Hot Topics in Machine Learning
09 – Topic Models

Prof. Dr. Rainer Gemulla, Dr. Laura Dietz

Universität Mannheim

HWS 2015
Why Topic Models?

- Topic models are a prime example of latent variable inference in Bayesian networks.
- Topic models are easy to customize to different properties of data sets.
- Revisit material from lecture 04:
  - Beta-Binomial
  - Dirichlet-Multinomial
- Focus on generative models, processes and Bayesian networks.
- Latent variable inference with Gibbs sampling.
- Practice expressing modeling assumptions in this mathematical framework.
Outline

1. Topic identification and applications

2. Modelling assumptions in Latent Dirichlet Allocation

3. Latent variable inference using Gibbs sampling

4. Extensions of topic models

5. Conclusion
Problem: Topic Identification

- Input: a corpus of documents.
- Each document is represented by a sequence of words.
- The goal is to identify reoccurring topics in the documents.
- A topic is represented as a distribution over words.

Idea:
- Assume a set of \( k \) latent topics.
- Each document has a distribution over characteristic topics.
- Each topic has a distribution over characteristic words.
Motivation: Exploration of publications

- Primary goal is to get an insight into an unknown document collection, e.g. [Blei 2009]

- Also see demo on visualizing publications below.
Exploration of publications

- Topics in science publications [Blei 2009]

  computer  chemistry  cortex  orbit  infection
  methods  synthesis  stimulus  dust  immune
  number  oxidation  fig  jupiter  aids
  two  reaction  vision  line  infected
  principle  product  neuron  system  viral
  design  organic  recordings  solar  cells
  access  conditions  visual  gas  vaccine
  processing  cluster  stimuli  atmospheric  antibodies
  advantage  molecule  recorded  mars  hiv
  important  studies  motor  field  parasite

Figure 1. Five topics from a 50-topic LDA model fit to Science from 1980–2002.
Model assumptions on how topics arise

- **Bag-of-words:**
  The order in which words occur does not contribute towards finding topics. Only their frequency matters.

- **Co-occurrence:**
  A strong signal for two words belonging to a topic is the co-occurrence of both words in many documents. Observing co-occurrence increasing the probability of both words under the topic-characteristic distribution $\phi$.

- **Sparseness:**
  Each document arises from only a few topics, i.e., $\theta$ is sparse.

- **Model can be extended with further information.**
- **Can also be applied to non-textual data**
  (replace ‘word’ with ‘item’ and ‘document’ with ‘container’).
Including Geographic word usage (1)

- Informing topics in twitter with geo-tags [Eisenstein 2010]

<table>
<thead>
<tr>
<th>Location</th>
<th>“basketball”</th>
<th>“popular music”</th>
<th>“daily life”</th>
<th>“emoticons”</th>
<th>“chit chat”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td><strong>PISTONS KOBE LAKERS</strong> game DUKE NBA CAVS STUCKEY JETS KNICKS</td>
<td>album music beats artist video #LAKERS ITUNES tour produced vol</td>
<td>tonight shop weekend getting going chilling ready discount waiting iam</td>
<td>:d haha :d :( ; :p xd :/ hahaha hahah</td>
<td>lol smh jk yea wdy coo ima wassup somethin jp</td>
</tr>
<tr>
<td>N. California</td>
<td>CELTICS victory BOSTON CHARLOTTE</td>
<td>playing daughter PEARL alive war comp</td>
<td>BOSTON</td>
<td>;p gna loveee</td>
<td>ese exam suttin sippin</td>
</tr>
<tr>
<td>New York</td>
<td>THUNDER KINGS GIANTS pimp trees clap</td>
<td>SIMON dl mountain seee</td>
<td>6am OAKLAND</td>
<td>pues hella koo SAN fckn</td>
<td>hella flirt hut iono OAKLAND</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>1KOBE 1LAKERS AUSTIN</td>
<td>#LAKERS load HOLLYWOOD imm MICKEY TUPAC</td>
<td>omw tacos hr HOLLYWOOD</td>
<td>af papi raining th bomb coo HOLLYWOOD</td>
<td>wyd coo af nada tacos messin fasho bomb</td>
</tr>
<tr>
<td>Lake Erie</td>
<td>CAVS CLEVELAND OHIO BUCKS od COLUMBUS</td>
<td>premiere prod joint TORONTO onto designer CANADA village Burr</td>
<td>stink CHIPOTLE tipsy</td>
<td>;d blvd BIEBER hve OHIO</td>
<td>foul WIZ salty excuses lames officer lastnight</td>
</tr>
</tbody>
</table>
Including Geographic word usage (2)

