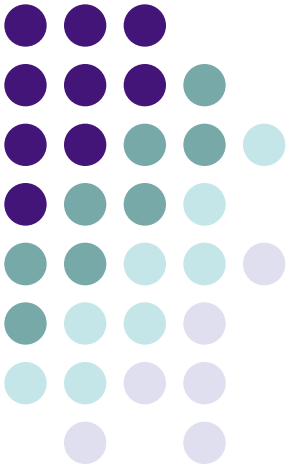
School of Computer Science
Carnegie Mellon

LDA, Sparse Coding, Matrix Factorization, and All That

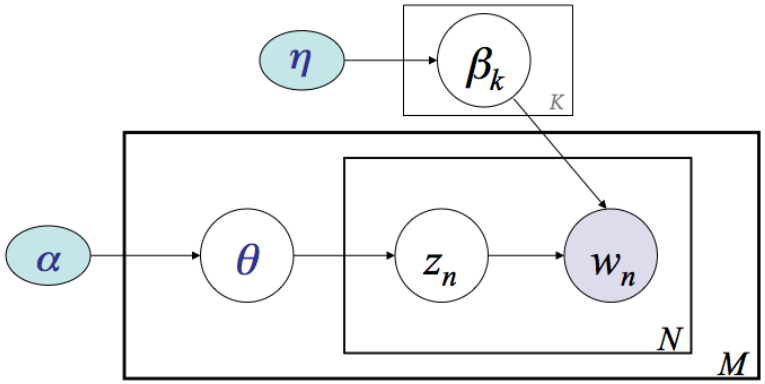


Yaoliang Yu

Carnegie Mellon University

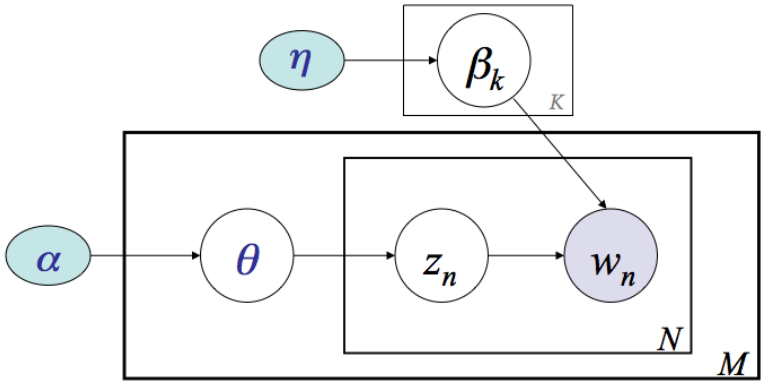
Take-home

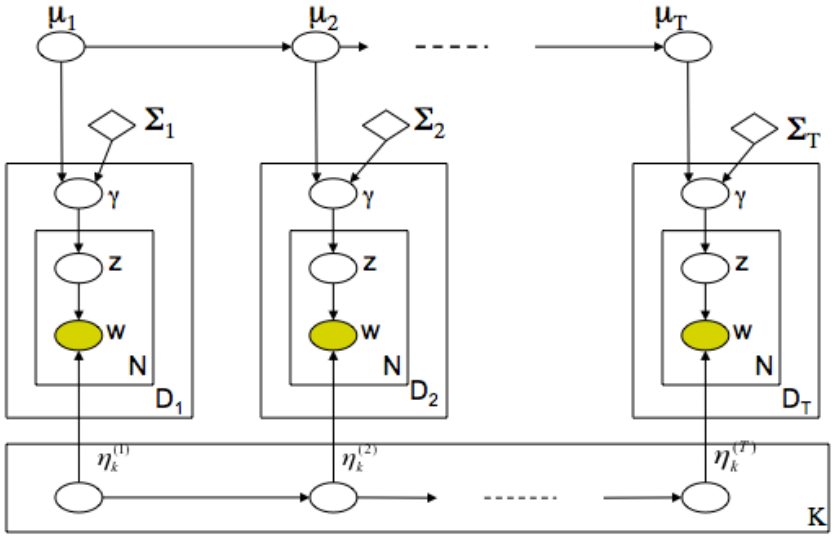
- Topic models are cool



Take-home

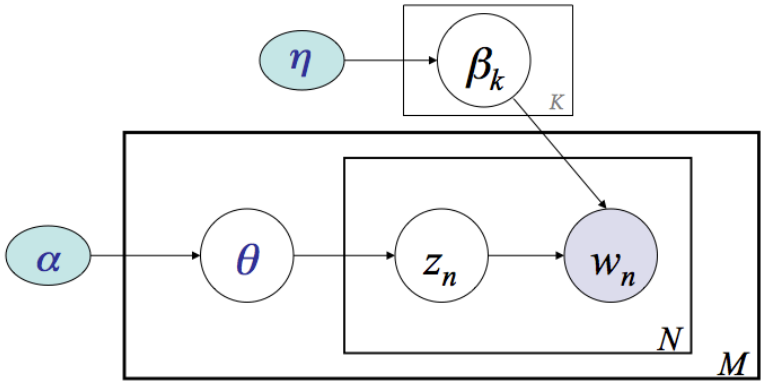
- Topic models are cool

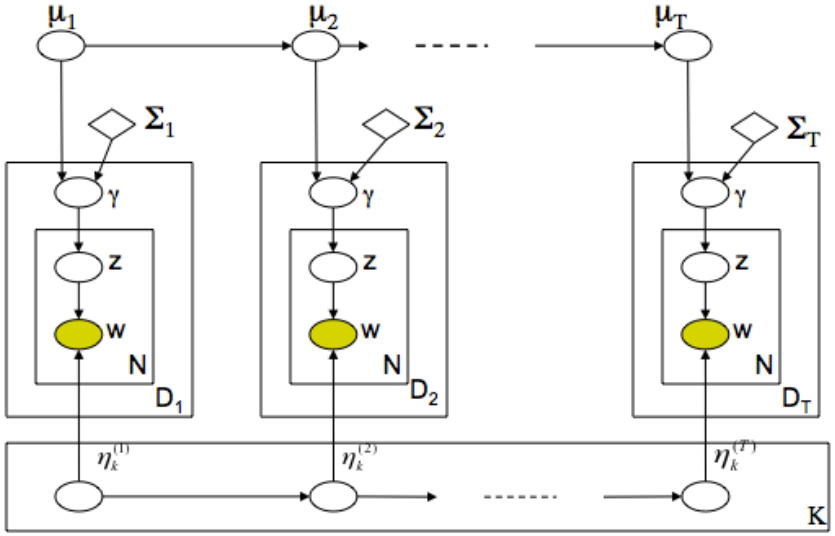




Take-home

- Topic models are cool



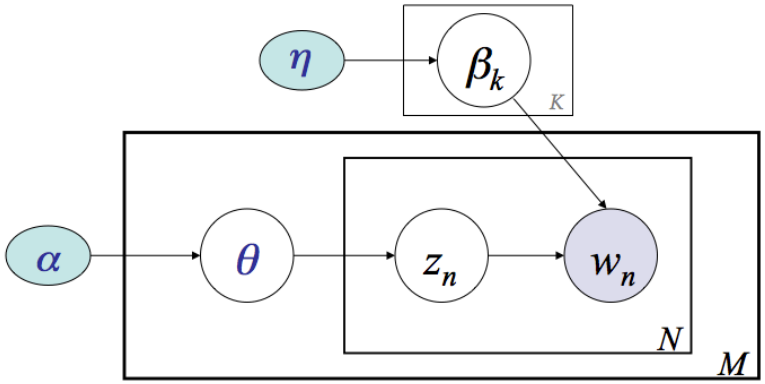


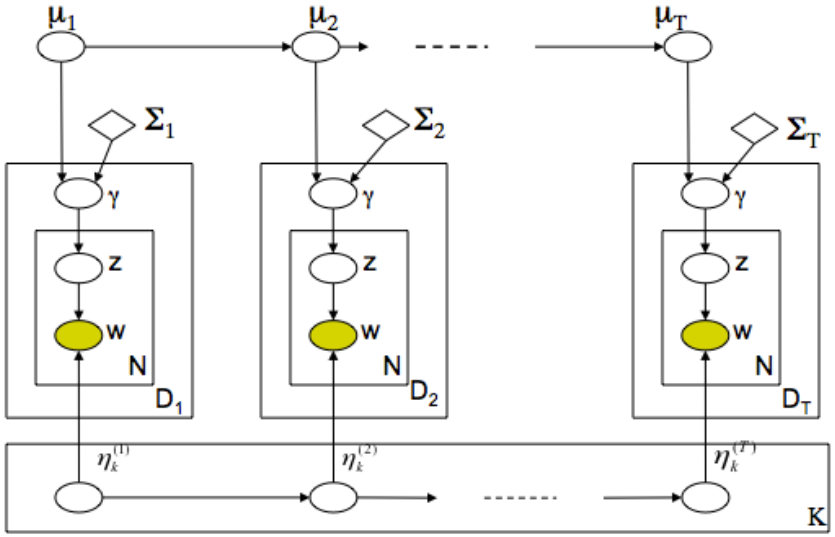
- Matrix factorizations are … simple

$$X \approx AB$$

Take-home

- Topic models are cool





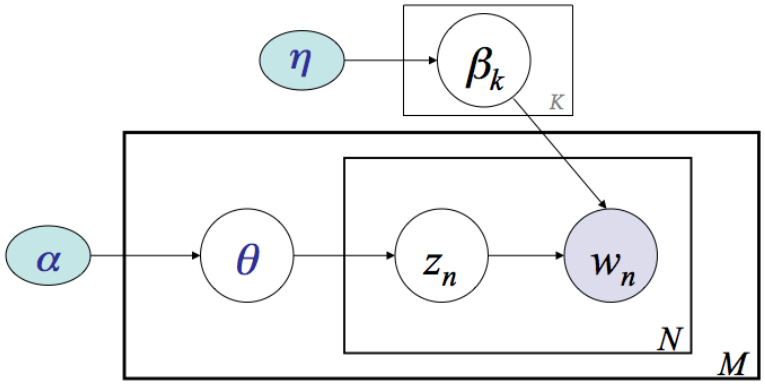
- Matrix factorizations are … simple

