JOANNEUM RESEARCH

DIGITAL

Institute for Information and Communication Technologies

Computational Intelligence Seminar F

Topic Models
LDA and the Correlated Topic Models

Claudia Wagner
Graz, 21.1.2011

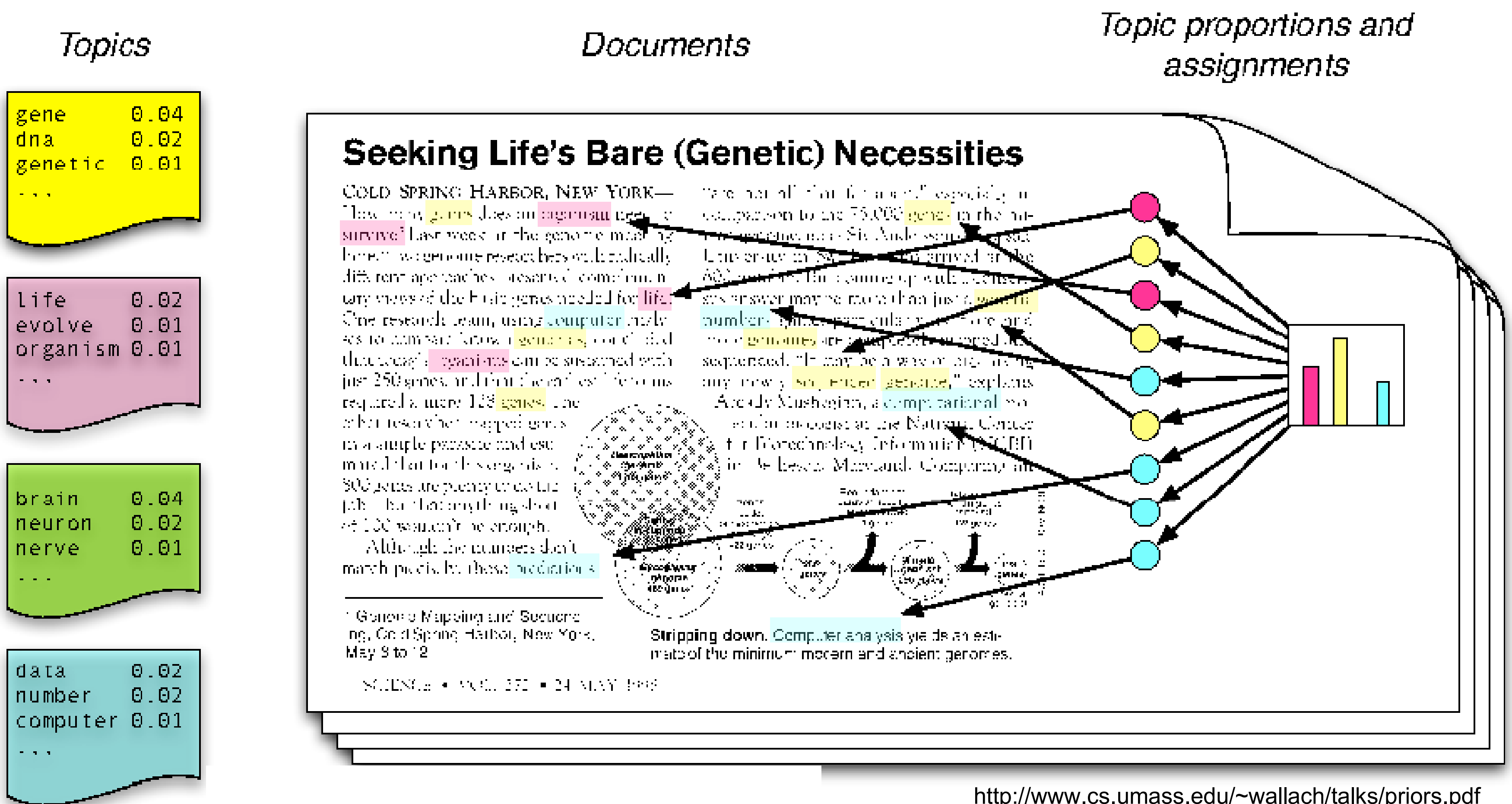
www.joanneum.at

Motivation

- Aim of Topic Models:
  - Large unstructured collection of document
  - Discover set of topics that generated the documents
  - Annotate documents with topics