- Informing topics in twitter with geo-tags [Eisenstein 2010]
Unsupervised texture learning

- Extracting patterns in image collections
Bioinformatics

- Associating genes to their biological functions

[Konietzny, 2011]
Different approaches towards topic identification

- **Matrix factorization (Latent Semantic Analysis)**
  \[ d \times w = d \times t \times t \times w \]

- **K-means**
  Represent each document with its bag-of-words vector (a vector of length \(|V|\)) Find \(k\) vectors of length \(|V|\) and assign documents to minimize the squared error. Each mean vector represents one topic.

- **word2vec or paragraph vector**
  Neural network based word embedding models

- **Topic Models**
  Generative model approach (Directed Bayesian Model)
Why not a discriminative model?

- Discriminative models covered:
  - Logistic Regression
  - linear-chain and general CRFs

- Discriminative models are successful when supervision (i.e. $Y$ labels) are available during training.

- But here supervision is not available. I.e., only have $X$ but no $Y$.

- CRFs are not applicable here.
Basic topic model

• Topic models are
  ▶ unsupervised model (i.e., no $\mathcal{Y}$ given during training)
  ▶ clustering model: given data $\mathcal{X}$, find topics $\mathcal{Y}$.
  ▶ generative model: $P(\mathcal{X}, \mathcal{Y}) = P(\mathcal{X}|\mathcal{Y})P(\mathcal{Y})$

• A class of topic models with many adaptations. (More later)

• Most basic topic model focuses on a corpus documents.

• Different variations on the basic model:
  ▶ Probabilistic Semantic Indexing [Hofmann, 1999].
    No priors. Trained with Expectation Maximization (EM).
    Prior on $\theta$. Trained with variational inference.
  ▶ Latent Dirichlet Allocation [Griffiths and Steyvers, 2007].
    Priors on $\theta$ and $\phi$. Trained with collapsed Gibbs sampler.

• Next we discuss the approach of Griffiths and Steyvers.
Outline

1. Topic identification and applications

2. Modelling assumptions in Latent Dirichlet Allocation

3. Latent variable inference using Gibbs sampling

4. Extensions of topic models

5. Conclusion
Modeling one word

- Consider modeling only one word $w$ in one document $d$.

\[
P(w \mid d) = \sum_{z} P(w, z \mid d) \tag{1}
\]

\[
= \sum_{z} \left( P(w \mid z, d) P(z \mid d) \right) \tag{2}
\]

\[
= \sum_{z} P(w \mid z) P(z \mid d) \tag{3}
\]

- Introduce two sets of multinomial parameters:
  - Word-distribution per topic $\phi_z = P(w \mid z)$
  - Topic-distribution per document $\theta_d = P(z \mid d)$

\[
P(w \mid d) = \sum_{z} \text{Mu}(w \mid \phi_z) \text{Mu}(z \mid \theta_d) \tag{4}
\]
Modeling the corpus

- Given a corpus of documents $\mathcal{D}$.
- $\mathbf{w}_d$ is a sequence of words contained in document $d$, e.g., $\mathbf{w}_d = ["topic", "models", "are", "fun"]$

\[
P(\mathcal{D}) = \prod_{d \in \mathcal{D}} P(d) P(\mathbf{w}_d | d) \tag{5}
\]

\[
= \prod_{d \in \mathcal{D}} \left( P(d) \prod_{\mathbf{w}_d \in d} \sum_z \text{Mu}(\mathbf{w}_{di} | \phi_z) \text{Mu}(z | \theta_d) \right) \tag{6}
\]

- Auxiliary variable trick: Instead of marginalizing over $z$, introduce a latent variable $Z_{di}$ for every word $W_{di}$.

\[
P(\mathcal{D}, \mathbf{Z}) = \prod_{d \in \mathcal{D}} \left( P(d) \prod_{\mathbf{w}_d \in d} \text{Mu}(\mathbf{w}_{di} | \phi_{Z_{di}}) \text{Mu}(Z_{di} | \theta_d) \right) \tag{7}
\]
Estimating topics

- For given $\phi$ and $\theta$ parameters, Equation 7 defines the probability of the data and $Z$. Called the **Forward model**.

\[ P(D, Z) = \prod_{d \in D} \left( P(d) \prod_{w_{di} \in d} \text{Mu} (w_{di} | \phi_{Z_{di}}) \text{Mu} (Z_{di} | \theta_{d}) \right) \]

- But how can we learn topics $\phi$?
  - and/or topic-mixtures $\theta$?
  - and/or topic assignments $Z_{di}$?

