An Introduction to Topic Modeling

Daniel W. Peterson

Department of Computer Science
University of Colorado at Boulder
daniel.w.peterson@colorado.edu

April 24, 2013

Latent Semantic Analysis

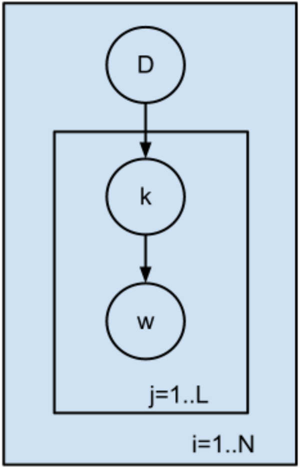
- Documents x Terms matrix: large and sparse
- Use SVD to decompose it into three matrices
- Keep only the "important" dimensions
- Assumptions:
  - Word order doesn't matter
  - Words are orthogonal dimensions in a high-dimensional space

Probabilistic Latent Semantic Analysis

- Documents are generated by a probabilistic process
  - Structure based on topics
  - Different topics make different words more likely
- Assumptions:
  - Word order doesn't matter
  - Each word is chosen as the result of exactly one topic

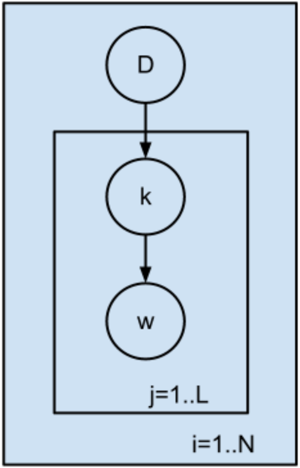
Probabilistic Latent Semantic Analysis

- N documents
- A document is L words long
- Each entry has an assignment to one of K topics



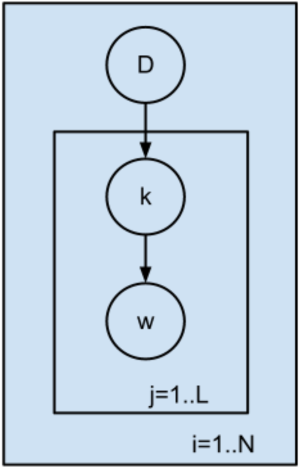
Probabilistic Latent Semantic Analysis

- How do we choose a topic?



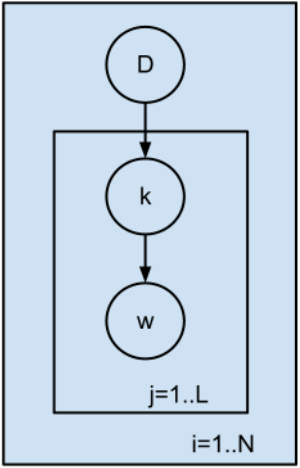
Probabilistic Latent Semantic Analysis

- How do we choose a topic?
  We sample from a distribution over topics.
- How do we choose a word?



Probabilistic Latent Semantic Analysis

- How do we choose a topic?
  We sample from a distribution over topics.
- How do we choose a word?
  We sample from a distribution over words.



Multinomial Distribution

- Select one of several possible outcomes

Multinomial Distribution

- Select one of several possible outcomes
- Outcomes may be equally likely (like dice)

Multinomial Distribution

- Select one of several possible outcomes
- Outcomes may be equally likely (like dice)
- OR: some outcomes may be more likely than others (load the dice)

Multinomial Distribution

- Select one of several possible outcomes
- Outcomes may be equally likely (like dice)
- OR: some outcomes may be more likely than others (load the dice)

- Looks like: a $1 \times n$ vector of probabilities
  - $[x_1, x_2, \ldots, x_n]$
  - $x_1 + x_2 + \ldots + x_n = 1$
  - every $x_i > 0$

Multinomial Distribution

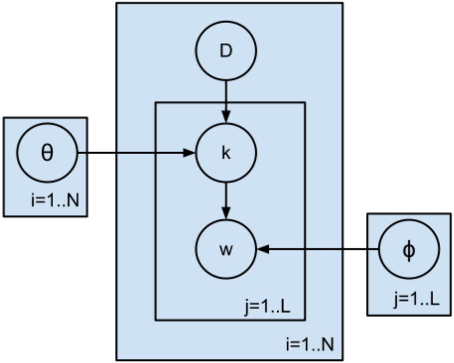
- Select one of several possible outcomes
- Outcomes may be equally likely (like dice)
- OR: some outcomes may be more likely than others (load the dice)

- Looks like: a $1 \times n$ vector of probabilities
  - $[x_1, x_2, \ldots, x_n]$
  - $x_1 + x_2 + \ldots + x_n = 1$
  - every $x_i > 0$
- A sample looks like: a number
  - The outcome of rolling the dice
  - Probability we get i is given by x_i

Probabilistic Latent Semantic Analysis

θ is a distribution over topics in a document

- One θ for each document
- θ is a $1 \times K$ vector
- Sum of θ is 1



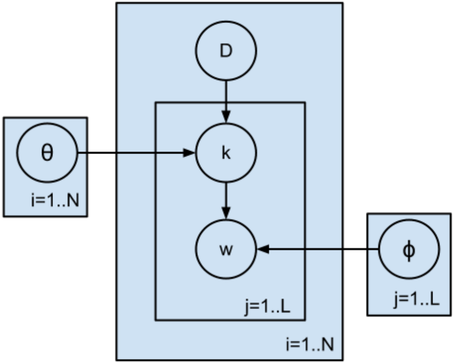
Probabilistic Latent Semantic Analysis

θ is a distribution over topics in a document

- One θ for each document
- θ is a $1 \times K$ vector
- Sum of θ is 1

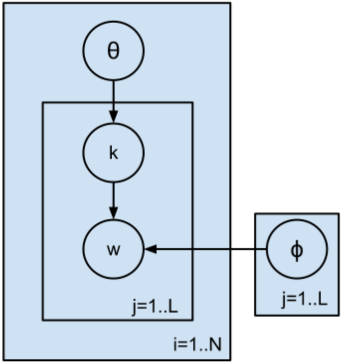
ϕ is a distribution over words in a topic

