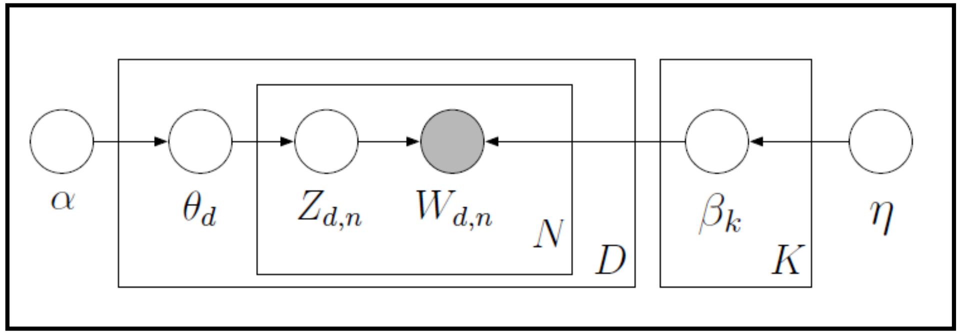
Correlated Topic Models

Authors: Blei and LaffertY, 2006

Reviewer: Casey Hanson

Recap Latent Dirichlet Allocation



- $D \equiv$ set of documents.
- $K =$ set of topics.
- $V =$ set of all words. $|N|$ words in each doc.
- $\theta_d \equiv$ Multi over topics for a document d $\in D$. $\theta_d \sim Dir(\alpha)$
- $\beta_k \equiv$ Multi over words in a topic, $k \in K$. $\beta_k \sim Dir(\eta)$
- $Z_{d,n} \equiv$ topic selected for word n in document d. $Z_{d,n} \sim \text{Multi}(\theta_d)$
- $W_{d,n} \equiv n_{th}$ word in document d. $W_{d,n} \sim \text{Multi}(B_{Z_{d,n}})$

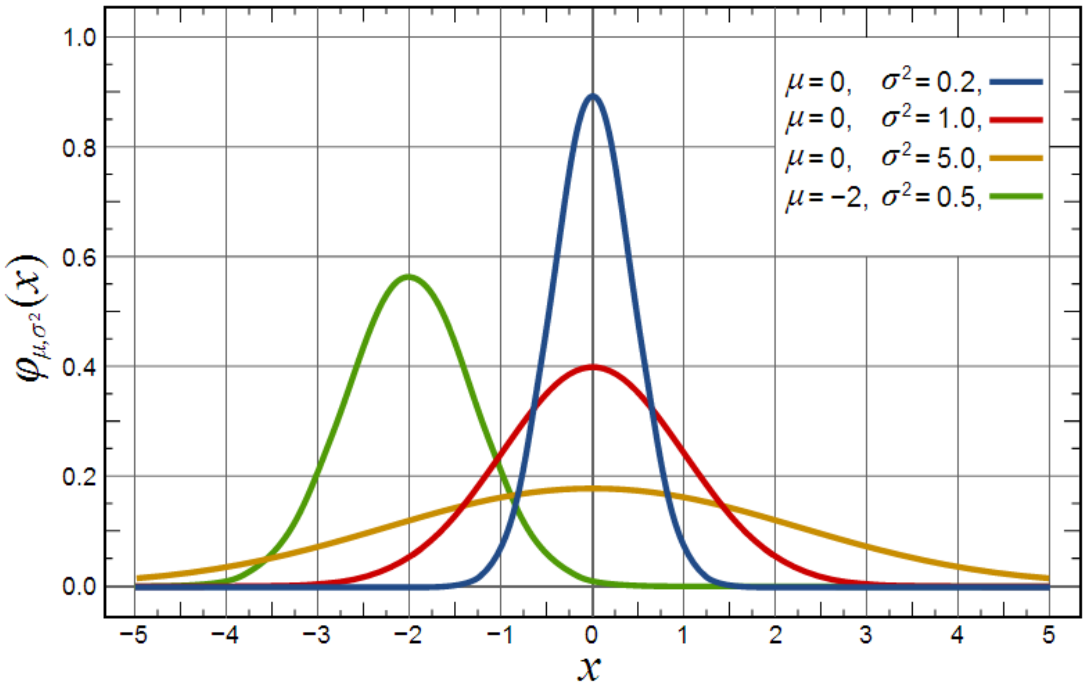
Latent Dirichlet Allocation

- Need to calculate posterior: $P(\theta_{1:D}, Z_{1:D,1:N}, \beta_{1:K} | W_{1:D,1:N}, \alpha, \eta)$
  - $\propto p(\theta_{1:D}, Z_{1:D,1:N}, \beta_{1:K}, W_{1:D,1:N}, \alpha, \eta)$
  - Normalization factor, $\int_\beta \int_\theta \sum_Z p(..)$, is intractable
  - Need to use approximate inference.
    - Gibbs Sampling
- Drawback
  - No intuitive relationship between topics.
- Challenge
  - Develop method similar to LDA with relationships between topics.

