Topic Modeling with Latent Dirichlet Allocation

Vineet Mehta

University of Massachusetts - Lowell

Contents

Contents

Contents

Contents

Background

- A number of techniques for text analysis and information retrieval have been developed over the past decades
- This presentation focuses on one such technique known as *Latent Dirichlet Allocation*
- *Latent Dirichlet Allocation* (LDA) was introduced by David Blei, Andrew Ng and Michael Jordan in a 2003 paper in Journal of Machine Learning Research
- Since its introduction LDA has been employed for applications beyond text analysis
- LDA has also seen a number of extensions

Topic Modeling

- LDA aims at classifying large collections of documents through statistical relationships amongst words, known as *topics*
- *Topics* are distributions over a vocabulary of words
- LDA employs Bayesian inference techniques to estimate the statistical quantities that are *topics*
- The LDA approach to text analysis does not assume any knowledge of language structure
- Documents in a text collection are treated as *bag-of-words*

Applications

A Small Sampling

- Exploring scientific, political, wikipedia articles
- Audio information retrieval using acoustic features
- Image segmentation using visual features
- Identifying surprising events in video data
- Analysis of stock categories using financial topic models
- Development of user recommendation systems in social media
- Topic models for gene expression analysis
- Analysis of twitter data for public health status and trends

Finding Out More

A really small sampling...so just google it!

- Workshops
  - Topic Models: Computation, Application, and Evaluation (NIPS 2013)
  - Applications for Topic Models: Text and Beyond (NIPS 2009)
  - Workshop on Topic Models: Structure, Applications, Evaluation, and Extensions (ICML 2010)
  - Topic Modeling for Humanities Research (MITH 2012)
- Topic Modeling Software
  - *lda-c*: C-code by David Blei
  - *lda*: R-language package at CRAN
  - *gensim*: Python package includes Latent Dirichlet Allocation
  - *mallet*: Java machine learning package, including topic modeling
  - *topictoolbox*: Matlab toolbox by UCI
- Data
  - UCI machine learning repository: `http://archive.ics.uci.edu/ml/datasets.html`
  - infochimps: `http://www.infochimps.com`
  - Enron dataset: `https://www.cs.cmu.edu/~enron/`
  - LDC (not free): `http://catalog.ldc.upenn.edu`

Bayes' Theorem

Consider the dataset $\mathcal{X}_N^{(K)}$ which consists of N samples

$$\mathcal{X}_N^{(K)} = \{\mathbf{x}_1 \cdots \mathbf{x}_N\}$$

where $\mathbf{x}_i = [x_{1,i} \cdots x_{K,i}]^T$ is sample from the random vector $\mathcal{X}_i^{(K)}$

The random vector $\mathcal{X}_i^{(K)}$ has a distribution $p(\mathcal{X}_i^{(K)}|\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$. And $\{\mathcal{X}_i^{(K)}\}$ are independent and identically distributed random vectors.

Bayes' Theorem

$$\text{posterior} \equiv p(\boldsymbol{\theta}|\mathcal{X}^{(K)}) = \frac{p(\mathcal{X}^{(K)}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{X}^{(K)})} \equiv \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

Key Distributions

Univariate Case

Binomial Distribution

$$\mathrm{Bin}(x|\theta, N) \triangleq p(\mathcal{X} = x|\theta, N) = \binom{N}{x}\theta^x(1-\theta)^{1-x} \qquad x \in \mathbb{N}_1$$

Bernoulli Distribution

$$\mathrm{Bern}(x|\theta) \triangleq p(\mathcal{X} = x|\theta) = \theta^x(1-\theta)^{1-x} \qquad x \in \{0, 1\}$$

Likelihood of N Bernoulli observations

$$p(\mathcal{X}_N^{(1)}|\theta) = \prod_{i=1}^{N} \theta^{\mathbb{I}(x_i=1)}(1-\theta)^{\mathbb{I}(x_i=0)} = \theta^{n_1}(1-\theta)^{n_0} \quad N = n_0 + n_1$$