- Trick: Inverse conditionals through Bayes’ rule.
- In the following: learning = estimating = inferring
Generative process

- Directed Bayesian models are also called generative models, because they are defined through a stochastic process that generates the data.
  - This process starts with known prior parameters,
  - from which hidden parameters and variables are drawn
  - Finally observed variables, which are tied to given data, are drawn from these parameters conditioned on hidden variables.
- During inference this process is inverted to estimate hidden parameters and variables given data and priors.
- Bayesian model can be read off the generative process.
- Each draw gives rise to one factor in the likelihood function (as in factor graphs).
- E.g. $X \sim \text{Mu}(\theta) \rightarrow P(X | \text{Mu}(\theta))$
Generative process for Latent Dirichlet Allocation

- Given Dirichlet parameters $\alpha$ and $\beta$, number of topics $K$.
- Given $\bar{N}$ (avg corpus size) and $\bar{n}$ (avg document length)

1. **for all** topics $k \in 1, \ldots, K$
2. **draw** $\phi_k \sim \text{Dir}(\beta)$  // distr over words
3. **draw** number of documents to generate $N \sim \text{Pois}(\bar{N})$
4. **for all** documents $d \in \{1 \ldots N\}$  // generating $d \in \mathcal{D}$
5. **draw** topic mix $\theta_d \sim \text{Dir}(\alpha)$  // distr over topics
6. **draw** doc length $n_d \sim \text{Pois}(\bar{n})$
7. **for all** word positions $i \in \{1, \ldots, n_d\}$
8. **draw** topic $Z_{di} \sim \text{Mu}(\theta_d)$ for position $di$
9. **draw** word $W_{di} \sim \text{Mu}(\phi_{Z_{di}})$ for position $di$
10. output words $W_{d,*}$ for document $d$

- When $\mathcal{D}$ is given, r.v.'s $N$ and $n_d$ (gray part) are conditionally independent of topics and can be omitted.
- Left: Topic’s word distribution $\phi_k$.
- Right: Documents’ topic distribution $\theta_d$.
- Colored circles, topic draws $Z_{di} \sim \theta_d$ (can take values $1, \ldots, K$).
- Colored words, drawn from the topic $W_{di} \sim \phi Z_{di}$.
- Document length $n_d$ and corpus size $N$, vocabulary size $V$. 
Bayesian Network

\[ \alpha \rightarrow \tilde{n} \rightarrow \tilde{N} \]

\[ \theta \rightarrow n \rightarrow N \]

\[ Z \rightarrow W \]

\[ i = 1 \ldots n_d \]

\[ d = 1 \ldots N \]

\[ \beta \rightarrow \phi \]

\[ 1 \ldots K \]
Factor graph (directed factor graph notation)

\[
\begin{align*}
\alpha &\quad \bar{n} &\quad \bar{N} \\
\text{Dir} &\quad \text{Pois} &\quad \text{Pois} \\
\theta &\quad n &\quad N \\
\text{Mu} &\quad \text{Mu} &\quad \beta \\
Z &\quad &\quad \\
W &\quad &\quad \\
\end{align*}
\]

\[i = 1 \ldots n_d\]
\[d = 1 \ldots N\]
Likelihood function

- Likelihood function, also called **joint distribution**, for LDA:

\[
\mathcal{L}(Z, \theta, \phi, W, n, N \mid \alpha, \beta, \bar{n}, \bar{N}, K) = \]

\[
\prod_{k=1}^{K} P(\phi_k \mid \text{Dir}(\beta)) \cdot P(N \mid \text{Pois}(\bar{N})) \cdot \prod_{d \in \{1 \ldots N\}} (P(\theta_d \mid \text{Dir}(\alpha)) \cdot P(n_d \mid \text{Pois}(\bar{n})) \cdot \prod_{i=1}^{n_d} (P(Z_{di} \mid \text{Mu}(\theta_d)) \cdot P(W_{di} \mid \text{Mu}(\phi Z_{di})))
\]
Outline