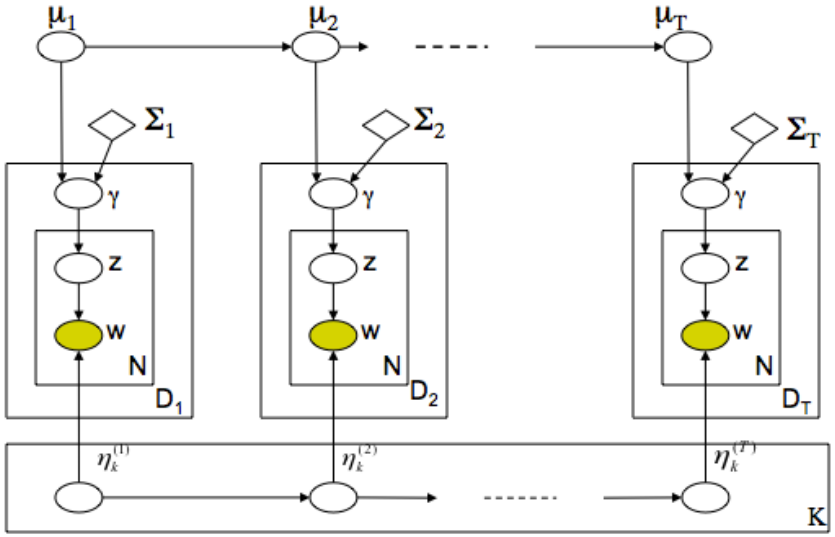
$$X \approx AB$$

- But, they are intimately related
- Allow knowledge transfer

Take-home

- Topic models are cool

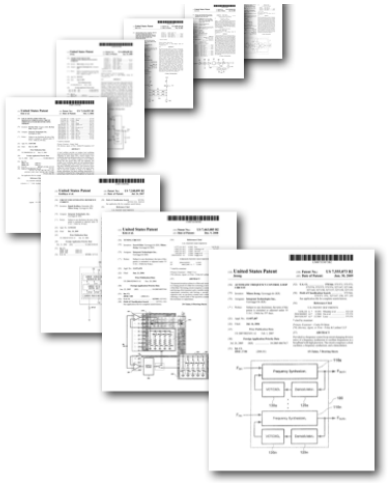




- Matrix factorizations are … simple

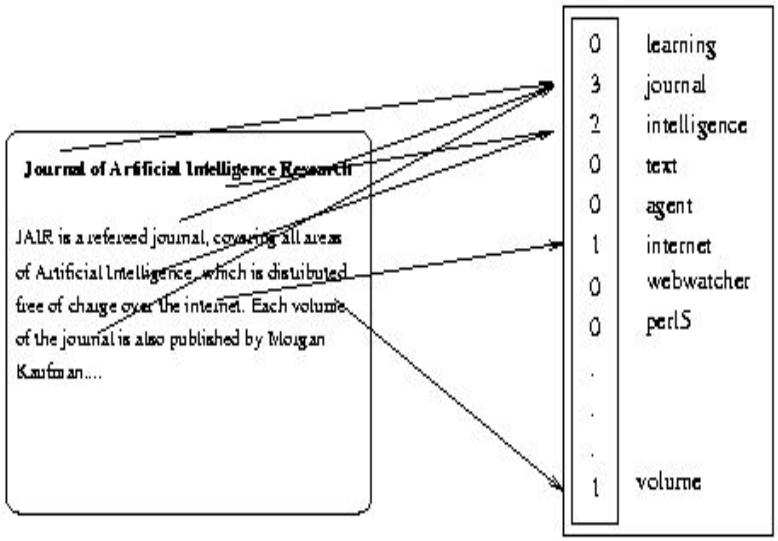
$$X \approx AB$$

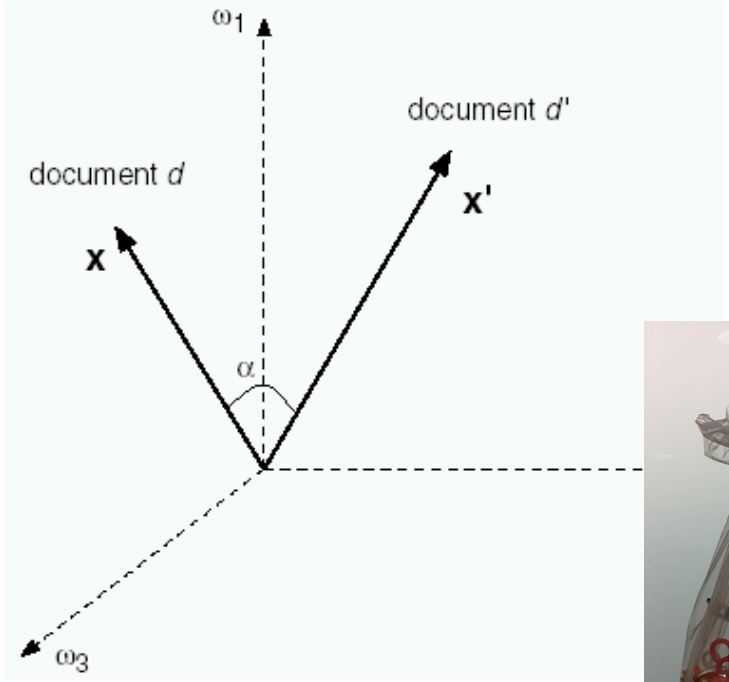
- But, they are intimately related
- Allow knowledge transfer



The Vector Space Model

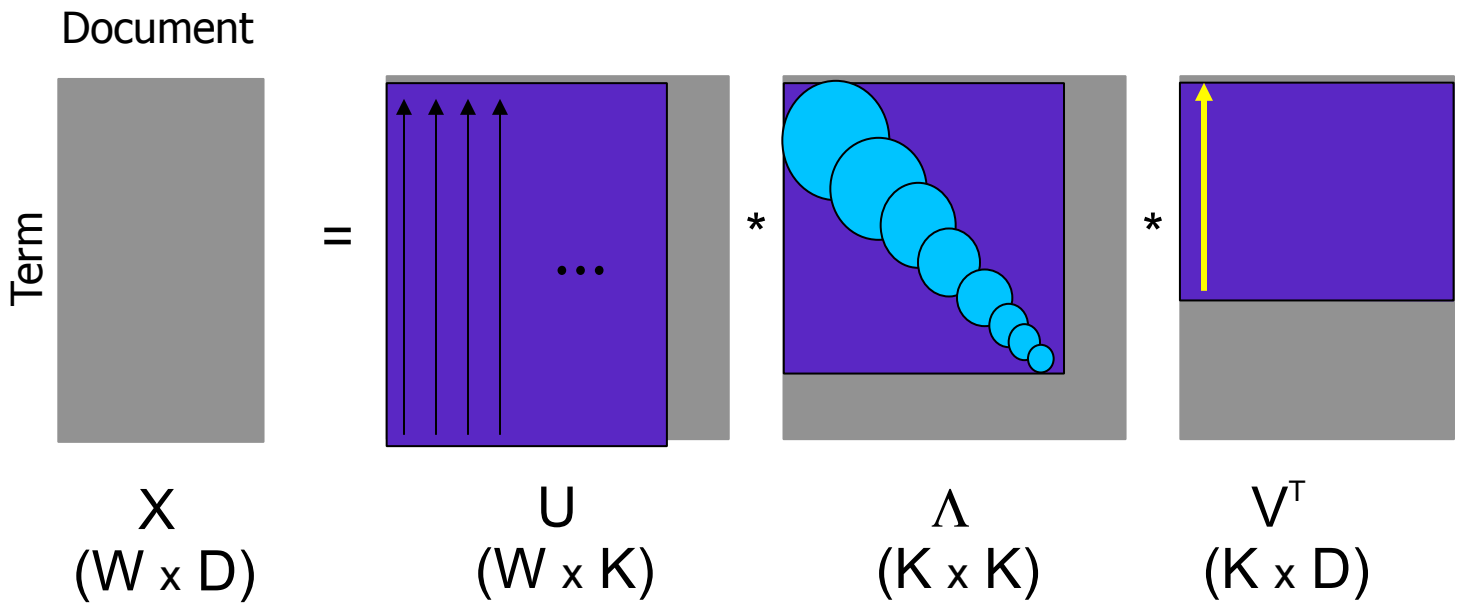
- Bag-of-words representation of doc
- Order ignored, only counts matter





Latent Semantic Indexing

(Deerwester et al., JASIS'90)



- Startling applications (Berry et al., SIAM Rev'95)
  - Cross-language retrieval
  - TOEFL/GRE synonym
  - Match paper submissions with reviewers (SIGIR'92!)