Key Distributions

Multivariate Case

Multinomial Distribution

$$\mathrm{Mult}(\mathbf{x}|\boldsymbol{\theta}, N) \triangleq p(\mathcal{X}^{(K)} = \mathbf{x}|\boldsymbol{\theta}, N) = \binom{N}{\mathbf{x}} \prod_{k=1}^{K} \theta_k^{x_k} \quad \mathbf{x} \in \mathbb{N}_1^K, \; \sum_{k=1}^{K} \theta_k = 1$$

Categorical Distribution

$$\mathrm{Cat}(\mathbf{x}|\boldsymbol{\theta}) \triangleq p(\mathcal{X}^{(K)} = \mathbf{x}|\boldsymbol{\theta}, 1) = \prod_{k=1}^{K} \theta_k^{x_k} = \theta_k \mathbb{I}(x_k = 1) \quad \mathbf{x} \in \{0, 1\}^K$$

Key Distributions

Multivariate Case (continued)

Likelihood of N Bernoulli observations

$$p(\mathcal{X}_N^{(K)}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{k=1}^{K} \theta_k^{\mathbb{I}(x_{i,k}=1)} = \prod_{k=1}^{K} \theta_k^{n_k}$$

where

$$\sum_{k=1}^{K} n_k = N$$

Parameterized Priors

Hyperparameters

Generalization: Parameter $\boldsymbol{\theta}$ depends on the hyperparameter $\boldsymbol{\vartheta}$

$$p(\boldsymbol{\theta}) \quad \Rightarrow \quad p(\boldsymbol{\theta}|\boldsymbol{\vartheta})$$

Bayes' Theorem

$$p(\boldsymbol{\theta}|\mathcal{X}_N^{(K)}, \boldsymbol{\vartheta}) = \frac{p(\mathcal{X}_N^{(K)}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\vartheta})}{\int p(\mathcal{X}_N^{(K)}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\vartheta})d\boldsymbol{\theta}}$$

Conjugate Priors

Univariate Case: $p(\theta|\boldsymbol{\vartheta})$

Beta Distribution: $\boldsymbol{\vartheta} = (\alpha, \beta)$

$$\mathrm{Beta}(\theta|\alpha,\beta) \triangleq p(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Posterior Distribution: $p(\theta|\mathcal{X}_N^{(1)}, \boldsymbol{\vartheta})$

$$\begin{aligned} p(\theta|\mathcal{X}_N^{(1)}, \boldsymbol{\vartheta}) &= \frac{\frac{1}{B(\alpha,\beta)}\theta^{n_1+\alpha-1}(1-\theta)^{n_0+\beta-1}}{\frac{1}{B(\alpha,\beta)}\int \theta^{n_1+\alpha-1}(1-\theta)^{n_0+\beta-1}d\theta} \\ &= \frac{1}{B(n_1+\alpha, n_0+\beta)}\theta^{n_1+\alpha-1}(1-\theta)^{n_0+\beta-1} \\ &\triangleq \mathrm{Beta}(\theta|n_1+\alpha, n_0+\beta) \end{aligned}$$

Prediction

Univariate Case

Marginalizing out likelihood parameters

$$p(\mathcal{X}_N^{(1)}|\alpha,\beta) = \int p(\mathcal{X}_N^{(1)}|\theta)p(\theta|\alpha,\beta)d\theta$$

$$= \frac{1}{B(\alpha,\beta)}\int \theta^{n_1+\alpha-1}(1-\theta)^{n_0+\beta-1}d\theta = \frac{B(n_1+\alpha, n_0+\beta)}{B(\alpha,\beta)}$$

$$= \frac{\Gamma(n_1+\alpha)\Gamma(n_0+\beta)\Gamma(\alpha+\beta)}{\Gamma(n_1+n_0+\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)}$$