www.betaversion.org/~stefano/linotype/news/26/

Topic Models
Generative Models

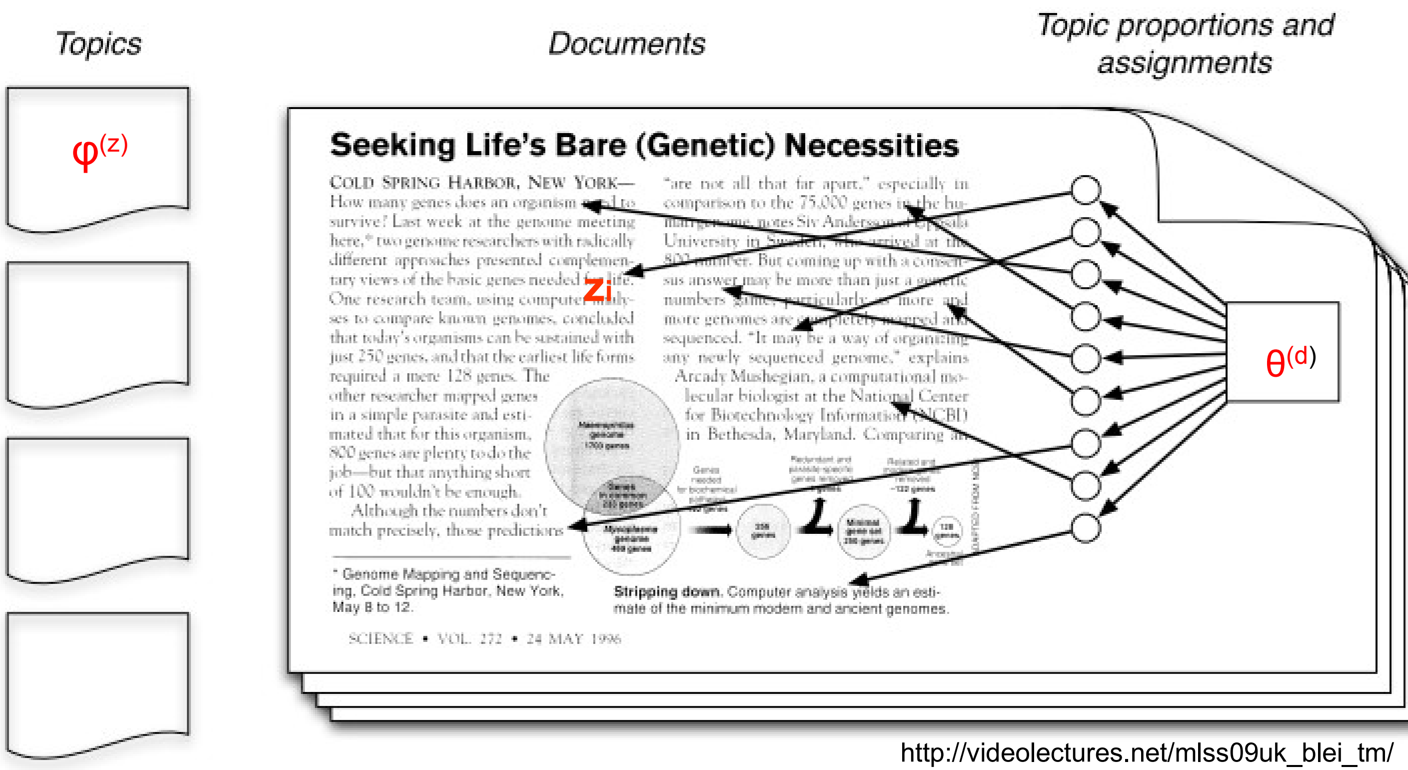


http://www.cs.umass.edu/~wallach/talks/priors.pdf

$$P(w_i) = \sum_{j=1}^{T} P(w_i | z_i = j) P(z_i = j)$$

JOANNEUM RESEARCH

Topic Models
Statistical Inference



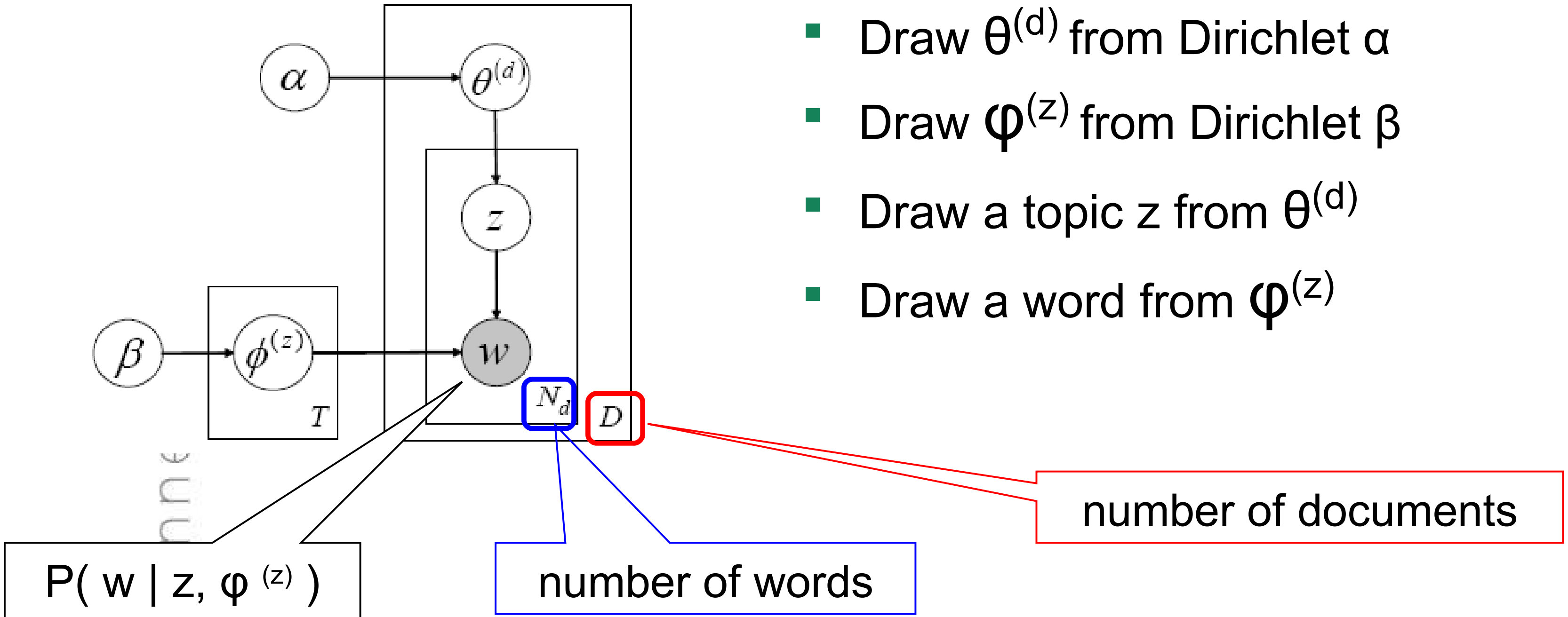
http://videolectures.net/mlss09uk_blei_tm/