What is really LSI?

- From an optimization point of view

$$\min_{U,V} \quad \|X - UV\|_F^2$$

$$\text{s.t.} \quad U \in \mathbb{R}^{W \times K}, V \in \mathbb{R}^{K \times D}$$

- Nothing but matrix factorization/approximation !
  - K << W, K << D, otherwise trivial
  - Solvable by SVD, even though not jointly convex
- Things can be improved:
  - Real-valued?
  - How to set K?
  - No probabilistic model?
  - Squared loss?

Towards A Prob. Model

$$X \sim \Pr(\cdot|\Theta), \text{where}$$

$$\Theta = UV^\top \text{ has low-rank}$$

- What prob. dist. should we use?
  - Exponential family!
- ML with independence:

$$L(\Theta) = -\log \mathrm{p}(X|\Theta) \propto -\langle X, \Theta \rangle + G(\Theta)$$

- Low-rank makes estimation possible
- Gaussian → $G(\Theta) = \frac{1}{2}\|\Theta\|_F^2$ → LSI

Exponential Family PCA

(Collins et al., NIPS'01)

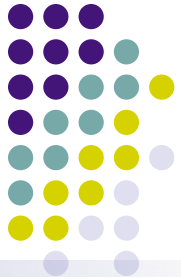
$$\min_{U,V} -\langle X, UV^\top \rangle + G(UV^\top)$$

- G: log-partition function, strictly convex and analytic
- Alternating minimization
  - Fix U, solve the convex subproblem w.r.t. V;
  - Fix V, solve the convex subproblem w.r.t. U;
  - Converge to a stationary point
- Amounts to:
  - Factorizing X under fancier loss than least squares
- Choose Poisson for count data, unconstrained

$$\min_{\Theta = UV^\top} \sum_{w,d} -X_{w,d}\Theta_{w,d} + \exp(\Theta_{w,d})$$

Nonnegative Matrix Factorization

(Lee & Seung, Nature'99)



- Count data X is nonnegative
  - Only (?) make sense to approximate with nonnegative numbers

$$\min_{U,V\geq 0} \|X - UV^\top\|_F^2$$

- Can also use KL divergence

$$\min_{U,V\geq 0} \sum_{w,d} -X_{w,d}\log\Theta_{w,d} + \Theta_{w,d}$$

- Very similar to EXP PCA !
  - Yet another fancier loss
- Multiplicative updates
  - Must initialize densely !

$$U_{w,k} \leftarrow U_{w,k}\frac{\sum_d V_{d,k}X_{w,d}/\Theta_{w,d}}{\sum_d V_{d,k}}$$

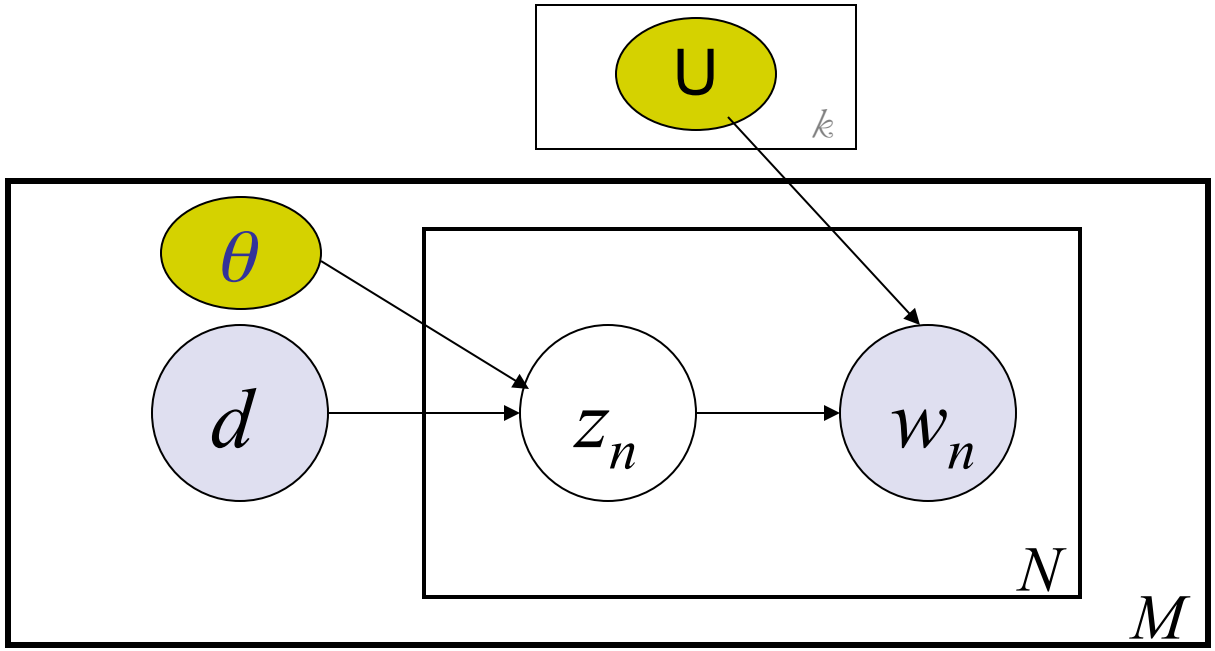
$$V_{d,k} \leftarrow V_{d,k}\frac{\sum_w U_{w,k}X_{w,d}/\Theta_{w,d}}{\sum_w U_{w,k}}$$

Synonymy vs. Polysemy

- LSI (and relatives aforementioned) is good at synonym
  - Similar words are close in latent semantic space

- But less so at polysemy:
  - "It was a nice **shot**."

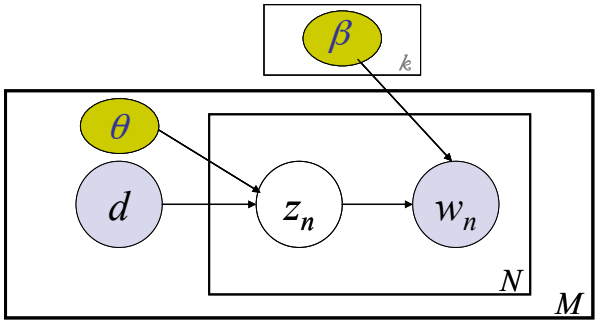
Probabilistic LSI (Hoffman, ML'01)



$$p(X) = \prod_{d=1}^{D} \mathrm{p}(d) \prod_{w=1}^{N_d} \sum_{k=1}^{K} \underbrace{\mathrm{p}(X_{w,d}|Z_{w,d}=k)}_{U_{w,k}} \underbrace{\mathrm{p}(Z_{w,d}=k|d)}_{V_{d,k}}$$

pLSI cont'

- 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).



pLSI cont''

- Maximize the marginal likelihood:

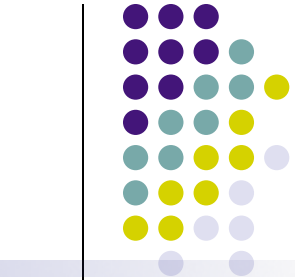
$$\min_{\Theta=UV} \quad \sum_{w,d} -X_{w,d} \log \Theta_{w,d}$$

$$\text{s.t.} \quad U, V \geq 0, U^\top 1 = 1, V1 = 1$$

- Similar to EXP PCA, NMF
  - But with more stringent normalization constraints
- Can use EM to optimize U, V (Hoffman'01)
  - Or simply alternate U, V
- Solves polysemy since same word can be drawn from different topics
- But may overfit !
  - Parameter V grows with doc size D

Latent Dirichlet Allocation

(Blei et al., JMLR'03)