1. Topic identification and applications

2. Modelling assumptions in Latent Dirichlet Allocation

3. Latent variable inference using Gibbs sampling
   3.1 Full Gibbs sampling
   3.2 Collapsed Gibbs Sampling

4. Extensions of topic models

5. Conclusion
Inference

- Given the data \( W, n, N \) and the priors \( \alpha, \beta, \tilde{n}, \tilde{N}, K \), infer estimates for \( Z, \theta, \phi \)
- This is done by applying Bayes’ rule on the generative process.

\[
P(Z, \theta, \phi \mid \ldots) = \frac{\mathcal{L}(Z, \theta, \phi, W, n, N \mid \alpha, \beta, \tilde{n}, \tilde{N}, K)}{\int \int \sum_{Z'} \mathcal{L}(Z', \theta', \phi', W, n, N \mid \alpha, \beta, \tilde{n}, \tilde{N}, K) \, d\theta' \, d\phi'}
\] (8)

- Previously gray factors cancel, obtaining:

\[
\prod_{k=1}^{K} P(\phi_k \mid \text{Dir}(\beta)) \cdot \prod_{d \in \{1 \ldots N\}} \left( P(\theta_d \mid \text{Dir}(\alpha)) \cdot \prod_{i=1}^{n_d} \left( P(Z_{di} \mid \text{Mu}(\theta_d)) \cdot P(W_{di} \mid \text{Mu}(\phi_{Z_{di}})) \right) \right)
\]

\[
\int \int \sum_{Z'} \ldots \, d\theta' \, d\phi'
\] (9)

- Closed-form inference not possible. Let’s try approximate inference!
Outline

1. Topic identification and applications

2. Modelling assumptions in Latent Dirichlet Allocation

3. Latent variable inference using Gibbs sampling
   3.1 Full Gibbs sampling
   3.2 Collapsed Gibbs Sampling

4. Extensions of topic models

5. Conclusion
Gibbs Sampling

- Goal: infer estimates for $Z, \theta, \phi$
- Gibbs sampling is an iterative process.
  - in each epoch all but one variable is held fixed, but the one variable is re-sampled.
    - Case 1: $\forall k$ resample $\phi_k$, fixing all other $\phi_{\neg k}$, as well as $Z, \theta$.
    - Case 2: $\forall d$ resample $\theta_d$, fixing all other $\theta_{\neg d}$, as well as $Z, \phi$.
    - Case 3: $\forall d \forall i$ resample $Z_{di}$, fixing all other $Z_{\neg di}, \theta$ and $\phi$.
- Notation: $\neg di$ stands for “all entries but omitting entry $di$”.
- These sampling equations are called full conditional distributions.
Full conditional for $\theta$

Recap: $P(\mathbf{Z}, \theta, \phi \mid \ldots) =$

$$\prod_{k=1}^{K} P(\phi_k \mid \text{Dir}(\beta)) \cdot \prod_{d \in \{1 \ldots N\}} \left( P(\theta_d \mid \text{Dir}(\alpha)) \cdot \prod_{i=1}^{n_d} \left( P(Z_{di} \mid \text{Mu}(\theta_d)) \cdot P(W_{di} \mid \text{Mu}(\phi_{Z_{di}})) \right) \right)$$

$$\int \int \sum_{\mathbf{Z}'} \ldots \ d\theta' d\phi'$$

$P(\theta_d \mid \theta_{\neg d}, \ldots) =$

$$\frac{P(\theta_d \mid \text{Dir}(\alpha)) \cdot \prod_{i=1}^{n_d} P(Z_{di} \mid \text{Mu}(\theta_d)))}{\int P(\theta'_d \mid \text{Dir}(\alpha)) \cdot \prod_{i=1}^{n_d} P(Z_{di} \mid \text{Mu}(\theta'_d))) d\theta'_d}$$

- This is known as the **Dirichlet-Multinomial compound** distribution, which is a generalization of the Beta-Bernoulli distribution to more than two dimensions.
Recap Lecture 04, Slide 17 (MAP estimation)

