Generative Models

Donald E. Brown

School of Data Science
University of Virginia
Charlottesville, VA 22904

Causal or Generative Models

- Generative models often contain latent variables or parameters
- In most practical problems, the parameters or latent variables will be the lower numbered indices and the higher numbered indices represent observations
- Can interpret the graphical model as producing or generating the data
- These generative graphical models can be viewed as causal (see Pearl, 1989)
- Causality flows from the relationships shown in the conditional distributions
- This process is sometimes called *ancestral sampling*

A Generative Model for Text

D.E. Brown

- Goal: Find topics in text
- Approach: Latent Dirichlet Allocation (LDA) (Blei, et al., 2003)
- Notation

$w \in \{1, \ldots, V\}$	Word in a vocabulary of length V
$\mathbf{w} = (w_1, , \ldots, w_N)$	Document with N words
$D = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$	Corpus of M documents

- “We wish to find a probabilistic model of a corpus that not only assigns high probability to members of the corpus, but also assigns high probability to other 'similar' documents."

LDA Processes

- LDA - generative probabilistic model of a corpus, where for each $\mathbf{w} \in D$:
  1. Choose $N \sim \text{Poisson}(\xi)$
  2. Choose $\theta \sim \text{Dir}(\alpha)$
  3. For each of the N words w_n choose
     1. Choose a topic $z_n \sim \text{Mutinom}(\theta)$
     2. Choose a word, w_n, from $p(w_n|z_n, \boldsymbol{\beta})$, a multinomial conditioned on z_n
- Assumptions
  - Dimensionality, k, of the Dirichlet is known (i.e., topics)
  - Word probabilities parameterized by $\boldsymbol{\beta}$ a $k \times V$ matrix, where $\beta_{ij} = p(w = i|z = j)$

LDA Distributions

The joint distribution of a topic mixture $\boldsymbol{\theta}$, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is

$$p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\alpha, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|\alpha) \prod_{n=1}^{N} p(z_n|\boldsymbol{\theta}) p(w_n|z_n \boldsymbol{\beta})$$

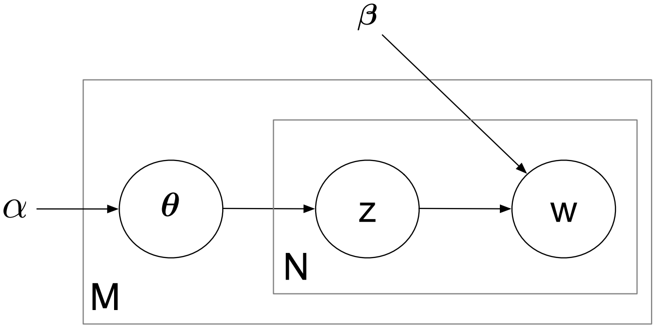
The marginal distribution of a document is

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

Probability of a corpus is

$$p(D|\alpha, \boldsymbol{\beta}) = \prod_{d=1}^{M} \int p(\boldsymbol{\theta}_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|z_{dn} \boldsymbol{\beta}) \right) d\boldsymbol{\theta}_d$$

LDA Graphical Model



LDA Inference

Compute the posterior distribution of the latent variables:

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\alpha, \boldsymbol{\beta})}{p(\mathbf{w}|\alpha, \boldsymbol{\beta})}$$

This is intractable; need to use sampling or variational inference.