Infer the hidden structure using posterior inference
P(hidden variables | observations, priors)

Situate new data into the estimated model
P(new doc | model)

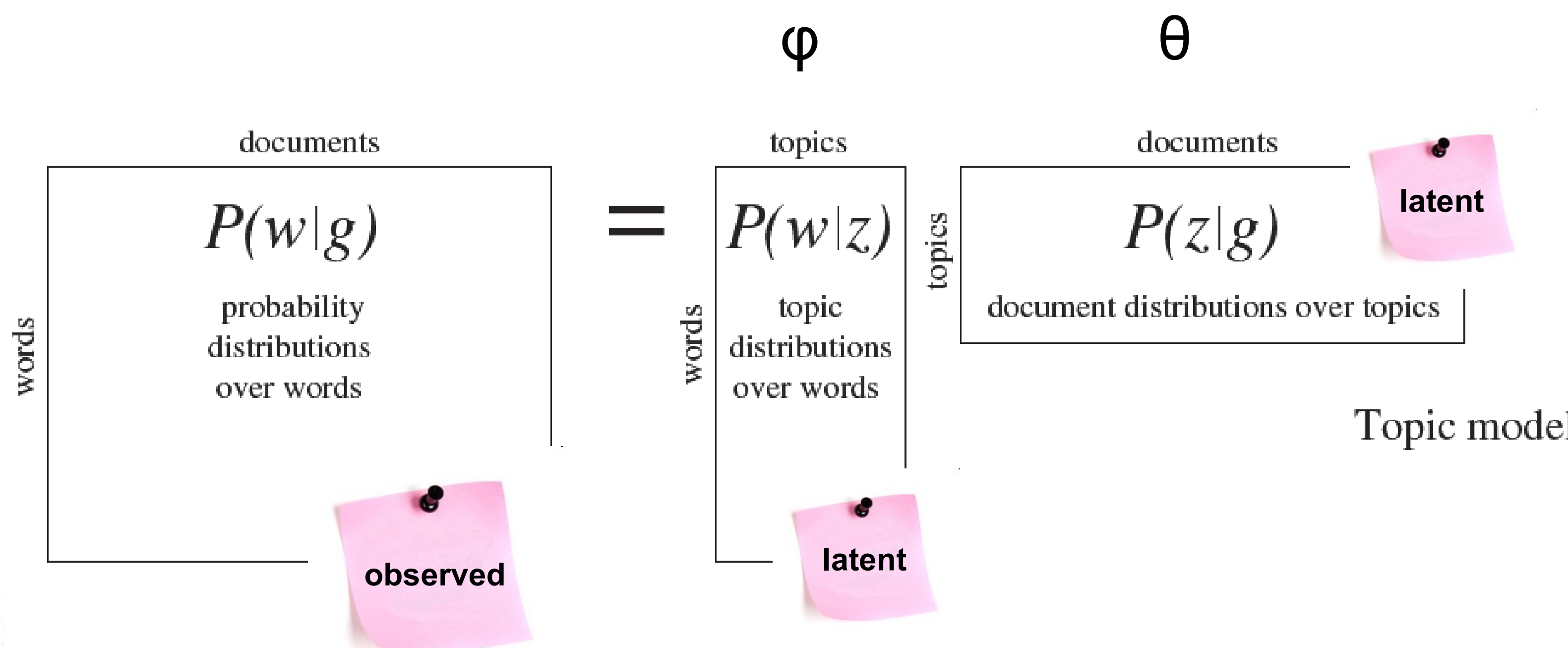
Latent Dirichlet Allocation (LDA)
(Blei et al, 2003)