New sample likelihood

$$p(\tilde{x}=1|\mathcal{X}_N^{(1)},\alpha,\beta) = \frac{p(\tilde{x}=1,\mathcal{X}_N^{(1)}|\alpha,\beta)}{p(\mathcal{X}_N^{(1)}|\alpha,\beta)} = \frac{n_1+\alpha}{n_1+n_0+\alpha+\beta}$$

Conjugate Priors

Multivariate Case: $p(\boldsymbol{\theta}|\boldsymbol{\vartheta})$

Dirichlet Distribution

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\vartheta}) \triangleq p(\boldsymbol{\theta}|\boldsymbol{\vartheta}) = \frac{1}{\Delta(\boldsymbol{\vartheta})}\prod_{k=1}^{K}\theta_k^{\vartheta_k - 1} \quad \Delta(\boldsymbol{\vartheta}) = \frac{\prod\limits_{k=1}^{K}\Gamma(\vartheta_k)}{\Gamma(\sum\limits_{k=1}^{K}\vartheta_k)}, \quad \sum_{k=1}^{K}\theta_k = 1$$

Posterior Distribution: $p(\boldsymbol{\theta}|\mathcal{X}_N^{(K)}, \boldsymbol{\vartheta})$

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{X}_N^{(K)}, \boldsymbol{\vartheta}) &= \frac{\frac{1}{\Delta(\boldsymbol{\vartheta})}\prod_{k=1}^{K}\theta_k^{n_k+\vartheta_k-1}}{\frac{1}{\Delta(\boldsymbol{\vartheta})}\int\prod_{k=1}^{K}\theta_k^{n_k+\vartheta_k-1}d\boldsymbol{\theta}} \\ &= \frac{1}{\Delta(\mathbf{n}+\boldsymbol{\vartheta})}\prod_{k=1}^{K}\theta_k^{n_k+\vartheta_k-1} \triangleq \mathrm{Dir}(\boldsymbol{\theta}|\mathbf{n}+\boldsymbol{\vartheta}) \end{aligned}$$

Modeling Text

Notation

$$\mathbf{x}_{i,m} \equiv i\text{-th word in } m\text{-th document}$$

$$\mathbf{z}_{i,m} \equiv \text{topic from which } i\text{-th workd in } m\text{-th document is drawn}$$

$$\boldsymbol{\omega}_j \equiv \text{value taken by } \mathbf{x}_{i,m}\text{, where } j \in [1, V] \text{ and } V \text{ is vocabulary size}$$

$$\boldsymbol{\xi}_k \equiv \text{value taken by } \mathbf{z}_{i,m}\text{, where } k \in [1, K] \text{ and } K \text{ is topic count}$$

$$\boldsymbol{\theta}_m \equiv \text{topic distribution for } m\text{-th document}$$

$$\boldsymbol{\phi}_k \equiv \text{word distribution for } k\text{-th topic}$$

$$\boldsymbol{\alpha} \equiv \text{hyperparameters for document topic distribution}$$

$$\boldsymbol{\beta} \equiv \text{hyperparameters for topic word distribution}$$

$$\mathrm{X}_{N_m}^{(V)} \equiv \text{words in document } m$$

$$\mathcal{X}_{\mathcal{N}}^{(V)} \equiv \text{words in all documents (corpus)}$$

$$\mathrm{Z}_{N_m}^{(K)} \equiv \text{topics associated with words in document } m$$

$$\mathcal{Z}_{\mathcal{N}}^{(K)} \equiv \text{topics associated with words in corpus}$$

Modeling Text

Latent Dirichlet Allocation

Generative Model

for all topics $k \in [1, K]$ **do**

  $\phi_k \sim \text{Dir}(\phi_k|\beta)$

for all documents $m \in [1, M]$ **do**

  $\boldsymbol{\theta}_m \sim \text{Dir}(\boldsymbol{\theta}_m|\boldsymbol{\alpha})$

