrics	corroon	oferrell	fabrics	fabrics	thatcher

ACK : acknowledges
Anticommu. : anticommunist

EM vs. DA-{GTM, PLSA}

		EM	DA
Optimization		Maximize log-likelihood *L*	Minimize free energy *F*
Objective Functions	GTM	$\sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{k=1}^{K} p(\boldsymbol{x}_n \mid \boldsymbol{y}_k) \right\}$	$-T \sum_{n=1}^{N} \ln \left\{ \left(\frac{1}{K} \right)^{\frac{1}{T}} \sum_{k=1}^{K} p(\boldsymbol{x}_n \mid \boldsymbol{y}_k)^{\frac{1}{T}} \right\}$
	PLSA	$\sum_{n=1}^{N} \ln \sum_{k=1}^{K} \{\psi_{nk} \mathrm{Multi}(\boldsymbol{x}_n \mid \boldsymbol{y}_k)\}$	$-T \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \{\psi_{nk} \mathrm{Multi}(\boldsymbol{x}_n \mid \boldsymbol{y}_k)\}^{\frac{1}{T}}$
		Note: When *T* = 1, *L* = -*F*. This implies EM can be treated as a special case in DA	
Pros & Cons		▪ Very sensitive ▪ Trapped in local optima ▪ Faster	▪ Less sensitive to an initial condition ▪ Find global optimum ▪ Require more computational time

I. Finite Mixture Model (FMM)

II. Model Fitting with Deterministic Annealing (DA)

III. Generative Topographic Mapping with Deterministic Annealing (DA-GTM)

IV. Probabilistic Latent Semantic Analysis with Deterministic Annealing (DA-PLSA)

V. Conclusion And Future Work

Conclusion

- Finite Mixture Model (FMM) problems
  - FMM-1 and FMM-2
  - Maximize log-likelihood for model fitting (MLE)
  - Traditional solutions use EM
- Solve FMMs with DA
  - Avoid local optimum problem
  - Find generalized (smoothed) solution
- Enhance and develop two data mining algorithms
  - DA-GTM
  - DA-PLSA

Future Work

- Determine number of components
  - Help to choose the right number of clusters, topics, the number of lower dimension, …
  - Bayesian model selection, minimum description length (MDL), Bayesian information criteria (BIC), …
  - Need to develop in a DA framework
- Quality study for DA-PLSA
  - Comparison with LDA
  - Precision and recall measurements
- Performance study for data-intensive analysis
  - MPI, MapReduce, PGAS, …

Related Publications

- [CCPE] **J. Y. Choi**, S.-H. Bae, J. Qiu, B. Chen, and D. Wild. *Browsing large scale cheminformatics data with dimension reduction*. *Concurrency and Computation: Practice and Experience,* 2011.
- [ECMLS] **J. Y. Choi**, S.-H. Bae, J. Qiu, G. Fox, B. Chen, and D. Wild. *Browsing large scale cheminformatics data with dimension reduction*. In *Workshop on Emerging Computational Methods for Life Sciences (ECMLS), in conjunction with the 19th ACM International Symposium on High Performance Distributed Computing (HPDC) 2010,* HPDC '10, pages 503–506, Chicago, Illinois, June 2010. ACM.
- [CCGrid] **J. Y. Choi**, S.-H. Bae, X. Qiu, and G. Fox. *High performance dimension reduction and visualization for large high-dimensional data analysis*. *Cluster Computing and the Grid, IEEE International Symposium on,* 0:331–340, 2010.
- [ICCS] **J. Y. Choi**, J. Qiu, M. Pierce, and G. Fox. *Generative Topographic Mapping by Deterministic Annealing*. In *Proceedings of the 10th International Conference on Computational Science and Engineering (ICCS 2010),* 2010.

Thank you!!

Question?

Email me at jychoi@cs.indiana.edu