LDA



- Draw $\theta^{(d)}$ from Dirichlet α
- Draw $\varphi^{(z)}$ from Dirichlet β
- Draw a topic z from $\theta^{(d)}$
- Draw a word from $\varphi^{(z)}$

Matrix Representation of LDA



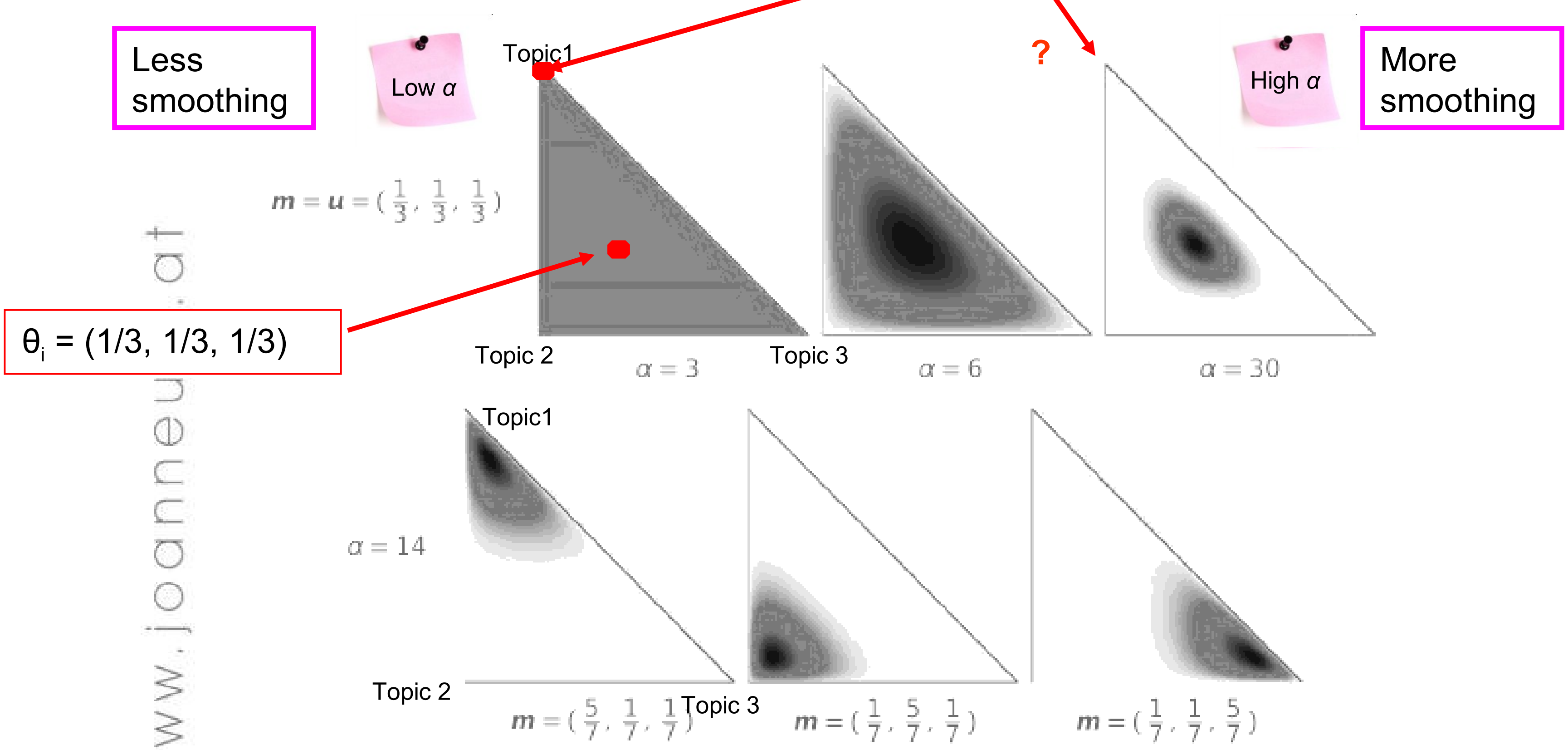
Dirichlet Distribution

- Dirichlet distribution is a "distribution over distributions"
- If we draw a sample from a Dirichlet distribution we get a positive vector that sums to one

- Parameters of Dirichlet:
  - Postive K-dimensional vector $\alpha = <\alpha_{1,}, \alpha_2, \ldots, \alpha_k>$
  - Concentration = $\sum_{i=0}^{K} \alpha_i$ → determines peakiness
  - Mean = $E[\theta_i | \alpha] = \frac{\alpha}{\sum_{i=0}^{K} \alpha_i}$ → determines peak location

Dirichlet Distribution

Examples with K=3



http://www.cs.umass.edu/~wallach/talks/priors.pdf

The larger the value of the concentration parameter alpha, the more evenly distributed is the resulting distribution!