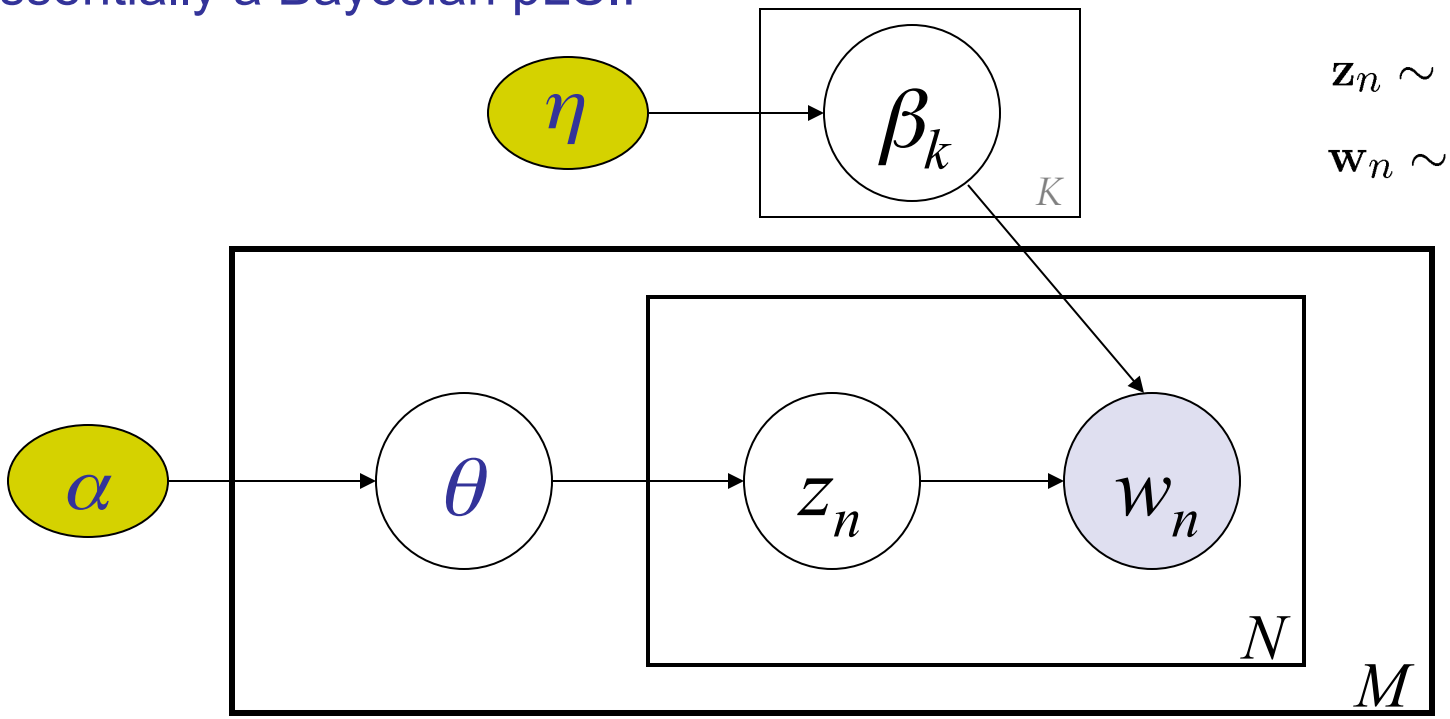


Essentially a Bayesian pLSI:

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

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

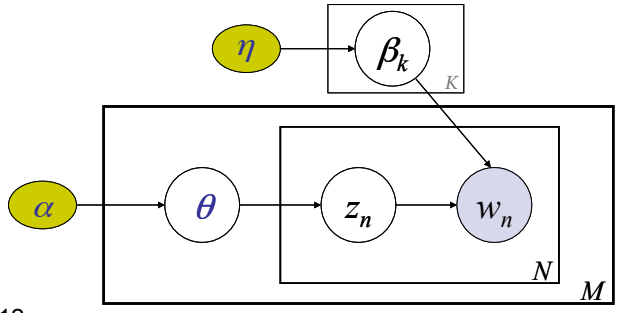
$$\mathbf{w}_n \sim p(\mathbf{w}_n|\mathbf{z}_n, \beta)$$



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

LDA

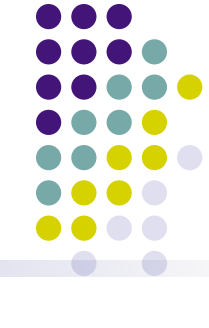
- Generative model
- Models each word in a document as a sample from a mixture model.
- Each word is generated from a single topic, different words in the document may be generated from different topics.
- A topic is characterized by a distribution over words.
- Each document is represented as a list 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
- How does LDA avoid overfitting?



PCA -- revisited

- PCA enjoys a set of salient properties:
  - Orthogonal basis
  - Implicit Gaussian assumption
  - 2nd order de-correlation
  - "noiseless"
- All above have been modified:
  - Overcomplete basis (Sparse coding)
  - Exponential family (more later)
  - High order de-correlation (ICA)
  - Probabilistic model
- LDA is Probabilistic PCA

Probabilistic PCA (Tipping & Bishop, JRSSB'99)



- Probabilistic model (U, sigma hyperparameter):

$$X|V \sim \mathcal{N}(UV^\top, \sigma^2 I_W), \quad V^\top \sim \mathcal{N}(0, I_K)$$

- Analytic marginalization:

$$X \sim \mathcal{N}(0, UU^\top + \sigma^2 I_W)$$

- MLE:

$$\min_{U,\sigma^2} \log\det(UU^\top + \sigma^2 I) + \langle S, (UU^\top + \sigma^2 I)^{-1} \rangle$$

  - Apply von Neuman's trace inequality:

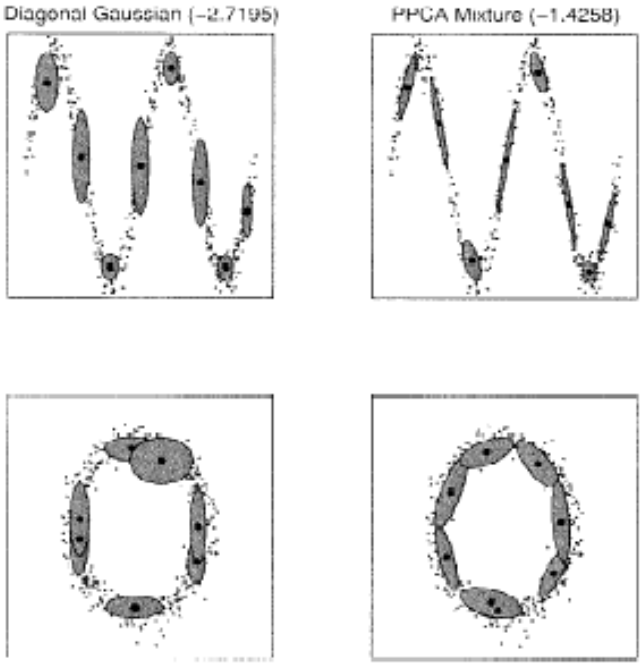
$$\min_{\sigma_k^2,\sigma^2} \sum_{k=1}^{K} \log(\sigma_k^2 + \sigma^2) + \frac{s_k^2}{\sigma_k^2 + \sigma^2} + \sum_{k=K+1}^{W} \log\sigma^2 + \frac{s_k^2}{\sigma^2}$$

  - Set derivative to 0:

$$\sigma_k^2 = s_k^2 - \sigma^2 \qquad \sigma^2 = \frac{1}{W-K} \sum_{k=K+1}^{W} s_k^2$$

PPCA cont'

- Similar to conventional PCA
  - Take sample covariance eigenvector
  - Shrink first K eigenvalues by the average of the tail eigenvalues
  - Recover PCA when sigma = 0, i.e., noiseless
- Many extensions
  - Mixture of PPCA (Tipping & Bishop, NC'99)
  - Hierarchical PPCA (Bishop & Tipping, PAMI'98)
  - Sparse PPCA (Guan & Dy, AISTATS'09)
  - Robust PPCA (Archambeau et al., ICML'06)
  - Infinite PPCA ??



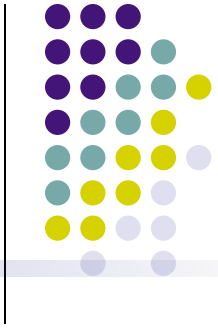
Multinomial PCA (Buntine, ECML'02)

- Apply PPCA to count data X:

$$X_{:,d}|V_{d,:} \sim \mathrm{Mul}(UV_{d,:}^{\top}, N_d), \quad V_{d,:} \sim \mathrm{Dir}(\alpha)$$

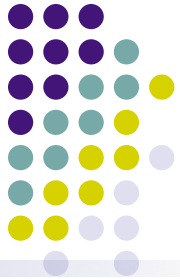
- Like before, U, alpha are hyperparameters
  - But each column in U is a probability distribution
- This is LDA !
  - Essentially marginalizing Z out
- But no longer can analytically marginalize out V
  - Solve by VI, SVI, or Gibbs sampling

What have learned so far



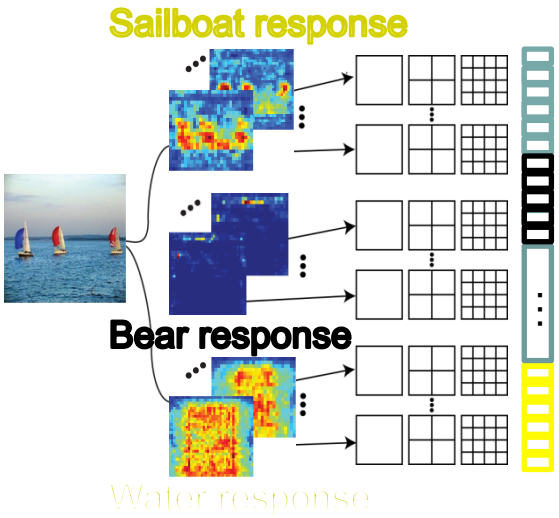
- Matrix factorization view of topic models
  - LSI, EXP PCA, NMF, pLSI are all matrix factorizations, under different loss / constraints
- Probabilistic view of matrix factorizations
  - PPCA, LDA = Multinomial PCA
- This connection is exploited in recent theoretical results
  - Papers have " .. provable ..." or "… spectral …"
- Ideas from matrix factorization easily translate to topic models, and vice versa
  - SVI, parallel implementation, etc.
- Issues have not been considered:
  - Sparsity: each doc contains few topics; each word appears in few topics
  - How to set K? (Yes, you've learned Dirichlet processes)

Strive after Sparsity



- Each doc contains few topics; each word appear in few topics
  - Each column in V or U needs to be sparse
- Cannot achieve exact sparsity with Bayesian methods
  - Bayesian estimates are conditional mean (under least squares risk)
  - Averaging never yields exact sparsity !
  - Can still achieve sparsity by ad hoc truncation
- But, hey, LASSO achieves exact sparsity
  - LASSO is MAP, not Bayesian
- Seen topic model = matrix factorization
  - There is sparse matrix factorization
  - Steal ideas from there!