  $N_m \sim \text{Pois}(N_m|\xi)$

  for all words $i \in [1, N_m]$ in document m **do**

    topic index $\mathbf{z}_{i,m} \sim \text{Mult}(\mathbf{z}_{i,m}|\boldsymbol{\theta}_m, 1)$

    word $\mathbf{x}_{i,m} \sim \text{Mult}(\mathbf{x}_{i,m}|\phi_{\{k:\mathbb{I}(\mathbf{z}_{i,m}=\boldsymbol{\xi}_k)\}}, 1)$

Latent Dirichlet Allocation

Joint Distribution of Known and Hidden Variables

i-th word in *m*-th document

$$p(\mathbf{x}_{i,m}, \mathbf{z}_{i,m}, \boldsymbol{\theta}_m, \mathbf{\Phi} | \boldsymbol{\alpha}, \boldsymbol{\beta})$$

All words in *m*-th document

$$p(\mathrm{X}_{N_m}^{(V)}, \mathrm{Z}_{N_m}^{(K)}, \boldsymbol{\theta}_m, \mathbf{\Phi} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{N_m} p(\mathbf{x}_{i,m}, \mathbf{z}_{i,m}, \boldsymbol{\theta}_m, \mathbf{\Phi} | \boldsymbol{\alpha}, \boldsymbol{\beta})$$

All words in corpus

$$p(\mathcal{X}_{\mathcal{N}}^{(V)}, \mathcal{Z}_{\mathcal{N}}^{(K)}, \mathbf{\Theta}, \mathbf{\Phi} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{m=1}^{M} p(\mathrm{X}_{N_m}^{(V)}, \mathrm{Z}_{N_m}^{(K)}, \boldsymbol{\theta}_m, \mathbf{\Phi} | \boldsymbol{\alpha}, \boldsymbol{\beta})$$

$$= \prod_{m=1}^{M} \prod_{i=1}^{N_m} p(\mathbf{x}_{i,m} | \mathbf{z}_{i,m}, \mathbf{\Phi}) p(\mathbf{z}_{i,m} | \boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m | \boldsymbol{\alpha}) p(\mathbf{\Phi} | \boldsymbol{\beta})$$

Latent Dirichlet Allocation

Conditional Distributions - Word Likelihoods

Word in Document

$$p(\mathbf{x}_{i,m}|\mathbf{z}_{i,m}, \mathbf{\Phi}) = \prod_{j=1}^{V}\prod_{k=1}^{K} \phi_{k,j}^{\mathbb{I}(\mathbf{x}_{i,m}=\boldsymbol{\omega}_j \ \wedge \ \mathbf{z}_{i,m}=\boldsymbol{\xi}_k)}$$

All Words in Document

$$\prod_{i=1}^{N_m} p(\mathbf{x}_{i,m}|\mathbf{z}_{i,m}, \mathbf{\Phi}) = \prod_{i=1}^{N_m}\prod_{j=1}^{V}\prod_{k=1}^{K} \phi_{k,j}^{\mathbb{I}(\mathbf{x}_{i,m}=\boldsymbol{\omega}_j \ \wedge \ \mathbf{z}_{i,m}=\boldsymbol{\xi}_k)}$$

$$= \prod_{j=1}^{V}\prod_{k=1}^{K} \phi_{k,j}^{\rho_{k,j}}$$

$\rho_{k,j}$ is the count of word j assigned to topic k

Latent Dirichlet Allocation

Conditional Distributions - Topic Likelihoods

Topic Likelihood for Single Word

$$p(\mathbf{z}_{i,m}|\boldsymbol{\theta}_m) = \prod_{k=1}^{K} \theta_{m,k}^{\mathbb{I}(\mathbf{z}_{i,m}=\boldsymbol{\xi}_k)}$$