Normal or Gaussian Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Continuous distribution
    - Symmetrical and defined for $-\infty < x < \infty$
- Parameters: $\mathcal{N}(\mu, \sigma^2)$
    - $\mu \equiv \text{mean}$
    - $\sigma^2 \equiv \text{variance}$
    - $\sigma \equiv \text{standard deviation}$
- Estimation from Data: $X = \{x_1 \ldots x_n\}$
    - $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$
    - $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$



Multivariate Gaussian Distribution: k dimensions

$$f(\boldsymbol{X}) = f_x(X_1 \ldots X_k) = \frac{1}{(2\pi)^{k/2}\sqrt{\det \Sigma}}\, e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X}-\boldsymbol{\mu})}$$

- $\boldsymbol{X} = [X_1 \ldots X_k]^T \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$
  - $\boldsymbol{\mu} \equiv k$ x 1 vector of means for each dimension
  - $\boldsymbol{\Sigma} \equiv k$ x k covariance matrix**.**

Example: 2D Case

- $\mu = E[\boldsymbol{X}] = \begin{bmatrix} E[x_1] \\ E[x_2] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$
- $\Sigma = \begin{bmatrix} E[(x_1-\mu_1)^2] & E[(x_1-\mu_1)(x_2-\mu_2)] \\ E[(x_1-\mu_1)(x_2-\mu_2)] & E[(x_2-\mu_2)^2] \end{bmatrix}$

2D Multivariate Gaussian:

- $\Sigma = \begin{bmatrix} \sigma_{X_1}^2 & \rho_{X_1,X_2}\sigma_{X_1}\sigma_{X_2} \\ \rho_{X_1,X_2}\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2 \end{bmatrix}$
  - Topic Correlations on Off Diagonal
- $\rho_{X_1,X_2}\sigma_{X_1}\sigma_{X_2} = E[(x_1 - \mu_1)(x_2 - \mu_2)] = \sum_{i=1}^{n} \frac{(X_{i,1}-\mu_1)(X_{i,2}-\mu_2)}{n}$
- Covariance matrix is diagonal!

Matlab Demo

…Back to Topic Models

- How can we adapt LDA to have correlations between topics.

- In LDA, we assume two things:
  - **Assumption 1:** Topics in a document are independent. $\theta_d \sim Dir(\alpha)$
  - **Assumption 2:** Distribution of words in a topic is stationary. $B_k \sim (\eta)$

- To sample topic distributions for topics that are correlated, we need to correct assumption 1.

Exponential Family of Distributions

- Family of distributions that can be placed in the following form:

$$f(x|\theta) = h(x) \cdot e^{\eta(\theta) \cdot T(x) - A(\theta)}$$

- Ex: Binomial distribution: $\theta = p$

$$f(x|\theta) = \binom{n}{x} p^x (1-p)^{n-x}, x \in \{0,1,2, \dots, n\}$$

- $\boxed{\eta(\theta) = \log \frac{p}{1-p}}$ $\quad h(x) = \binom{n}{x}, \quad A(\theta) = n \log 1 - p, \quad T(x) = x$

$$f(x) = \binom{n}{x} e^{x \cdot \log\left(\frac{p}{1-p}\right) + n \cdot \log(1-p)}$$

Natural Parameterization

Categorical Distribution

- Multinomial n=1:
  - $f(x_1) = \theta_1;\ \ f(Z_1) = \theta^T \cdot Z_1$
  - where $Z_1 = [1\ 0\ 0..0]^T$ (**Iverson Bracket or Indicator Vector)**
  - $z_i = 1$
- Parameters: θ
  - $\theta = [p_1\ p_2\ p_3]$, where $\sum_i p_i = 1$
  - $\theta' = \left[\frac{p_1}{p_k} \frac{p_2}{p_k}\ 1\right]$
  - $\log \theta' = \left[\log \frac{p_1}{p_k}\ \log \frac{p_2}{p_k}\ \ 1\right]$

Exponential Family Multinomial With N=1

- $\boldsymbol{Recall}$: $f(Z_i|\theta) = \theta^T \cdot Z_i$
- We want: $f(x|\theta) = h(x) \cdot e^{\eta(\theta)\cdot T(x) - A(\theta)}$

- $f(Z_i|\eta) = e^{\eta^T Z_i - \log \sum_{i=1} e^{\eta_i}} = \frac{e^{\eta^T \cdot Z_i}}{\sum_{i=1} e^{\eta_i}}$
- *Note: k-1 independent dimensions in Multinomial*

- $\eta' = [\log \frac{p_1}{p_k} \ \log \frac{p_2}{p_k} \ldots .0], \eta'_i = \log \frac{p_i}{p_k}$