- Recall: posterior $\propto$ likelihood $\times$ prior
- Can show: posterior is $\text{Dir}(N_1 + \alpha_1, \ldots, N_k + \alpha_K)$
  - Thus the $\alpha_k$ can be interpreted as pseudo-counts (as in the beta-binomial model)
- The maximum a posteriori estimate is the categorical distribution ($\theta$) that maximizes the posterior
- We have

$$
\hat{\theta}_{\text{MAP}} = \arg\max_{\theta \in S_K} \text{Mu}(n | n, \theta) \text{ Dir}(\theta | \alpha)
$$

$$
= \arg\max_{\theta \in S_K} \text{Dir}(N_1 + \alpha_1, \ldots, \alpha_K + N_k)
$$

$$
= \arg\max_{\theta \in S_K} \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}
$$

$$
= \left( \frac{n_1 + \alpha_1 - 1 \ n_2 + \alpha_2 - 1 \ \cdots \ n_K + \alpha_K - 1}{n + \alpha_0 - K} \right)^T
$$

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$
Full conditional for $\theta$ (2)

Recap: $P(\mathbf{Z}, \theta, \phi \mid \ldots) = \\
\prod_{k=1}^{K} P(\phi_k \mid \text{Dir}(\beta)) \cdot \prod_{d \in \{1\ldots N\}} \left( P(\theta_d \mid \text{Dir}(\alpha)) \cdot \prod_{i=1}^{n_d} (P(Z_{di} \mid \text{Mu}(\theta_d)) \cdot P(W_{di} \mid \text{Mu}(\phi_{Z_{di}}))) \right) \\
\int \int \sum_{\mathbf{Z}'} \ldots d\theta' d\phi' \\

P(\theta_d \mid \theta_{-d}, \ldots) = \\
\frac{P(\theta_d \mid \text{Dir}(\alpha)) \cdot \prod_{i=1}^{n_d} P(Z_{di} \mid \text{Mu}(\theta_d)))}{\int P(\theta'_d \mid \text{Dir}(\alpha)) \cdot \prod_{i=1}^{n_d} P(Z_{di} \mid \text{Mu}(\theta'_d))) d\theta'_d} \\
(10) \\

\arg \max_{\theta_d} P(\theta_d \mid \ldots) = \\
\frac{(n_1 + \alpha_1 - 1 \cdots n_K + \alpha_K - 1)^T}{n + \alpha_0 - K} \\
(11) \\

for a given $d$: $n_k = \#\{i \mid Z_{di} = k, \forall 1 \leq i \leq n_d\}$ \\
(12)
Full conditional for $\phi$

Recap:  
\[ P(Z, \theta, \phi | \ldots) = \prod_{k=1}^{K} P(\phi_k | \text{Dir}(\beta)) \cdot \prod_{d \in \{1 \ldots N\}} \left( P(\theta_d | \text{Dir}(\alpha)) \cdot \prod_{i=1}^{n_d} \left( P(Z_{di} | \text{Mu}(\theta_d)) \cdot P(W_{di} | \text{Mu}(\phi_{Z_{di}})) \right) \right) \int \int \sum_{z'} \ldots d\theta d\phi' \]

- Exercise!

\[ P(\phi_k | \phi_{\neg k}, \ldots) = \ldots? \quad (13) \]

- Hint: re-arrange the product over $i$

- Solution:

\[ \arg \max_{\phi_k} P(\phi_k | \ldots) = \left( \frac{n_1 + \beta_1 - 1}{n + \beta_0 - V} \cdots \frac{n_V + \beta_V - 1}{n + \beta_0 - V} \right)^T \quad (14) \]

\[ n_v = \# \{ i, d \mid Z_{di} = k, W_{di} = v, \forall d, \forall 1 \leq i \leq n_d \} \quad (15) \]
Full conditional for $Z$