Sparse Coding



$$= \left(\quad \cdots \quad \right) \times$$

$$X \approx UV^\top$$

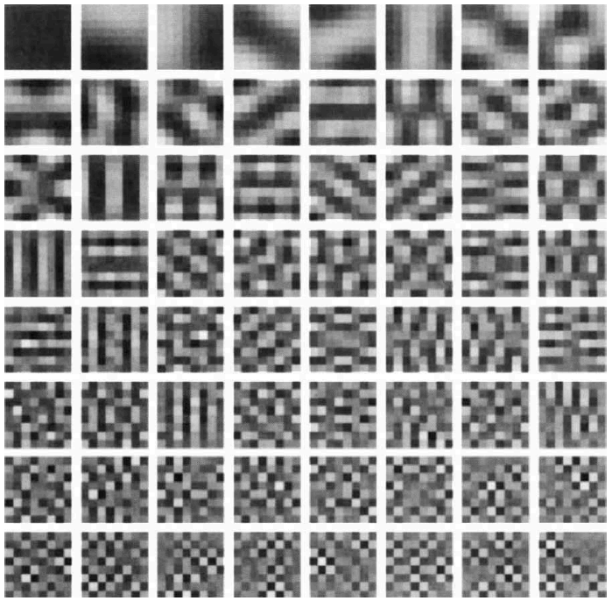
- U are called dictionary, e.g., topic distributions
- V are called coefficients (coding / loading / mixing proportion)
  - For each column $\mathbf{x} \approx U\mathbf{v}$, encoding x with v, under dictionary U
- For v to be sparse, U needs to be overcomplete
  - Wavelets, random matrices
  - Can be learned jointly with V

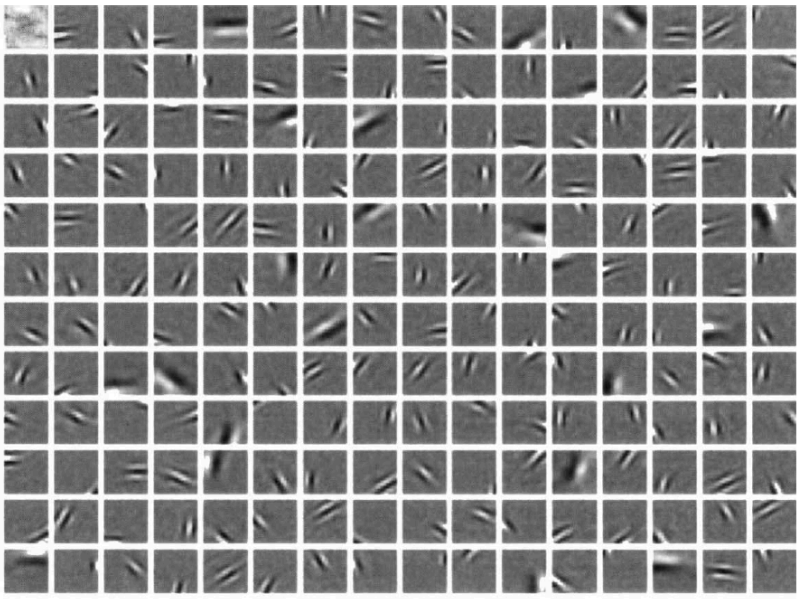
Sparse coding cont'

(Olshausen and Field, Nature'96)

$$\min_{U,V} \underbrace{\ell(X, UV^\top)}_{Reconstruction} + \lambda \sum_d \underbrace{g(V_{d,:})}_{sparseness}$$

- ell: likelihood, g: prior on V, MAP estimate of U, V
  - Need to constrain U due to indeterminism
  - Choose g e.g. the 1-norm





Sparse coding cont''

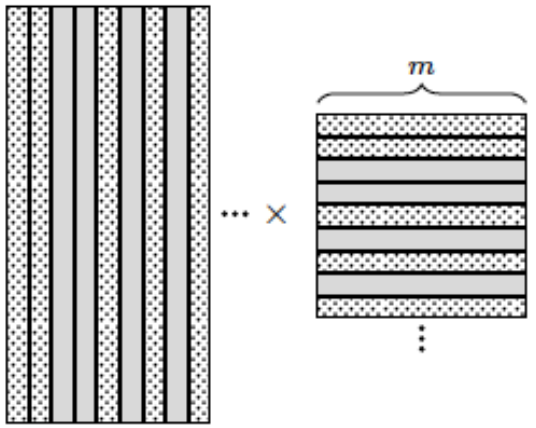
$$\min_{U,V} \underbrace{\ell(X, UV^\top)}_{Reconstruction} + \lambda \sum_d \sum_k \underbrace{g(V_{d,k})}_{sparseness}$$

- Exponential family ell (Lee et al., AAAI'10)
- Solve by alternating U and V
  - Converge to stationary point
  - In general NP-hard, recent global convergence guarantees
- Group sparse coding (Bengio et al., NIPS'09)
  - Group along the doc dim, e.g., l1/l2

Letting K → infty
(Zhang et al., AAAI'11, NIPS'12)

- Take g = || . ||

$$\min_{\Theta} \ell(X,\Theta) + \lambda \underbrace{\min_{\Theta=UV^\top, |U_{:,k}|\le 1} \sum_k \|V_{:,k}\|}_{|||\Theta|||}$$

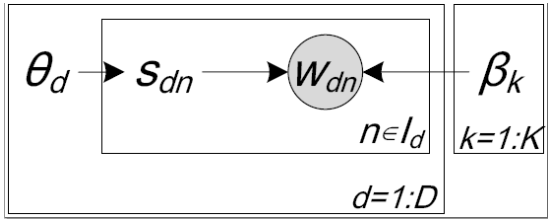


- K → infty, $|||\Theta|||$ becomes a norm (convex!)
  - Gauge induced by the set $\{\mathbf{u}\mathbf{v}^\top : |\mathbf{u}| \le 1, \|\mathbf{v}\| \le 1\}$
  - Dual norm $|||\Theta|||^\circ = \max\{\mathbf{u}^\top \Theta \mathbf{v} : |\mathbf{u}| \le 1, \|\mathbf{v}\| \le 1\}$
- For both $|\mathbf{u}| = \|\mathbf{u}\|_2, \|\mathbf{v}\| = \|\mathbf{v}\|_2$
  - $|||\Theta|||^\circ = \|\Theta\|_{\mathrm{sp}}$
  - Thus need to solve $\min_{\Theta} \ell(X,\Theta) + \lambda \|\Theta\|_{\mathrm{tr}}$
  - Induce low-rank, convex, no local minima !
- Can play with other choices of norms (e.g., White et al., NIPS'12)

Sparse Topical Coding
(Zhu & Xing, UAI'11)