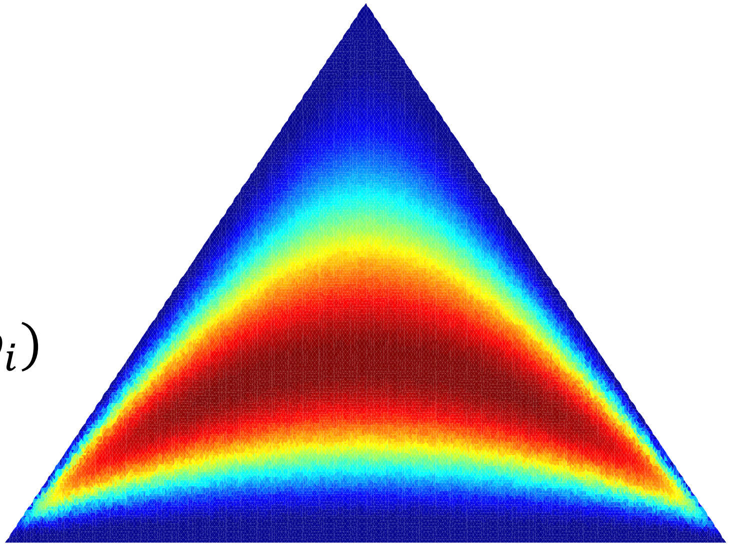
- $f(Z_i|\eta') =\cdot \frac{e^{\eta'^T \cdot Z_i}}{1+\sum_{i=1}^{k-1} e_i^{\eta_i'}}$

Verify: Classroom participation

- Given: $\eta = [\log\frac{p_1}{p_k}\ \log\frac{p_2}{p_k} \ldots 0]$
- Show: $f(Z_i|\theta) = \theta^T \cdot Z_i = e^{\eta^T Z_i - \log\sum_{i=1} e^{\eta_i}}$

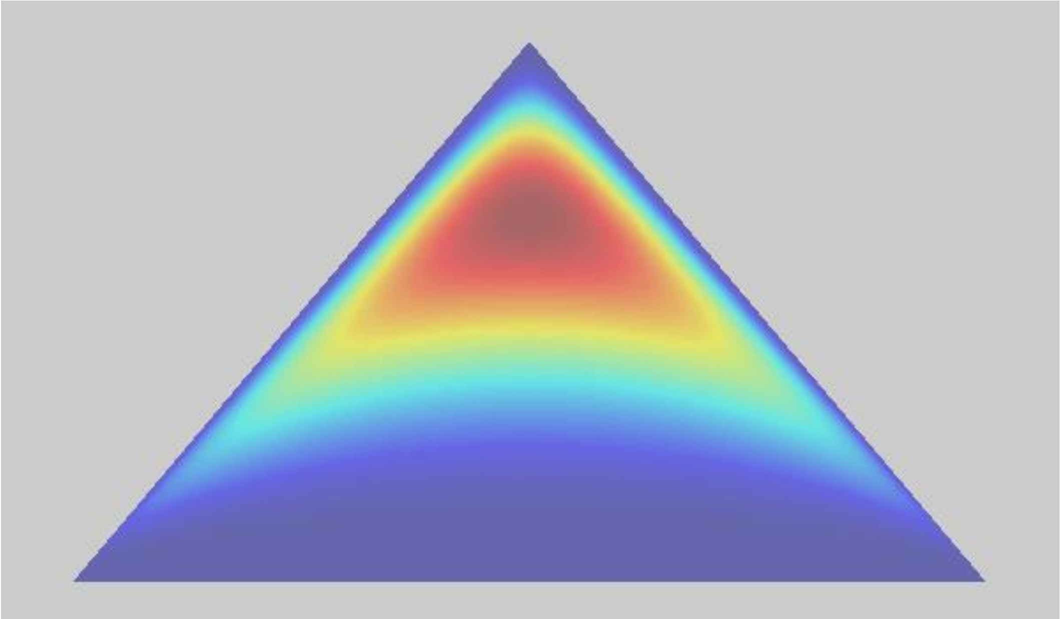
Intuition and Demo

- Can sample η from any number of places.
  - Choose normal (allows for correlation between topic dimensions)
- Get a topic distribution for each document by sampling: $\eta \sim \mathcal{N}_{k-1}(\mu, \sigma)$
  - What is the μ
    - Expected deviation from last topic: $\log\left(\frac{p_i}{p_k}\right)$
    - Negative means push density towards last topic ($\eta_i < 0, p_k > p_i$)
  - What about the covariance
    - Shows variability in deviation from last topic between topics.