Topic Likelihood for all Words in Document

$$\prod_{i=1}^{N_m} p(\mathbf{z}_{i,m}|\boldsymbol{\theta}_m) = \prod_{i=1}^{N_m} \prod_{k=1}^{K} \theta_{m,k}^{\mathbb{I}(\mathbf{z}_{i,m}=\boldsymbol{\xi}_k)}$$

$$= \prod_{k=1}^{K} \theta_{m,k}^{\upsilon_{m,k}}$$

$\upsilon_{m,k}$ is the count of words in document m assigned to topic k

Latent Dirichlet Allocation

Priors

Topic Distribution for Document

$$p(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) = \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k - 1}$$

Word Distribution over Topics

$$p(\boldsymbol{\Phi}|\boldsymbol{\beta}) = \frac{1}{\Delta(\boldsymbol{\beta})} \prod_{k=1}^{K} \prod_{j=1}^{V} \phi_{k,j}^{\beta_j - 1}$$

Latent Dirichlet Allocation

Full Joint Distribution

$$
\begin{aligned}
& p(\mathcal{X}_{\mathcal{N}}^{(V)}, \mathcal{Z}_{\mathcal{N}}^{(K)}, \mathbf{\Theta}, \mathbf{\Phi} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \prod_{m=1}^{M} \prod_{i=1}^{N_m} p(\mathbf{x}_{i,m} | \mathbf{z}_{i,m}, \mathbf{\Phi}) p(\mathbf{z}_{i,m} | \boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m | \boldsymbol{\alpha}) p(\mathbf{\Phi} | \boldsymbol{\beta}) \\
&= \frac{1}{\Delta(\boldsymbol{\alpha}) \Delta(\boldsymbol{\beta})} \prod_{m=1}^{M} \prod_{j=1}^{V} \prod_{k=1}^{K} \phi_{k,j}^{\rho_{k,j} + \beta_j - 1} \theta_{m,k}^{\upsilon_{m,k} + \alpha_k - 1}
\end{aligned}
$$

Latent Dirichlet Allocation

Integrating out $\mathbf{\Theta}$ and $\mathbf{\Phi}$

$$
\begin{aligned}
&p(\mathcal{X}_{\mathcal{N}}^{(V)}, \mathcal{Z}_{\mathcal{N}}^{(K)} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \frac{1}{\Delta(\boldsymbol{\alpha})\Delta(\boldsymbol{\beta})} \int \prod_{m=1}^{M} \prod_{j=1}^{V} \prod_{k=1}^{K} \phi_{k,j}^{\rho_{k,j}+\beta_j-1} \theta_{m,k}^{\upsilon_{m,k}+\alpha_k-1} d\mathbf{\Theta} d\mathbf{\Phi} \\
&= \prod_{m=1}^{M} \frac{\Delta(\boldsymbol{\upsilon}_m + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \prod_{k=1}^{K} \frac{\Delta(\boldsymbol{\rho}_k + \boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}
\end{aligned}
$$

Gibbs Sampling

Sampling the Posterior $p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$

Sampling Algorithm

initialize $\mathcal{Z}$ to $\mathcal{Z}^{(0)} = \{\mathbf{z}_1^{(0)} \ldots \mathbf{z}_N^{(0)}\}$ at interation $l = 0$

for $l \in [0, L]$ **do**

  for $n \in [1, N]$ **do**

    sample $\mathbf{z}_n^{(l+1)} \sim p(\mathbf{z}_n^{(l+1)} | \{\mathbf{z}_1^{(l+1)} \ldots \mathbf{z}_{n-1}^{(l+1)}, \mathbf{z}_{n+1}^{(l)} \ldots \mathbf{z}_N^{(l)}\}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$

After sufficient iterations the sampler converges, and the samples $\mathbf{z}_n^{(l)}$ are instances of $p(\mathcal{Z}|\mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$