- 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 $\boldsymbol{s}$



reconstruction loss

sparse codes

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

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

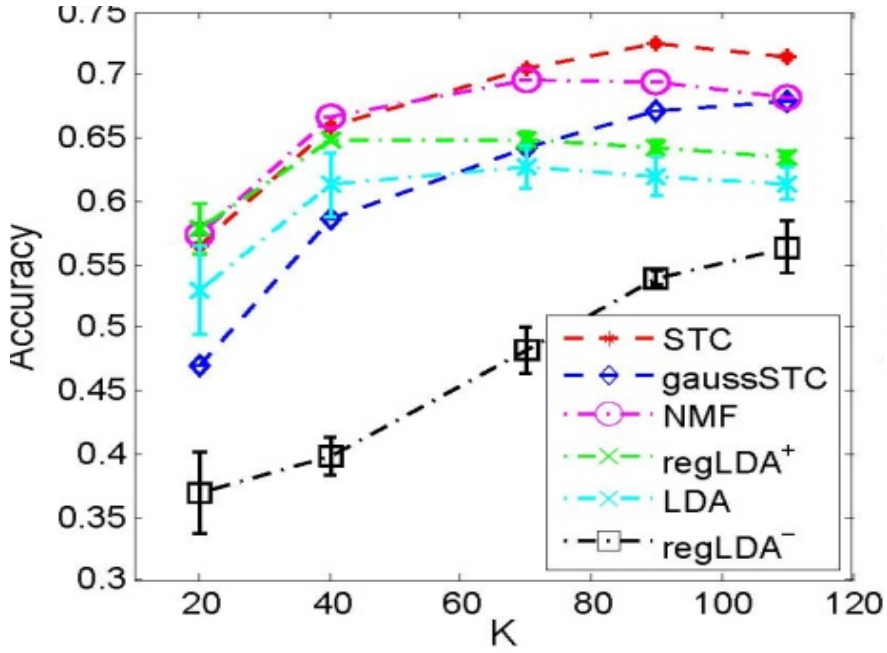
truncated aggregation

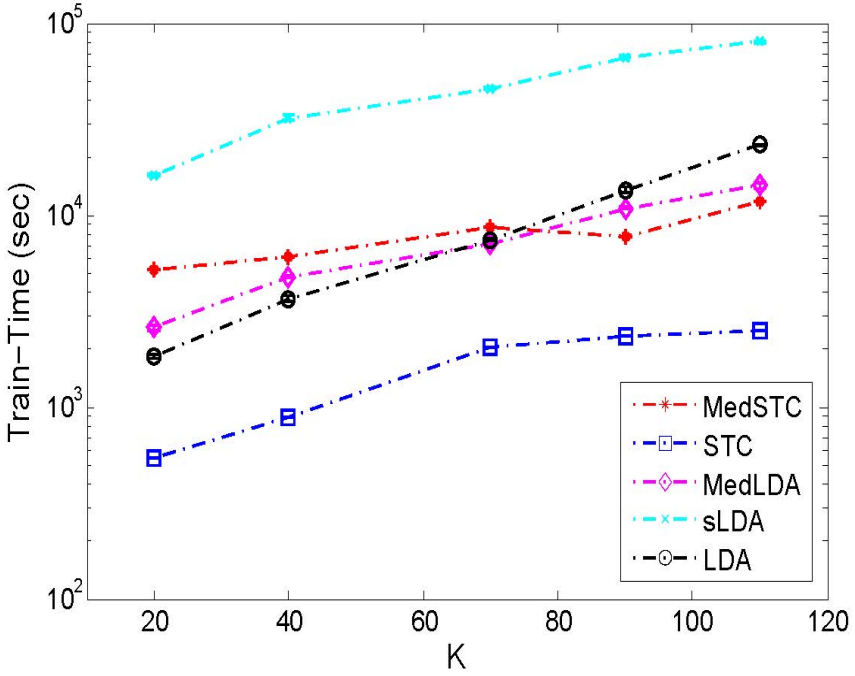
non-negative codes

topical bases

Comparisons

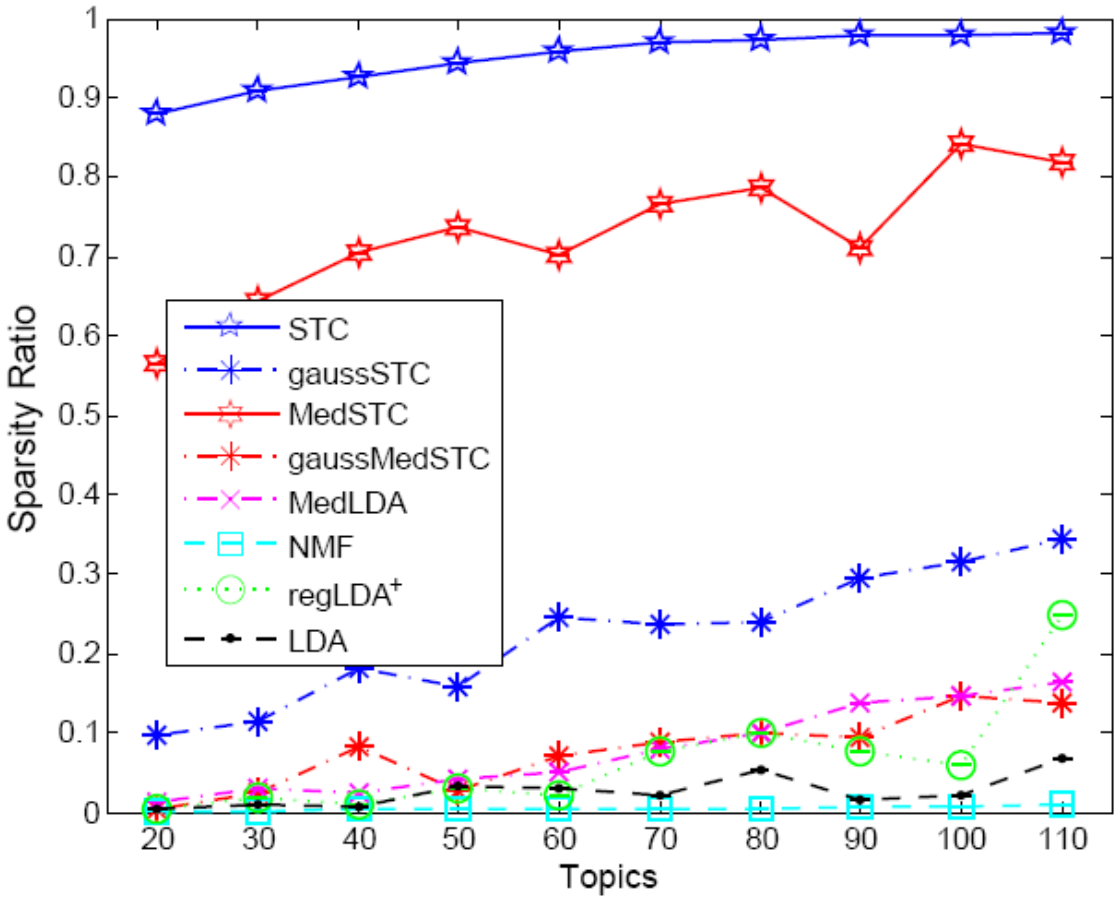
LDA vs. STC





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

Computation on STC

- 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

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

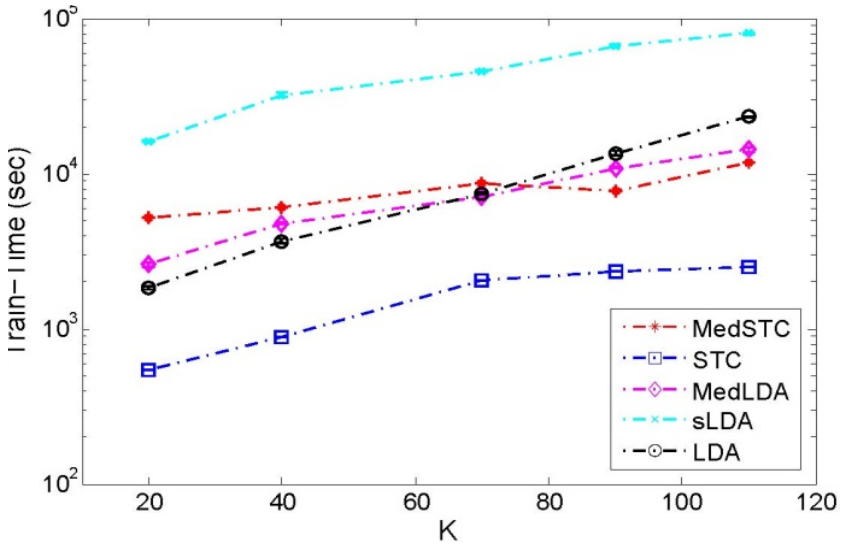
$$\text{where } 2\gamma\beta_{kn}\nu_k^2 + (2\gamma\mu + \beta_{kn}\eta)\nu_k + \mu\eta - w_n\beta_{kn} = 0$$

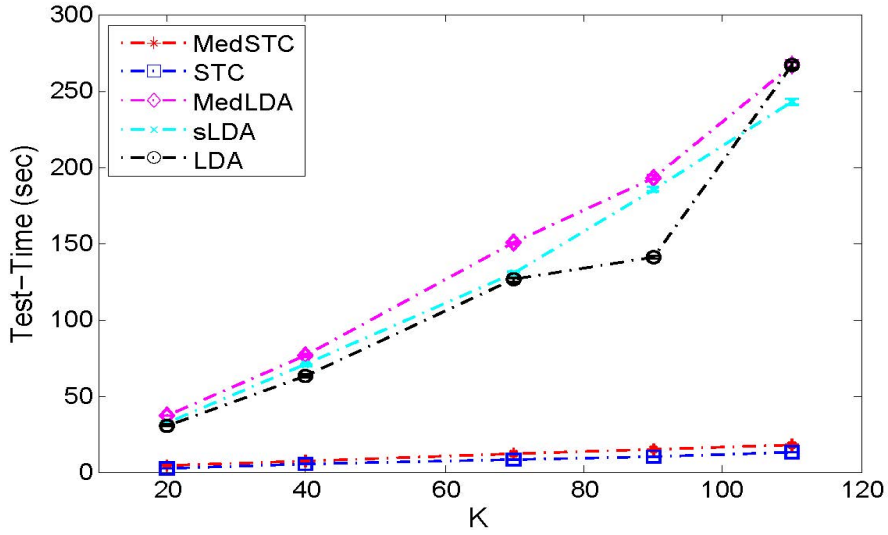
  - Document code (truncated averaging)