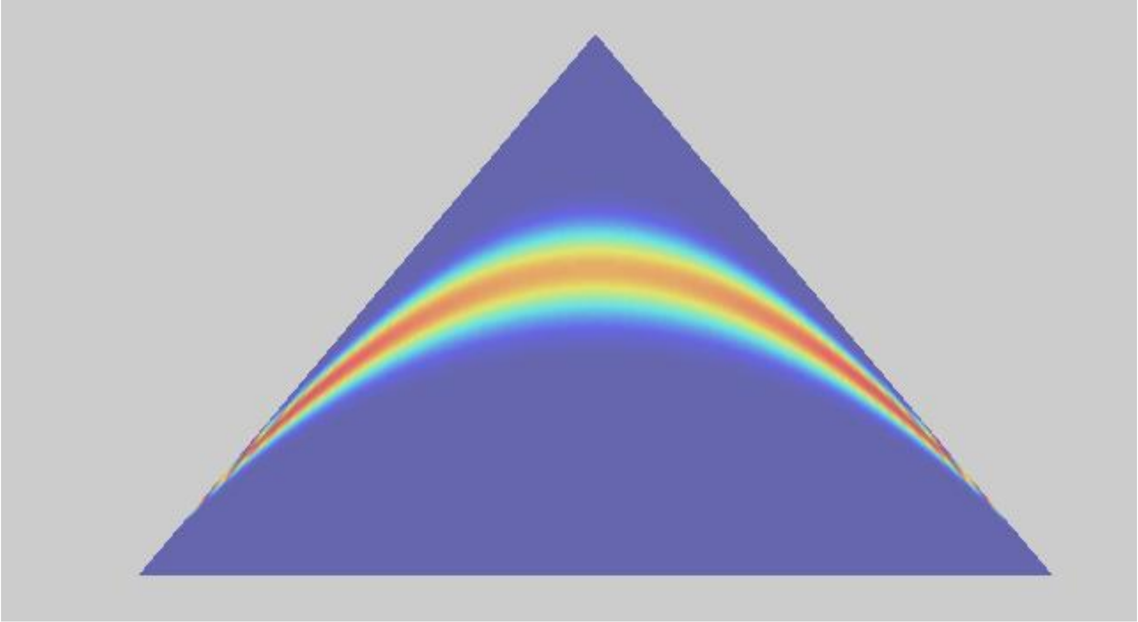


$\mu = [0\ 0]^T, \sigma = [1\ 0; 0\ 1]$

Favoring Topic 3

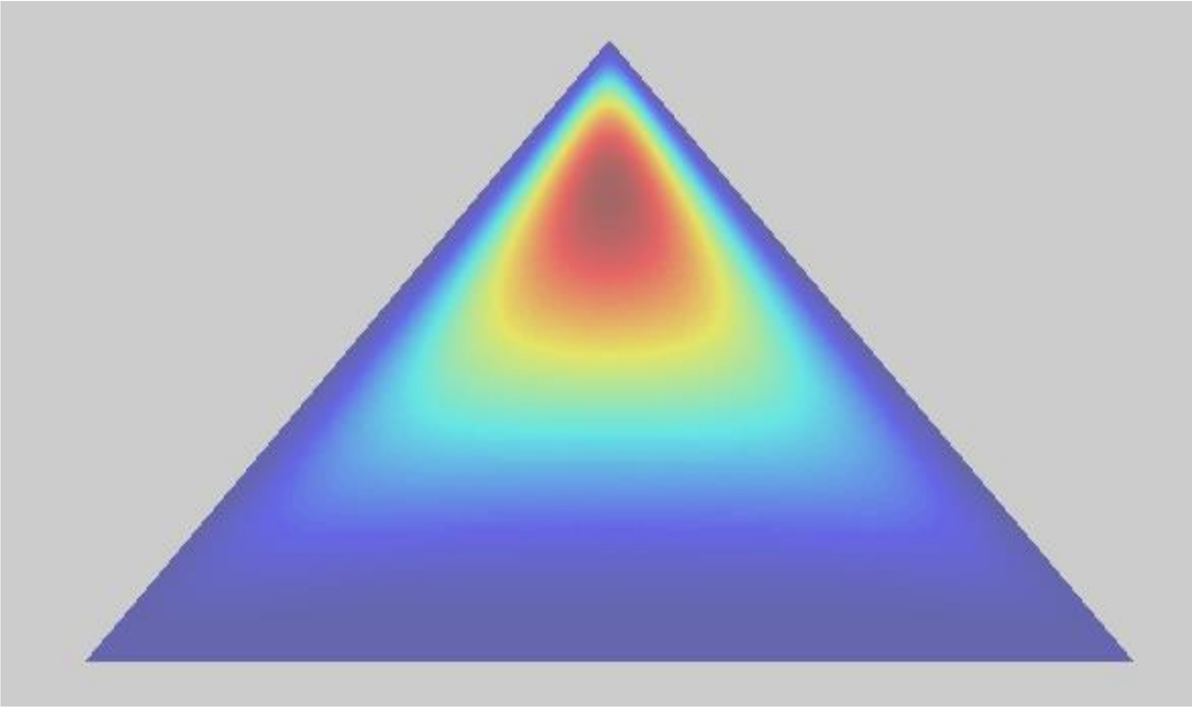


$\mu = [-0.9, -0.9], \qquad \Sigma = [1\ 0; 0\ 1]$