Constructing the Posterior for Gibbs Sampler

$$p(\mathbf{z}_n|\mathcal{Z}_{\neg n}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\mathcal{X}, \mathcal{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathcal{X}_{\neg n}, \mathcal{Z}_{\neg n}|\boldsymbol{\alpha}, \boldsymbol{\beta})p(\mathbf{x}_n|\boldsymbol{\alpha}, \boldsymbol{\beta})} \propto \frac{p(\mathcal{X}, \mathcal{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathcal{X}_{\neg n}, \mathcal{Z}_{\neg n}|\boldsymbol{\alpha}, \boldsymbol{\beta})}$$

$$p(\mathcal{X}_{\neg n}, \mathcal{Z}_{\neg n}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\mathcal{X}_{\neg n}|\mathcal{Z}_{\neg n}, \boldsymbol{\Phi})p(\mathcal{Z}_{\neg n}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\boldsymbol{\alpha})p(\boldsymbol{\Phi}|\boldsymbol{\beta})\ \ d\boldsymbol{\Theta}d\boldsymbol{\Phi}$$

$$p(\mathcal{X}_{\neg n}|\mathcal{Z}_{\neg n}, \boldsymbol{\Phi}) = \prod_{\substack{m=1 \\ n=(q,r,s,t) \\ \neg n \to (m,i,j,k) \neq (q,r,s,t)}}^{M} \prod_{i=1}^{N_m} \prod_{j=1}^{V} \prod_{k=1}^{K} \phi_{k,j}^{\mathbb{I}(\mathbf{x}_{i,m}=\boldsymbol{\omega}_j\ \wedge\ \mathbf{z}_{i,m}=\boldsymbol{\xi}_k)} = \prod_{j=1}^{V} \prod_{k=1}^{K} \phi_{k,j}^{\rho_{k,j}^{(\neg n)}}$$

$$p(\mathcal{Z}_{\neg n}|\boldsymbol{\Theta}) = \prod_{\substack{m=1 \\ n=(q,r,s,t) \\ \neg n \to (m,i,k) \neq (q,r,t)}}^{M} \prod_{i=1}^{N_m} \prod_{k=1}^{K} \theta_{m,k}^{\mathbb{I}(\mathbf{z}_{i,m}=\boldsymbol{\xi}_k)} = \prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{m,k}^{\upsilon_{m,k}^{(\neg n)}}$$

Defining Counts for Posterior in Gibbs Sampler

Counts words assigned to topics k: $\boldsymbol{\rho}_k$, $\boldsymbol{\rho}_k^{(\neg n)}$

$$\rho_{k,j}^{(\neg n)} = \begin{cases} \rho_{k,j} & (j,k) \neq (s,t) \\ \rho_{k,j} - 1 & (j,k) = (s,t) \end{cases}$$

Counts words in document m assigned to topics: $\boldsymbol{\upsilon}_m$, $\boldsymbol{\upsilon}_m^{(\neg n)}$

$$\upsilon_{m,k}^{(\neg n)} = \begin{cases} \upsilon_{m,k} & (m,k) \neq (q,t) \\ \upsilon_{m,k} - 1 & (m,k) = (q,t) \end{cases}$$

Joint Distributions and Posterior

$$p(\mathcal{X}, \mathcal{Z} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{m=1}^{M} \frac{\Delta(\boldsymbol{\upsilon}_m + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \prod_{k=1}^{K} \frac{\Delta(\boldsymbol{\rho}_k + \boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}$$

$$p(\mathcal{X}_{\neg n}, \mathcal{Z}_{\neg n} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{m=1}^{M} \frac{\Delta(\boldsymbol{\upsilon}_m^{(\neg n)} + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \prod_{k=1}^{K} \frac{\Delta(\boldsymbol{\rho}_k^{(\neg n)} + \boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}$$

$$p(\mathbf{z}_n | \mathcal{Z}_{\neg n}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{p(\mathcal{X}, \mathcal{Z} | \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathcal{X}_{\neg n}, \mathcal{Z}_{\neg n} | \boldsymbol{\alpha}, \boldsymbol{\beta})} = \frac{\Delta(\boldsymbol{\upsilon}_q + \boldsymbol{\alpha})\Delta(\boldsymbol{\rho}_t + \boldsymbol{\beta})}{\Delta(\boldsymbol{\upsilon}_q^{(\neg n)} + \boldsymbol{\alpha})\Delta(\boldsymbol{\rho}_t^{(\neg n)} + \boldsymbol{\beta})}$$