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

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

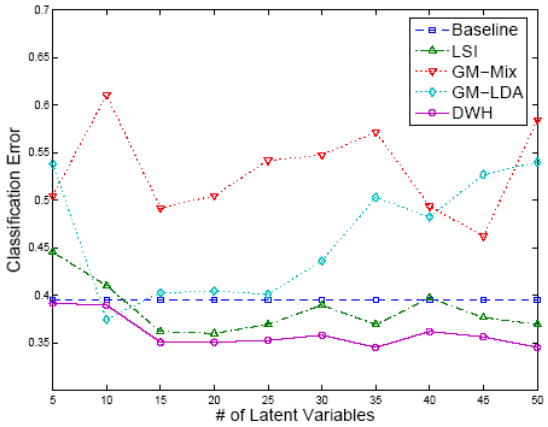
Performance



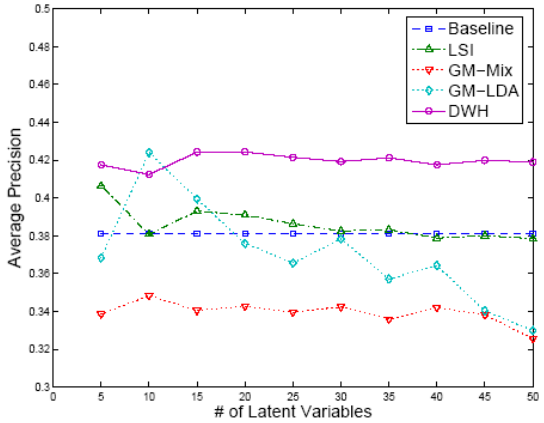


~ 10 times speed up in train &test

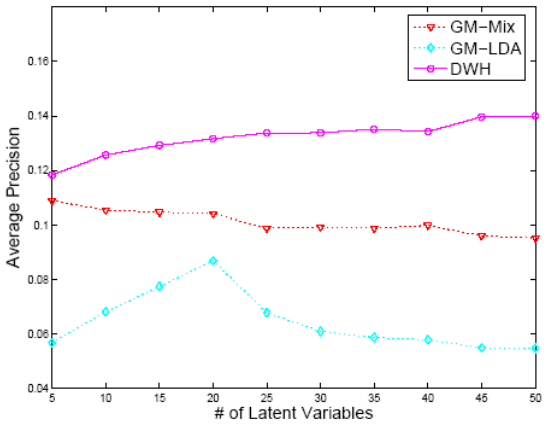
The shocking results on LDA



Classification



Retrieval

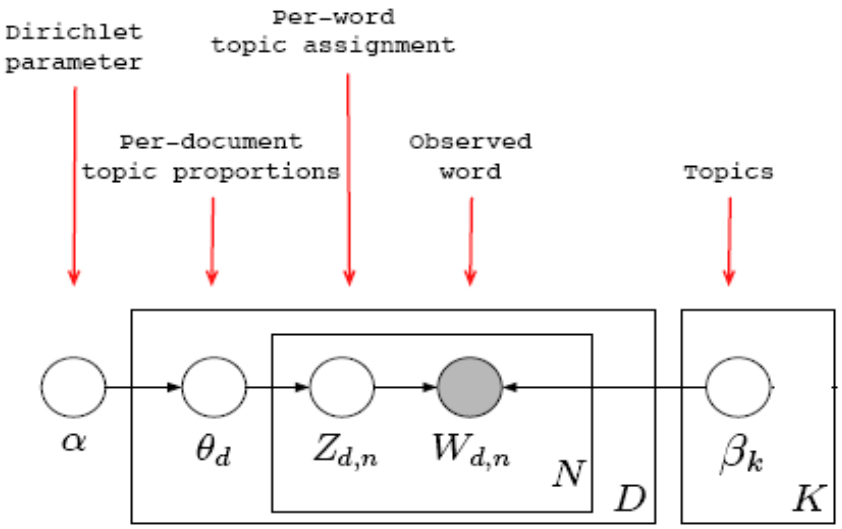


Annotation

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

Why?

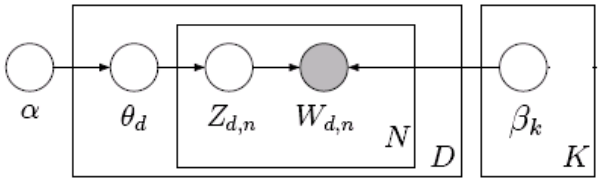
- LDA is not designed, nor trained for such tasks, such as classification, there is no warrantee that the estimated topic vector θ is good at discriminating documents



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

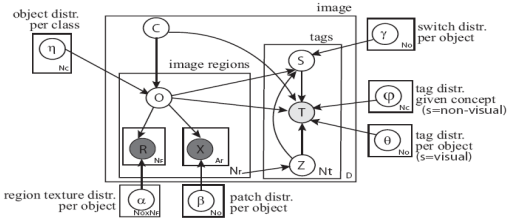
Unsupervised Latent Subspace Discovery

- 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]



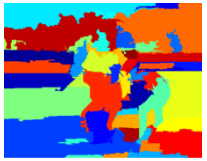
⇒

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

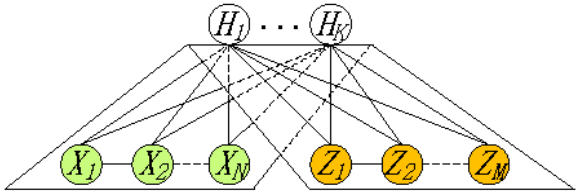


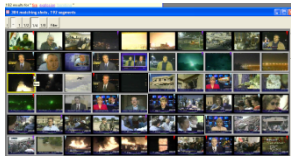
Athlete
Horse
Grass
Trees
Sky
Saddle

⇒

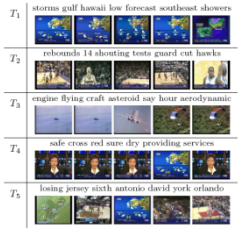


  - Multi-view latent Markov models [Xing et al 2005]



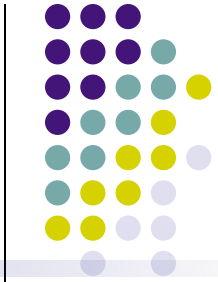


⇒



  - 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)

"Lovely welcomming staff, good rooms that give a good nights sleep, downtown location"

Meramees Hostel

Save Review

SheikhSahib 10 contributions
London

Jul 7, 2009 | Trip type: Friends getaway

This hotel is just of the side streets of Talat Harb, one of the main arteries to downtown Cairo. It is walking distance to the Nile, riverfront hotels, Egyptian Museum, and there are many eateries in the area at night when it is still bustling. Only a short cab ride away from the Old Fatimid Cairo.

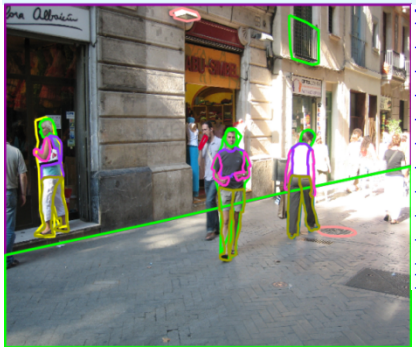
The staff are young and very friendly and able to sort out things like mobile chargers, internet, and they have skype installed on their computers which is brilliant. The rooms are nicer then the Luna (nearby) and much quieter as well.

My ratings for this hotel

Value, Service, Rooms, Location, Cleanliness

Date of stay February 2009
Visit was for Leisure
Traveled with With Friends
Member since July 03, 2009
Would you recommend this hotel to a friend? Yes