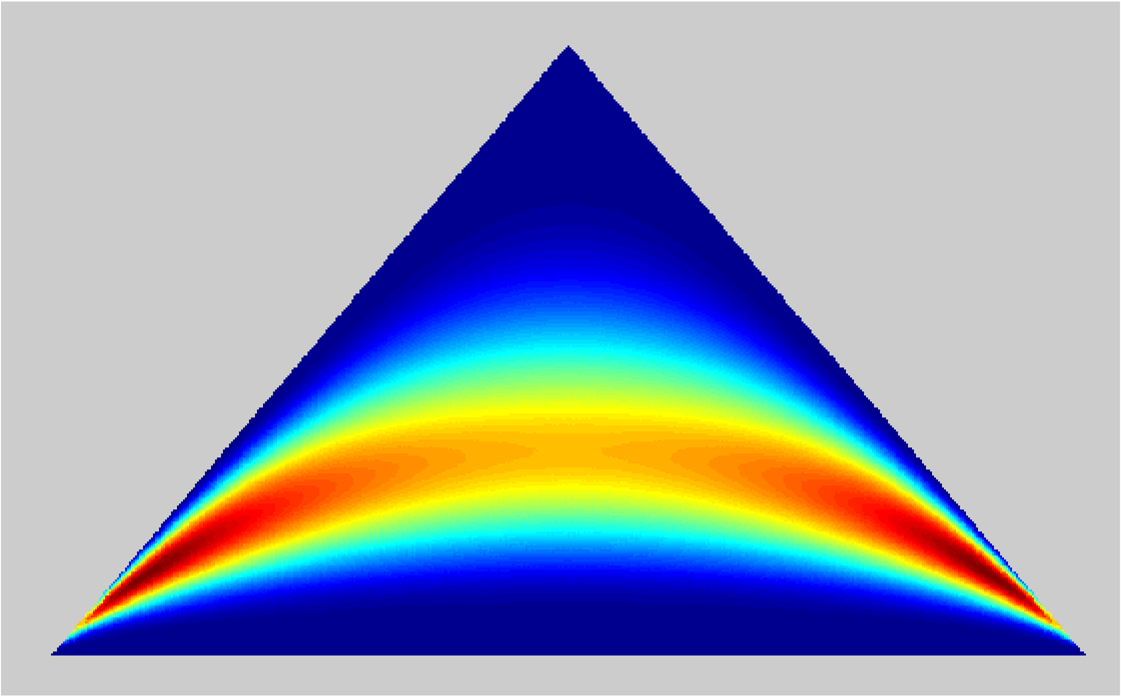
$\mu = [-0.9, -0.9], \qquad \Sigma = [1\ \ -0.9; -0.9\ 1]$

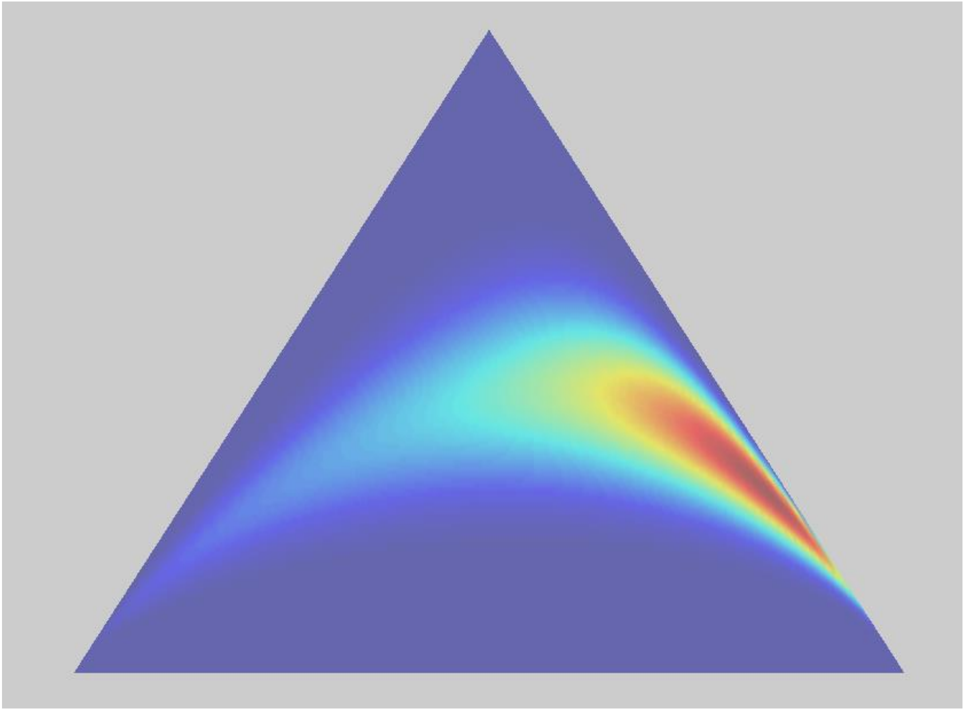
Favoring Topic 3:



$$\mu = [-0.9, -0.9], \qquad \Sigma = [1\ 0.4; 0.4\ 1]$$

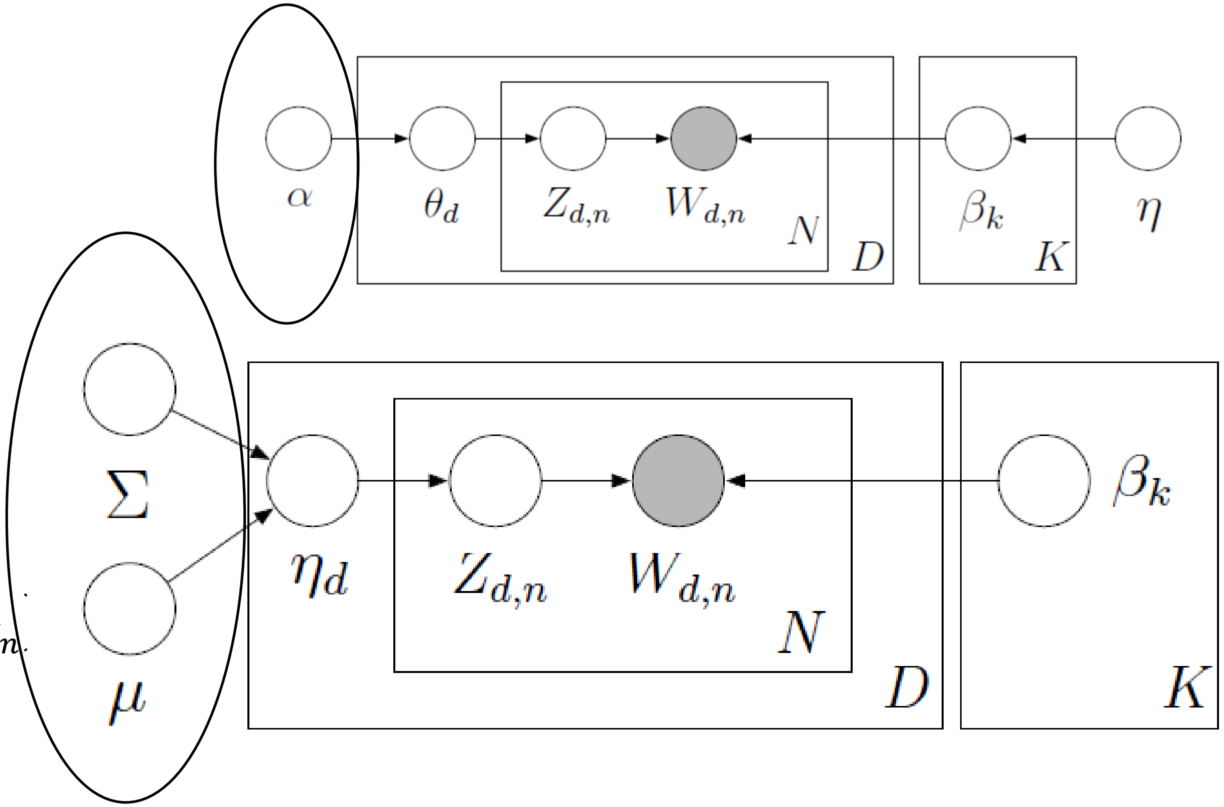
Exercises





Correlated Topic Model

- Algorithm:
- $\forall d \in D$
  - Draw $\eta_d | \{\mu, \Sigma\} \sim \mathcal{N}(\mu, \Sigma)$
  - $\forall\, n \in \{1 \ldots N\}$:
    - Draw topic assignment
      - $Z_{n,d} | \eta_d \sim \text{Categorical}\big(f(\eta_d)\big)$
    - Draw word
      - $W_{d,n} | \{Z_{d,n}, \beta_{1:K}\} \sim \text{Categorical}\big(\beta_{Z_n}$



- Parameter Estimation:
  - Intractable
  - User variational inference (later)