JOANNEUM RESEARCH

Dirichlet Priors
α and β

- Dirichlet priors α and β are a conjugate priors of the parameters of the multinomial distribution over topics/words

- α is a force on the topic combinations
  - Low α forces to pick for each doc a topic distribution which favors few topics
  - High α allows documents to have similar, smooth topic proportions

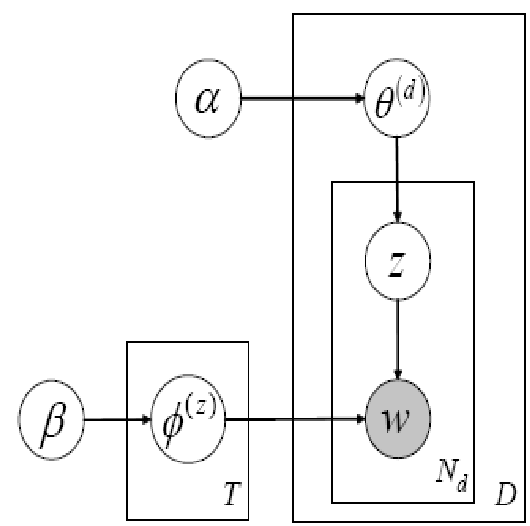
- β is a force on the word combinations
  - Low β forces each topic to favors few words
  - High β allows topics to be less distinct

Posterior Distribution of LDA

- Posterior distribution is the cond. distribution of the hidden variable given the observations
- P (θ,φ, z | w, α, β, .)
- Per document posterior

$$\frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_{\theta} p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

- LDA posterior is intractable
  - Note: Hidden variables are dependent when conditioned on data.
- Approximate LDA posterior
  - Variational Methods
  - Gibbs Sampling
  - …



(Collapsed) Gibbs Sampling

- Define a Markov Chain whose stationary distribution is the posterior of interest
- Space of Markov Chain is space of possible configurations of hidden variables
- Draw iteratively independent samples from the conditional distribution of each hidden variable given the observations and the current state of all other hidden variables

$$P\left(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot\right) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i t}^{DT} + T\alpha}$$

- When chain has “burned in” collect samples to approximate posterior

Collapsed Gibbs Sampling

- Exploit conjugacy

$$P(\theta \mid z_i, w_i) \sim Dir(\alpha + n(z_i))$$

Current state of hidden vars

Observation

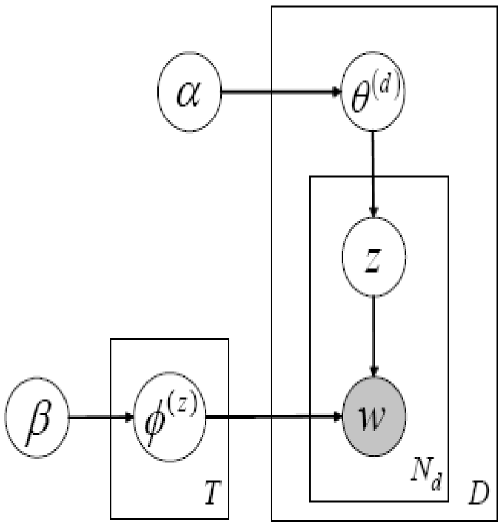
Topic counts

- We can integrate out θ if we condition on all other topic assignments z_{-i} while sampling z_i

$$P(z_i = t \mid z_{-i}, w_i) \sim$$
$$P(w_i \mid z_i = t, z_{-i}, \beta_{1..k}) \; * \; \Pi_{i=1.K} \; \alpha + n(z_i)$$

How likely is the word wi for topic t?

How likely is topic t?



Variational Methods

- Introduced a proposal distribution of the latent variables with free variational parameters v
- The latent variables are independent in proposal distribution
- Each latent variable has its own variational parameter
- Optimize those parameters to tighten this bound

$$\log p(x_{1:N}) \geq \mathrm{E}_{q_\nu}[\log p(z_{1:M}, x_{1:N})] - \mathrm{E}_{q_\nu}[\log q_\nu(z_{1:M})]$$

$$\nu_m = \mathrm{E}_{q_\nu}[g_m(\mathbf{Z}_{-m}, \mathbf{x})]$$

Why does LDA work?

- Semantically related words tend to co-occur
- LDA performs a smooth co-occurence analysis

- Why does the LDA posterior put “topical” words together? What keeps us away from putting all words in all topics?
  - **Priors** ensure that documents are penalized for equaliy favoring all topics and that topics are penalized for equaliy favoring all words.
  - **Likelihood of data** → P(w | z) → Word probabilities are maximized by dividing the words among the topics. If many words are likely for one topic, they will have all small probability → cond. probabilities of words given a topic sum to 1.