Recap: $P(Z, \theta, \phi \mid \ldots) =$

\[
\prod_{k=1}^{K} P(\phi_k \mid \text{Dir}(\beta)) \cdot \prod_{d \in \{1 \ldots N\}} \left( P(\theta_d \mid \text{Dir}(\alpha)) \cdot \prod_{i=1}^{n_d} \left[ P(Z_{di} \mid \text{Mu}(\theta_d)) \cdot P(W_{di} \mid \text{Mu}(\phi_{Z_{di}})) \right] \right) \\
\int \int \sum_{Z'} \ldots \ d\theta' d\phi'
\]

\[
P(Z_{di} \mid Z_{\neg di}, \ldots) = \frac{P(Z_{di} \mid \text{Mu}(\theta_d)) \cdot P(W_{di} \mid \text{Mu}(\phi_{Z_{di}}))}{\sum_{z'_{di}=1}^{K} P(z'_{di} \mid \text{Mu}(\theta_d)) \cdot P(W_{di} \mid \text{Mu}(\phi_{z'}))}
\]

(16)

\[
P(Z_{di} = k \mid Z_{\neg di}, \ldots) \propto \theta_d(k) \phi_k(w_{di})
\]

(17)

- Compute for all $k$, then recompute the normalizer to obtain a multinominal distribution $\mu$ of range $K$.
- Sample a new setting for $Z_{di} \sim \text{Mu}(\mu)$. 
Outline

1. Topic identification and applications

2. Modelling assumptions in Latent Dirichlet Allocation

3. Latent variable inference using Gibbs sampling
   3.1 Full Gibbs sampling
   3.2 Collapsed Gibbs Sampling

4. Extensions of topic models

5. Conclusion
Integrating out multinomials

- If settings for $Z_{di}$ are known the likelihood function can be arranged into individual products over all $\phi_k$ and $\theta_d$.

\[
P(Z | \ldots) = \prod_{k=1}^{K} \int P(\phi_k | \beta) \prod_{??} P(W_{di} | \phi_k) d\phi_k \prod_{d=1}^{N} \int P(\theta_d | \alpha) \prod_{??} P(Z_{di} | \theta_d) d\theta_d
\]

Closed form solution for these integrals ($n_0 = \sum n_k, \alpha_0 = \sum \alpha_k$)

\[
\int P(\theta_d | \text{Dir}(\alpha)) \cdot \prod_{i=1}^{n_d} P(Z_{di} | \text{Mu}(\theta_d))) d\theta_d
\]

\[
= \frac{\Gamma(\alpha_0)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(n_k + \alpha_k)}{\Gamma(n_0 + \alpha_0)}
\]
Analytic solution for one integral

- Resampling $P(Z_{di}|Z_{\neg di}, \ldots)$ on this collapsed form amounts to adjusting counts $n_k$.
- In particular the old setting of $Z_{di}$ is removed from the count statistics.
- In order to re-sample, it is test-wise added to each topic $k$ to compute factors that look like

$$P(Z_{di} = k \mid \ldots) = \frac{\theta \text{ with } n_k + 1}{\theta \text{ with } n_k} \ldots$$ (23)

$$(k' \in \neg k) = \frac{\Gamma(\alpha_0) \prod_{k'} \Gamma(n_{k'} + \alpha_{k'}) \cdot \Gamma(n_k + \alpha_k + 1)}{\Gamma(\alpha_0) \prod_{k'} \Gamma(n_{k'} + \alpha_{k'}) \cdot \Gamma(n_0 + \alpha_0 + 1)} \ldots$$ (24)

$$= \frac{\Gamma(n_k + \alpha_k + 1)}{\Gamma(n_k + \alpha_k)} \frac{1}{\Gamma(n_0 + \alpha_0 + 1)} \ldots$$ (25)
Gamma trick

- The remaining trick lies in the definition of the Gamma function

\[
\Gamma(x + 1) = x\Gamma(x) \iff \frac{\Gamma(x + 1)}{\Gamma(x)} = x
\]

\[
P(Z_{di} = k \mid \ldots) = \frac{\theta \text{ with } n_k + 1}{\theta \text{ with } n_k} \cdot \ldots
\]

\[
= \frac{\Gamma(n_k + \alpha_k + 1)}{\Gamma(n_k + \alpha_k)} \cdot \frac{1}{\Gamma(n_0 + \alpha_0 + 1)} \cdot \ldots
\]

\[
= (n_k + \alpha_k) \frac{1}{n_0 + \alpha_0} \cdot \ldots
\]

- This represents some quantity that is related to \(\theta_d\). We refer to it as \(\tilde{\theta}_d(k)\). Equivalent expressions for \(\phi\) are derived analogously.
Collapsed Gibbs sampler