Evaluation I: CTM on Test Data

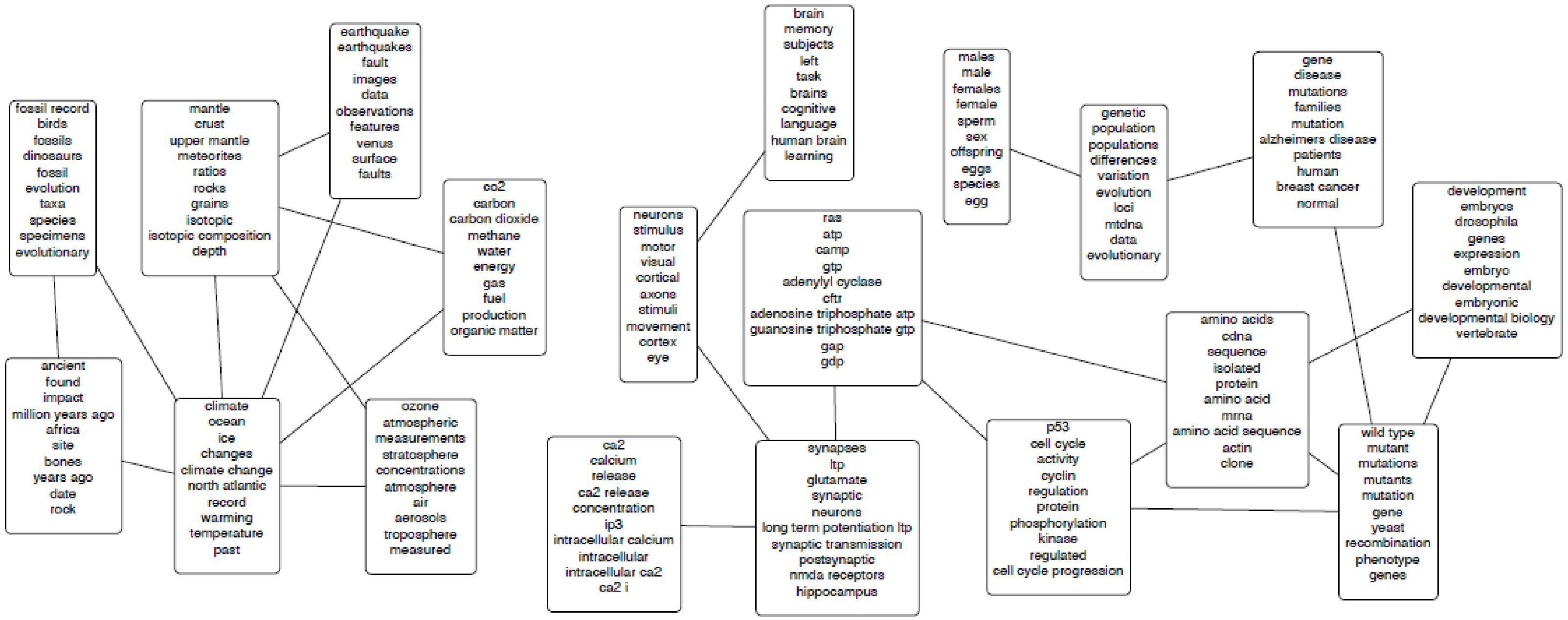
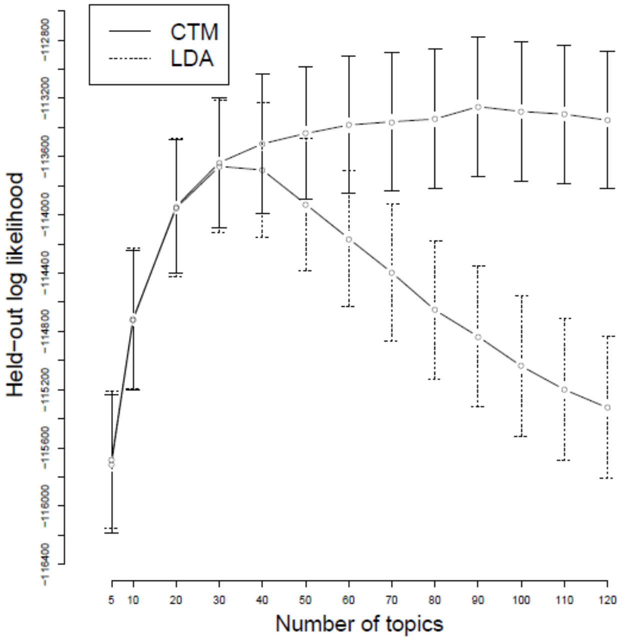
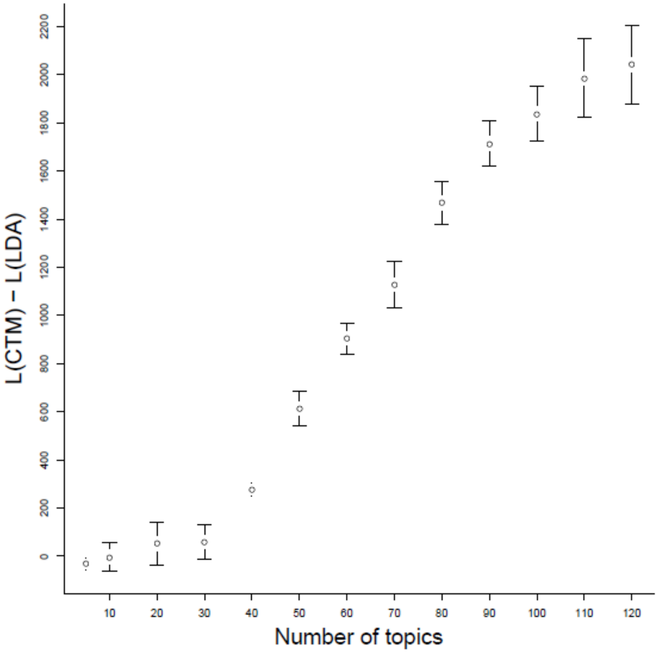