$$\Delta(\mathbf{y}) = \frac{\prod_{k=1}^{K} \Gamma(y_k)}{\Gamma(\sum_{k=1}^{K} y_k)} \qquad \Gamma(y+1) = y\Gamma(y)$$

Simplifing Expression for Posterior

$$p(\mathbf{z}_n|\mathcal{Z}_{\neg n}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{\frac{\prod_{k=1}^{K}\Gamma(\upsilon_{q,k}+\alpha_k)}{\Gamma(\sum_{k=1}^{K}\upsilon_{q,k}+\alpha_k)} \cdot \frac{\prod_{j=1}^{V}\Gamma(\rho_{t,j}+\beta_j)}{\Gamma(\sum_{j=1}^{V}\rho_{t,j}+\beta_j)}}{\frac{\prod_{k=1}^{K}\Gamma(\upsilon_{q,k}^{(\neg n)}+\alpha_k)}{\Gamma(\sum_{k=1}^{K}\upsilon_{q,k}^{(\neg n)}+\alpha_k)} \cdot \frac{\prod_{j=1}^{V}\Gamma(\rho_{t,j}^{(\neg n)}+\beta_j)}{\Gamma(\sum_{j=1}^{V}\rho_{t,j}^{(\neg n)}+\beta_j)}}$$

$$\propto \frac{\frac{\Gamma(\upsilon_{q,t}+\alpha_t)}{\Gamma(\sum_{k=1}^{K}\upsilon_{q,k}+\alpha_k)} \cdot \frac{\Gamma(\rho_{t,s}+\beta_s)}{\Gamma(\sum_{j=1}^{V}\rho_{t,j}+\beta_j)}}{\frac{\Gamma(\upsilon_{q,t}^{(\neg n)}+\alpha_t)}{\Gamma(\sum_{k=1}^{K}\upsilon_{q,k}^{(\neg n)}+\alpha_k)} \cdot \frac{\Gamma(\rho_{t,s}^{(\neg n)}+\beta_s)}{\Gamma(\sum_{j=1}^{V}\rho_{t,j}^{(\neg n)}+\beta_j)}}$$

$$\propto \frac{\frac{\Gamma(\upsilon_{q,t}+\alpha_t)}{\Gamma(\sum_{k=1}^{K}\upsilon_{q,k}+\alpha_k)} \cdot \frac{\Gamma(\rho_{t,s}+\beta_s)}{\Gamma(\sum_{j=1}^{V}\rho_{t,j}+\beta_j)}}{\frac{\Gamma(\upsilon_{q,t}+\alpha_t-1)}{\Gamma(\sum_{k=1}^{K}\upsilon_{q,k}+\alpha_k-1)} \cdot \frac{\Gamma(\rho_{t,s}+\beta_s-1)}{\Gamma(\sum_{j=1}^{V}\rho_{t,j}+\beta_j-1)}}$$

$$\propto \frac{\upsilon_{q,t}+\alpha_t-1}{\sum_{k=1}^{K}\upsilon_{q,k}+\alpha_k-1} \cdot \frac{\rho_{t,s}+\beta_s-1}{\sum_{j=1}^{V}\rho_{t,j}+\beta_j-1}$$

Note that the counts $\upsilon_{m,k}$ and $\rho_{k,j}$ are updated over the Gibbs sampling iterations