- As these integrals can be solved in closed form, where the counts follow from settings of \( Z \)
  - Topic \( k \), document \( d \): \( \tilde{\theta}_d(k) \propto \#\{k \mid Z_{di} = k\} + \alpha_k \)
  - Topic \( k \), word \( v \): \( \tilde{\phi}_k(v) \propto \#\{k, v \mid Z_{di} = k, W_{di} = v\} + \beta_v \)

- As a conclusion, it is sufficient to only re-sample \( Z_{di} \) given remaining \( Z_{\neg di} \) and \( W_{di} \) through Equation 17:
  \[
P(Z_{di} = k \mid Z_{\neg di}, \ldots) \propto \tilde{\theta}_d(k)\tilde{\phi}_k(w_{di})
  \]

- After the Gibbs sampler is finished final estimates for \( \hat{Z} \) are available.

- Estimates for parameters \( \hat{\phi} \) and \( \hat{\theta} \) can be obtained using the expressions of full conditionals from the full Gibbs sampler.
Collapsed blocked Gibbs sampler and Pros and Cons

- **Collapsed sampler**
  Analytically integrate over parameters \((\phi, \theta)\) instead of sampling.
  - Pro: Reported to have faster sampling convergence (i.e., better mixing)
  - Con: But full sampler is easy to parallelize (YahooLDA)

- **Blocked sampler**
  Sample groups of hidden variables as a block.
  - Problem: Small groups of latent variables with strong dependencies may prevent good mixing.
  - Intuition: “variables hold hands, either all change or none.”
  - Solution: Resample all variables in one block together through their joint distribution.
  - Con: Such groups are not always known beforehand.
Outline

1. Topic identification and applications

2. Modelling assumptions in Latent Dirichlet Allocation

3. Latent variable inference using Gibbs sampling

4. Extensions of topic models

5. Conclusion
Customizing topic models

- Main advantage of topic models is their extensibility.
- Different generative models can be composed to a larger model.
- New variables can be added to account for additional data
  - authors of a document
  - links between documents
  - date of a document
- Distribution assumptions can be varied (e.g., multinominal versus multi-variate Gaussian)
- Different configurations of priors, e.g., dense versus sparse.
Examples of extended topic models

- Geographics topics: Close Proximity leads to more similar topics.
- Labeled LDA, Disc LDA, Supervised LDA: Inform topics through document classes.
- Ideal point topic model: Each topic has two extreme perspectives, documents are in between.
- Collaborative filtering pLSI: Learn topics to predict ratings.
- Topics over time, dynamic topic model: Each document has a time stamp. Topics change words gradually over time.
- Author-topic model: Each document has multiple known authors. Each author has one topic mixture.
- Network topic models: Document network informs topics, makes predictions about links and influence strengths.
- Dirichlet-multinomial Regression: Inform topics from meta-data and content.
Network topic models

- See slide deck B!
Outline

1. Topic identification and applications
2. Modelling assumptions in Latent Dirichlet Allocation
3. Latent variable inference using Gibbs sampling
4. Extensions of topic models
5. Conclusion
**Lessons learnt**

- Topic models are one approach towards detection of topics in text collections.
- More general: Identify re-occurring patterns in data collections.
- Can incorporate additional data into the inference process.
- Unsupervised machine learning with Bayesian networks.

- Topic model inference with Gibbs sampling in a nutshell:
  - Devise a generative process. Priors to data.
  - Derive Bayesian network and/or factor graph.
  - Derive joint distribution. Optionally integrate out parameters (collapsing).
  - Derive full conditionals for all latent variables.
  - Random initialization, then iteratively re-sample all variables.
Topic model toolkits

● Particular topic models
  ▶ Stanford topic model toolbox
    http://nlp.stanford.edu/software/tmt
  ▶ Topic modeling at Princeton
  ▶ MALLET (Java) http://mallet.cs.umass.edu
  ▶ Network topic models: Bayes-stack
    https://github.com/bgamari/bayes-stack
  ▶ Gensim (Python) http://radimrehurek.com/gensim/

● Frameworks for generative models
  ▶ Variational inference: Infer.net
    http://research.microsoft.com/infernet/
  ▶ Gibbs sampling: OpenBUGS http://openbugs.net/
Suggested reading

- Introductions to topic models

- The maths behind topic model inference

- Topic model bibliography
  http://mimno.infosci.cornell.edu/topics.html