Correlated Topic Models (CTM)
(Blei et al, 2007)

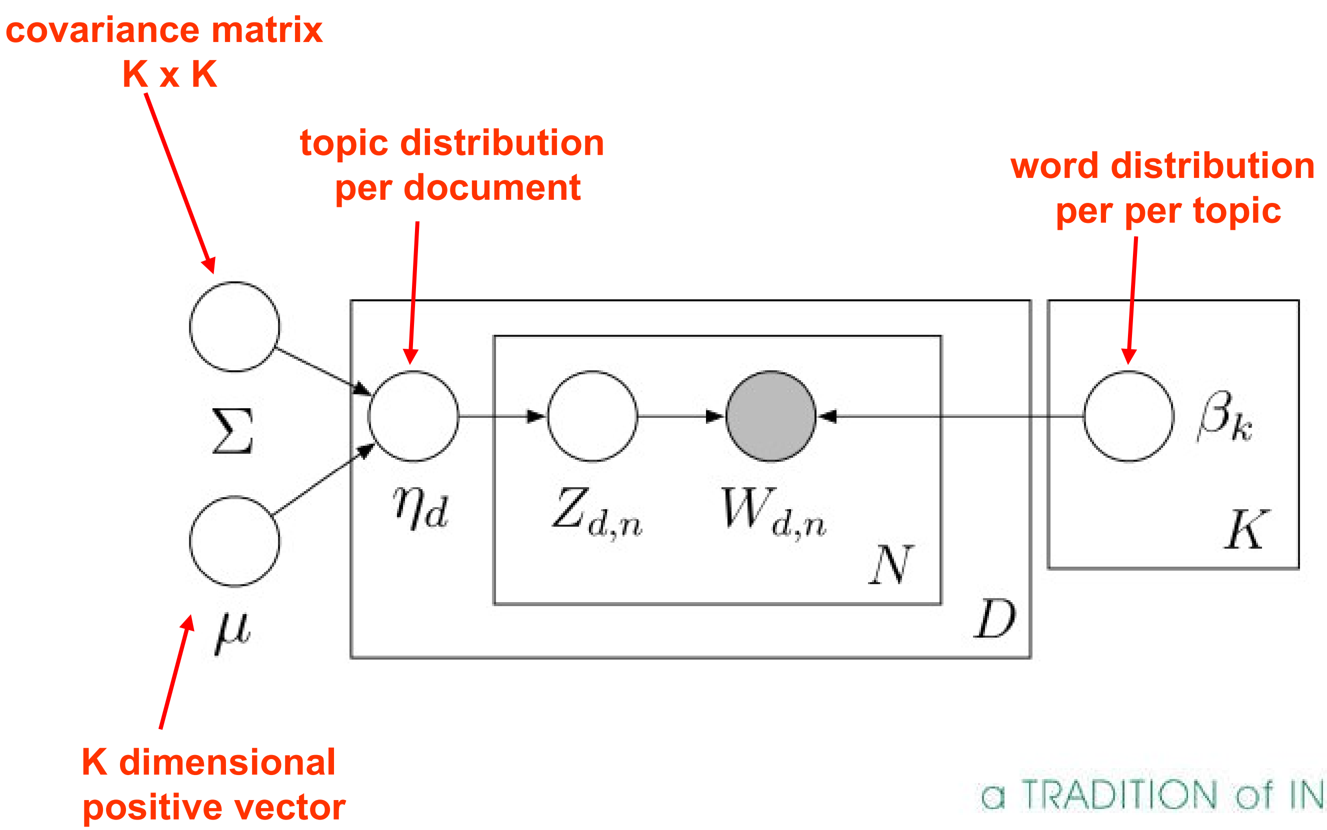
CTM

- Limitations of LDA:
  - LDA fails to directly model correlation between the occurrence of topics!
  - LDA makes an independence assumption between topics
  - Why?
    - Dirichlet prior on topic proportion
    - Under a Dirichlet prior the components of the proportion vector are nearly independent

- Intuition behind CTM:
  - Presence of one latent topic may be correlated with the presence of another
  - e.g., a document about the topic “semantic web” is more likely to be also about the topic “information retrieval” than about the topic “genetics”

CTM

- CTM is equal to LDA except that topic proportions are drawn from a logistic normal rather than a Dirichlet



Logistic normal distribution

- Logistic normal distribution is obtained by
  - Drawing for each doc a K-dimensional vector η_d from a multivariate Gaussian distr with mean μ and covariance matrix Σ

    $\eta_d \sim N(\mu, \Sigma)$

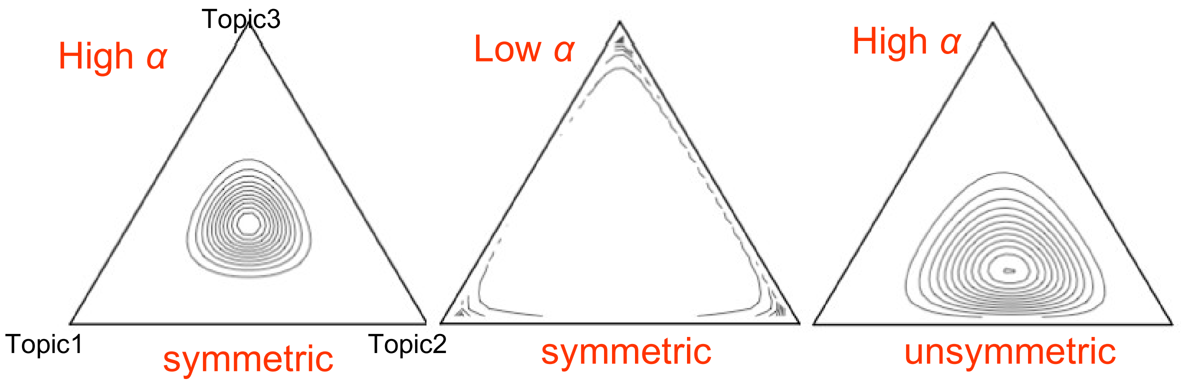
  - f (η) maps a natural parameterization of the topic proportions to the mean parameterization:

$$\theta = f(\eta) = \frac{\exp\{\eta\}}{\sum_i \exp\{\eta_i\}}$$

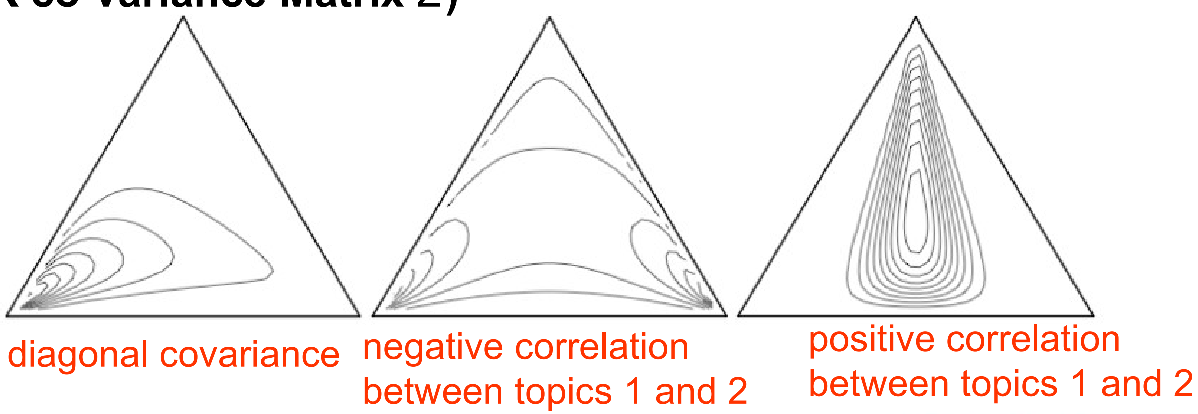
  - i.e., map η onto a simplex so that it sums to 1
- The covariance of the Gaussian induces dependencies between the components of the transformed vector

Dirichlet versus Logistic Normal

Dirichlet distribution (Paramter: postive K-dim vector α)



Logistic Normal distribution (Paramter: postive K-dim vector μ, K x K co-variance Matrix Σ)



http://www.cs.princeton.edu/~blei/modeling-science.pdf

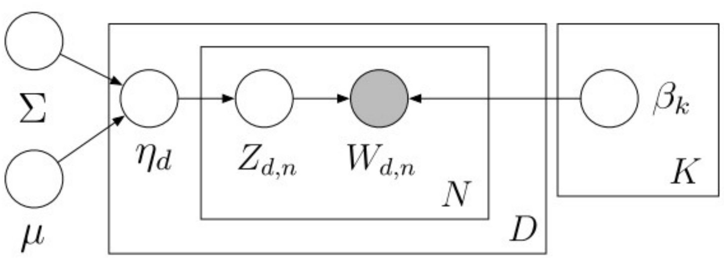
Posterior of CTM

N … Num of words
K … Num of topics

$$p(\boldsymbol{\eta}, z | \boldsymbol{w}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$= \frac{p(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} p(z_n|\boldsymbol{\eta}) p(w_n|z_n, \boldsymbol{\beta}_{1:K})}{\int p(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \underline{\prod_{n=1}^{N} \sum_{z_n=1}^{K}} p(z_n|\boldsymbol{\eta}) p(w_n|z_n, \boldsymbol{\beta}_{1:K})\, d\boldsymbol{\eta}}$$

- Not tractable
- Why?
  - Sum over the K values of each z occurs inside the product over words
  - Logistic normal is not conjugate to the multinomial



Approximate Posterior of CTM

Consequences of non-conjugacy:

- We cannot use many of the MCMC sampling techniques that have been developed for Dirichlet-based mixed membership models → we cannot integrate out topic mixture η (previously θ)

- Use Variational methods (Blei et al, 2007)

- Gibbs Sampling for logistic normal Topic Models (Mimno et al., 2008)

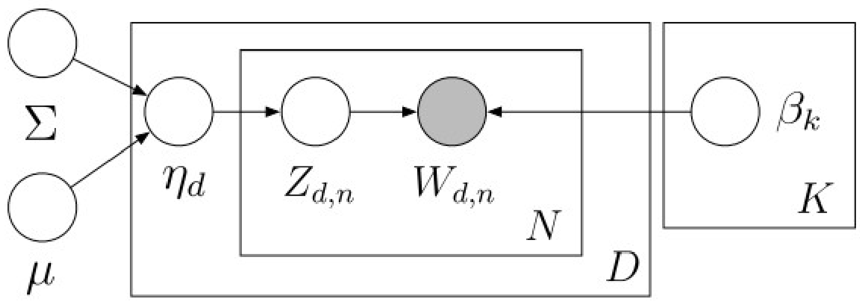
Gibbs Sampling

- We cannot integrate out η if we condition on all other topic assignments z_{-i} while sampling z_i

How likely is the word wi for topic t?

$$P(z_i = t \mid z_{-i}, w_{i,} \{\eta_d\}_{d=1..D}, \beta) \sim P(w_i \mid z_i = t, z_{-i}, w_{-i}, \beta) * \exp(\eta_{d,t})$$

How likely is topic t for this document?