- 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

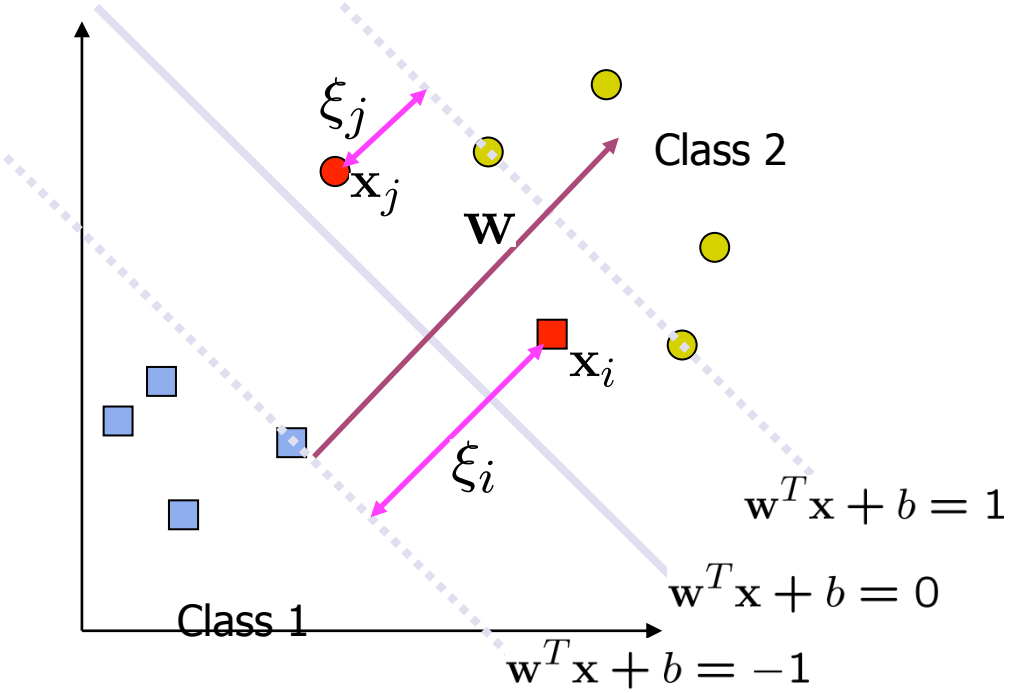
- Many others

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**

Support vector machines

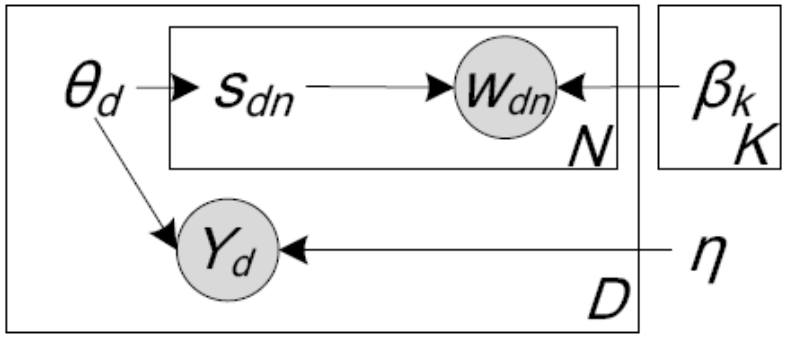


$$\min_{w,b} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$$

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

$$\xi_i \geq 0, \quad \forall i$$

Supervised STC



- Joint loss minimization

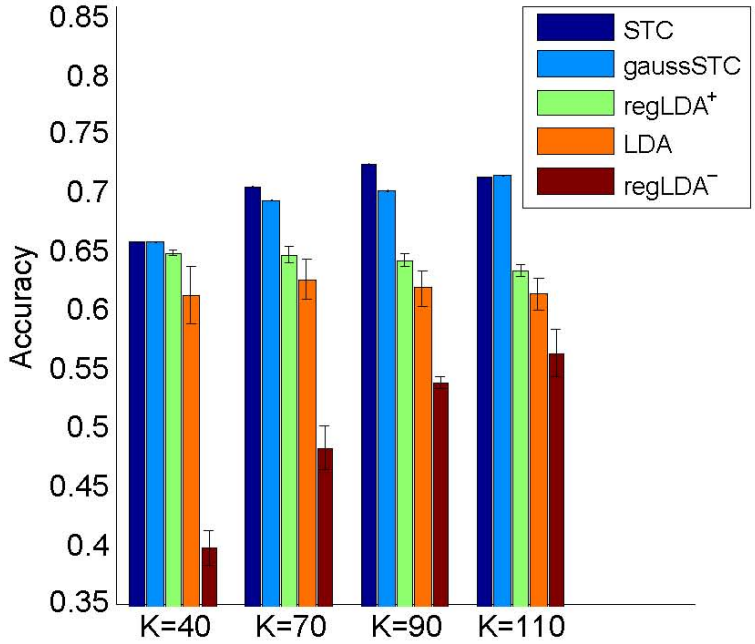
$$\min_{\{\theta_d\},\{\mathbf{s}_d\},\beta,\eta} \quad f(\{\theta_d\},\{\mathbf{s}_d\},\beta) + C\mathcal{R}_h(\{\theta_d\},\eta) + \frac{1}{2}\|\eta\|_2^2$$

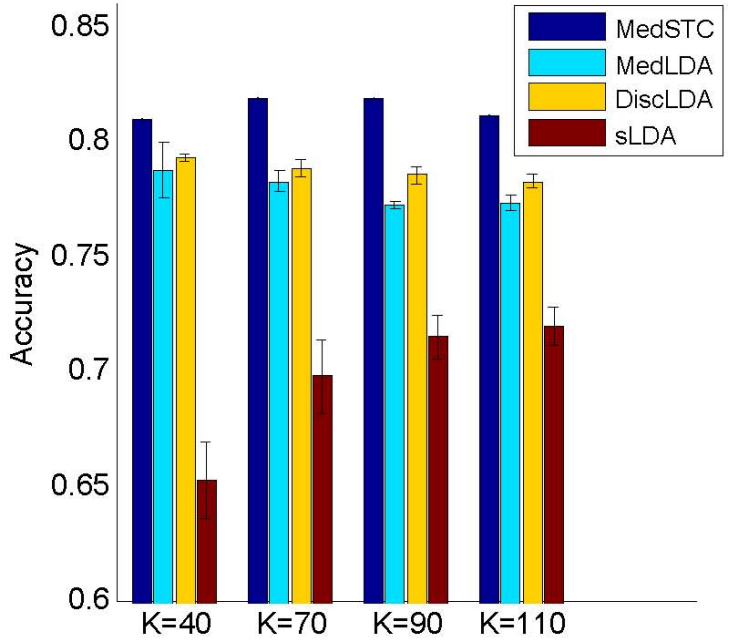
$$\text{s.t.}: \quad \theta_d \geq 0,\ \forall d;\ \mathbf{s}_{dn} \geq 0,\ \forall d, n \in I_d;\ \beta_k \in \mathcal{P},\ \forall$$

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

Classification accuracy

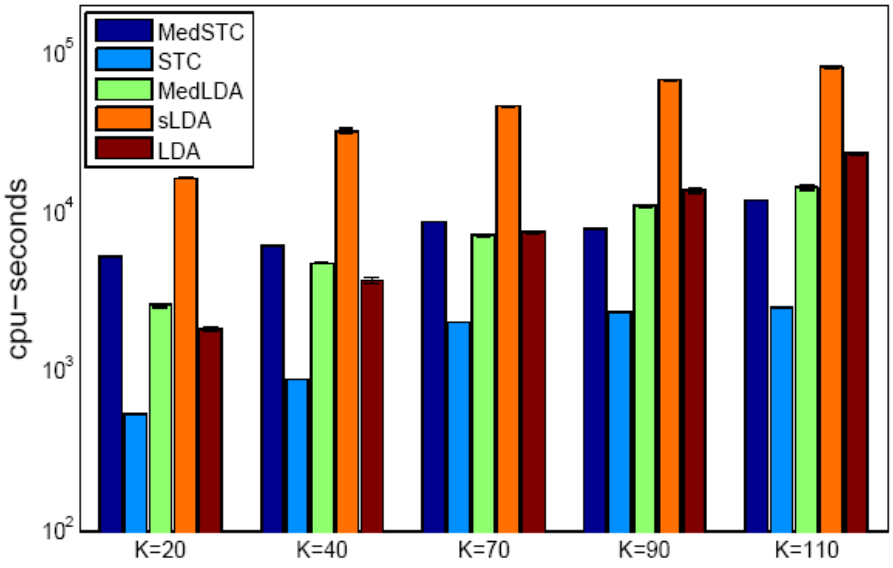
- 20 newsgroup data:

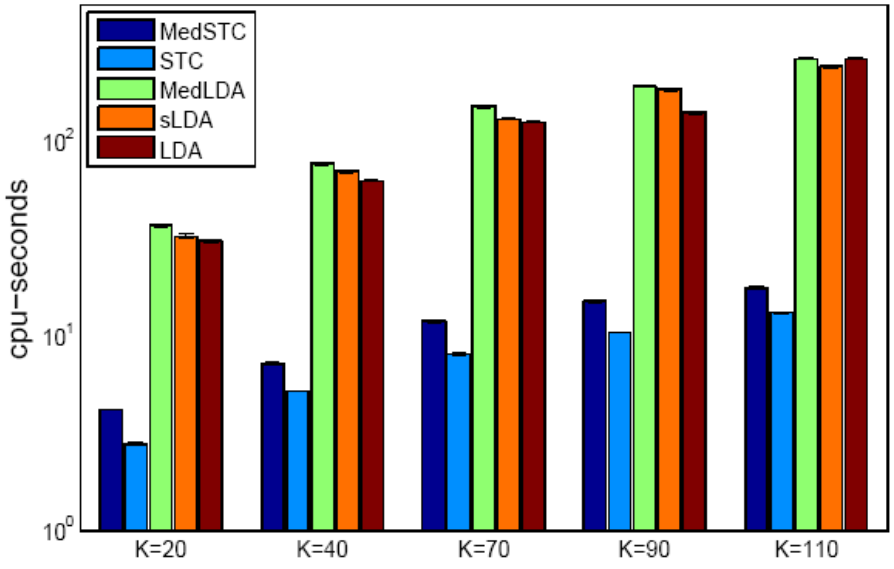




Time efficiency

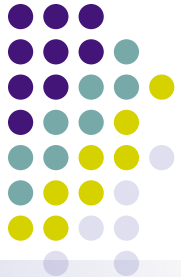
- training & testing time





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

Summary



- Topic models are intimately related to matrix factorization
- LDA = Multinomial PCA
- Exploring the connection can be beneficial
  - Understanding
  - Sparsity
  - Efficient inference
  - Supervision
- More elaborate matrix factorizations can be devised
  - Until next time !