Estimating Topic Model Parameters

Distribution of topics in documents

$$p(\boldsymbol{\theta}_m|\mathrm{Z}_{N_m},\boldsymbol{\alpha}) = \frac{1}{C_{\boldsymbol{\theta}_m}} p(\mathrm{Z}_{N_m}|\boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m|\boldsymbol{\alpha})$$
$$= \frac{1}{C_{\boldsymbol{\theta}_m}\Delta(\boldsymbol{\alpha})}\prod_{i=1}^{N_m}\prod_{k=1}^{K}\theta_{m,k}^{\mathbb{I}(\mathbf{z}_{(i,m)}=\boldsymbol{\xi}_k)}\theta_{m,k}^{\alpha_k-1}$$
$$= \frac{1}{C_{\boldsymbol{\theta}_m}\Delta(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_{m,k}^{\upsilon_{m,k}}\theta_{m,k}^{\alpha_k-1} = \mathrm{Dir}(\boldsymbol{\theta}_m|\boldsymbol{\upsilon}_m+\boldsymbol{\alpha})$$

Estimating Topic Model Parameters

Continued

Distribution of words in topics

Let $N(\boldsymbol{\xi}_k) = \{(i, m) : \mathbf{z}_{i,m} = \boldsymbol{\xi}_k\}$

$$p(\phi_k | \mathcal{X}_{N(\boldsymbol{\xi})}, \mathcal{Z}_{N(\boldsymbol{\xi})}, \boldsymbol{\beta}) = \frac{1}{C_{\phi_k}} p(\mathcal{X}_{N(\boldsymbol{\xi})} | \phi_k) p(\phi_k | \boldsymbol{\beta})$$

$$= \frac{1}{C_{\phi_k}} \prod_{N(\boldsymbol{\xi}_k)} p(\mathbf{x}_{i,m} | \phi_k) p(\phi_k | \boldsymbol{\beta})$$

$$= \frac{1}{C_{\phi_k} \Delta(\boldsymbol{\beta})} \prod_{N(\boldsymbol{\xi}_k)} \prod_{j=1}^{V} \phi_{k,j}^{\mathbb{I}(\mathbf{x}_{i,m} = \boldsymbol{\omega}_j)} \phi_{k,j}^{\beta_j - 1}$$

$$= \frac{1}{C_{\phi_k} \Delta(\boldsymbol{\beta})} \prod_{j=1}^{V} \phi_{k,j}^{\rho_{k,j} + \beta_j - 1} = \mathrm{Dir}(\phi_k | \boldsymbol{\rho}_k + \boldsymbol{\beta})$$

Estimating Topic Model Parameters

Continued

Given $\mathbf{x} = (x_1 \dots x_K) \sim \mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha})$

$$\mathrm{E}[x_i] = \frac{\alpha_i}{\bar{\alpha}} \qquad \mathrm{Var}[x_i] = \frac{\alpha_i(\bar{\alpha} - \alpha_i)}{\bar{\alpha}^2(\bar{\alpha} + 1)} \qquad \bar{\alpha} = \sum_{i=1}^{K} \alpha_i = \boldsymbol{\alpha}^T \mathbf{1}$$

Estimate for distribution of topics in documents $[a_k = (\boldsymbol{\upsilon}_m + \boldsymbol{\alpha})^T \mathbf{1}]$

$$\mathrm{E}[\theta_{m,k}] = \frac{\upsilon_{m,k} + \alpha_k}{a_k} \qquad \mathrm{Var}[\theta_{m,k}] = \frac{(\upsilon_{m,k} + \alpha_k)[a_k - (\upsilon_{m,k} + \alpha_k)]}{a_k^2(a_k + 1)}$$

Estimate for distribution of words in topics $[b_k = (\boldsymbol{\rho}_k + \boldsymbol{\beta})^T \mathbf{1}]$

$$\mathrm{E}[\phi_{k,j}] = \frac{\rho_{k,j} + \beta_j}{b_k} \qquad \mathrm{Var}[\phi_{k,j}] = \frac{(\rho_{k,j} + \beta_j)[b_k - (\rho_{k,j} + \beta_j)]}{b_k^2(b_k + 1)}$$