Empirical Results
Comparing CTM and LDA
(Blei et al, 2007)

Experimental Setup

- 16 351 Science articles
- Estimated a CTM and LDA model with 100 topics

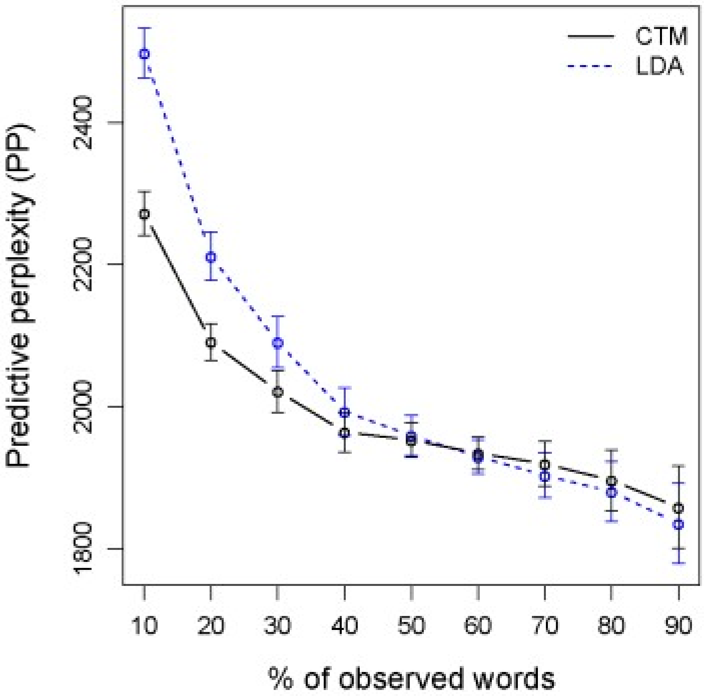
- Compare predictive performance of CTM and LDA
  - Observe P words from a document
  - Which model provides a better predictive distribution of the remaining words $P(w|w_{1:P})$?

$$\mathrm{Perp}(\Phi) = \left(\prod_{d=1}^{D} \prod_{i=P+1}^{N_d} p(w_i|\Phi, w_{1:P}) \right)^{-1/(\sum_{d=1}^{D}(N_d - P))}$$

Num of words per document

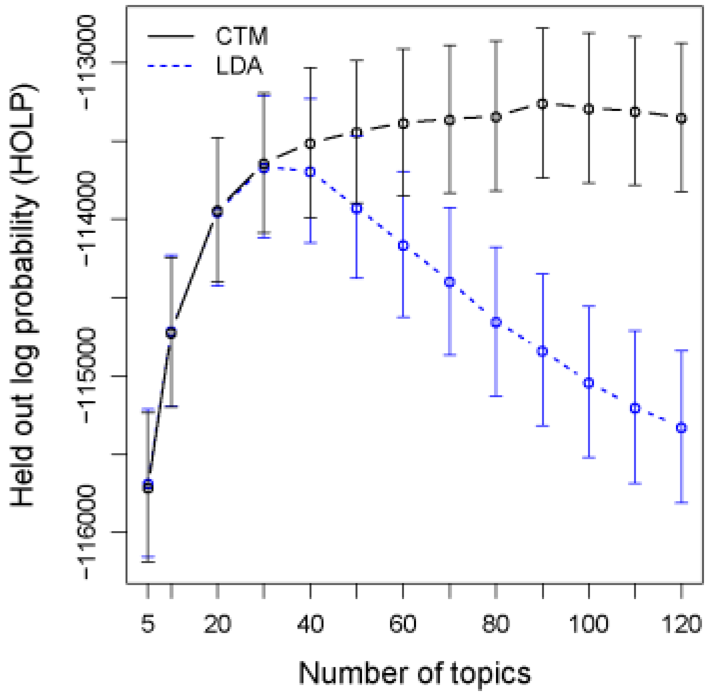
Num of observed words

Results



- When a small number of words have been observed, there is less uncertainty about the remaining words under the CTM than under LDA
- CTM seems to be able to use its knowledge about topic correlations and better predicts words

Results



- Held-out log probability
  - Is the probability that the model gives to hold-out-data
  - The higher the better
- The CTM provides a slightly better fit than LDA and supports more topics;
- The likelihood for LDA peaks near 30 topics
- The likelihood for the CTM peaks close to 90 topics.
- Interpretation: corpus contains around 30 independent topics

References

- **David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3: 993-1022.**
- **Blei, D. and Lafferty, J. (2007). A correlated topic model of Science. Annals of Applied Statistics, 1(1):17–35.**
- **David Mimno, Hanna Wallach, Andrew McCallum (2008). Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors. NIPS Workshop on Analyzing Graphs.**