Figure 2: A portion of the topic graph learned from 16,351 OCR articles from *Science*. Each node represents a topic, and is labeled with the five most probable phrases from its distribution (phrases are found by the "turbo topics" method [3]). The interested reader can browse the full model at `http://www.cs.cmu.edu/~lemur/science/`.

Evaluation II: 10-Fold Cross Validation LDA vs CTM

- ~1500 documents in corpus.
- ~5600 unique words
  - After pruning
- Methodology:
  - Partition data into 10 sets
  - 10 fold cross validation
  - Calculate the log likelihood of a set, given you trained on the previous 9 sets, for both LDA and CTM.
- Right(L(CTM) - L(LDA))
- Left(L(CTM) – L(LDA))

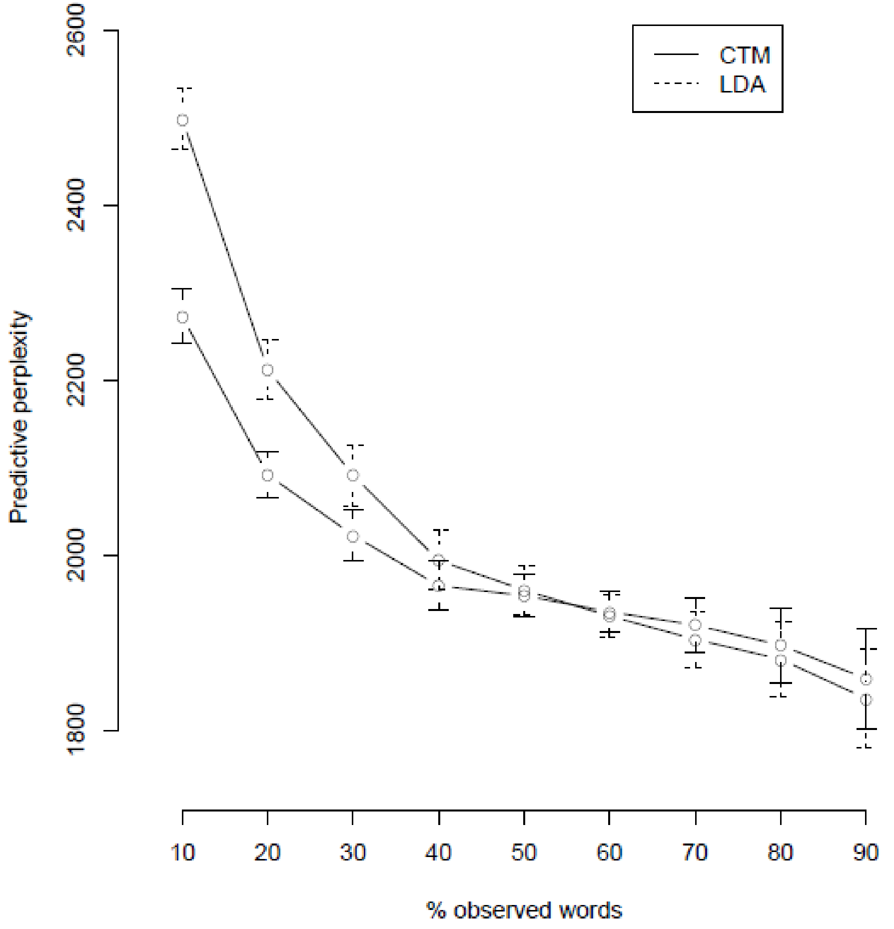




CTM shows a much higher log likelihood as the number of topics increases.

Evaluation II: Predictive Perplexity

- Perplexity measure ≡ expected number of equally likely words
  - Lower perplexity means higher word resolution.
- Suppose you see a percentage of words in a document, how likely is the rest of the words in the document according to your model?
- CTM does better with lower #'s of observed words.
  - Able to infer certain words given topic probabilities.



Conclusions

- CTM changes the distribution from which hyper parameters are drawn, from a Dirichlet to a logistic normal function.
    - Very similar to LDA
- Able to model correlations between topics.
- For larger topic sizes, CTM performs better than LDA.
- With known topics, CTM is able to infer words associations better than LDA.