- One ϕ for each topic
- ϕ is a $1 \times W$ vector
- Sum of ϕ is 1



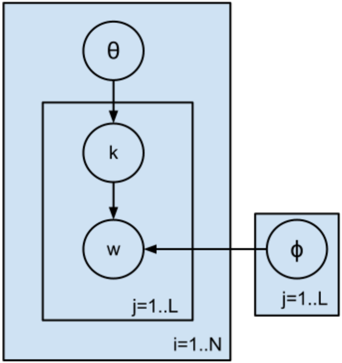
Probabilistic Latent Semantic Analysis

- Fold θ into graphical model



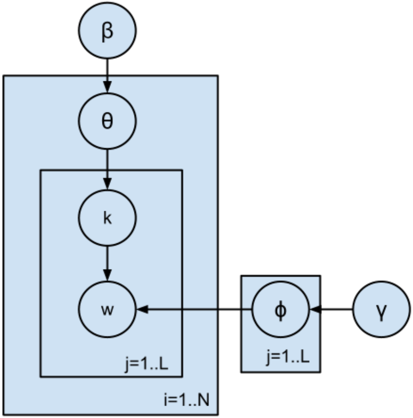
Probabilistic Latent Semantic Analysis

- Fold θ into graphical model
- Where do θ and ϕ come from?



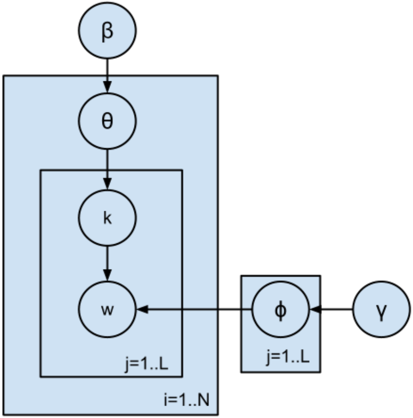
Topic Modeling

- Sample θ and ϕ from an appropriate distribution



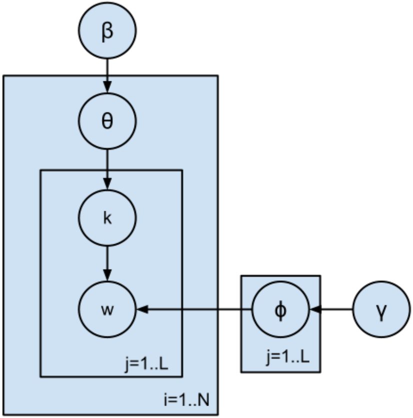
Topic Modeling

- Sample θ and ϕ from an appropriate distribution
- Dirchlet: a distribution over distributions



Topic Modeling

- Sample θ and ϕ from an appropriate distribution
- Dirchlet: a distribution over distributions
- Incorporating Dirichlet prior provides smoothing



Dirichlet Distribution

- Takes n parameters $\alpha_1, \alpha_2, \ldots, \alpha_n$
- Distribution over $1 \times n$ vectors with sum of 1
- α_i are called concentration parameters

Dirichlet Distribution with 2 Parameters

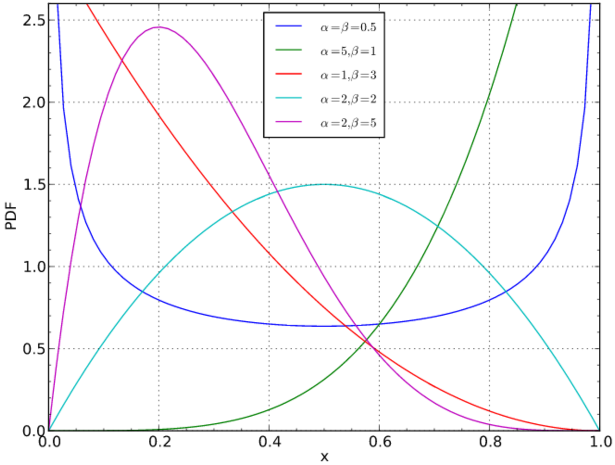


Figure: Image source: Wikipedia

Dirichlet Distribution with 3 Parameters

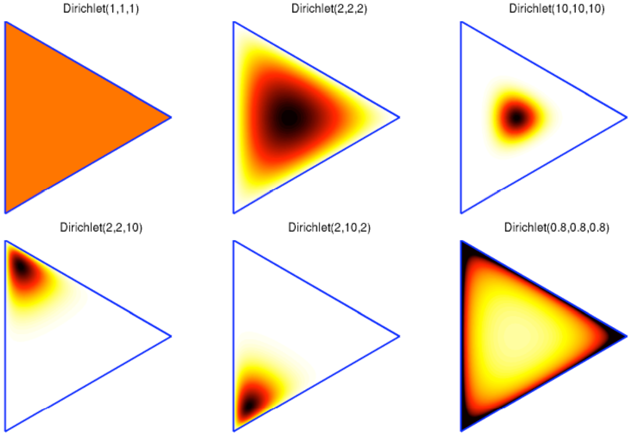


Figure: Image source: Yee Whye Teh

A Sample from a Dirichlet

- A particular $1 \times n$ vector with sum of 1

A Sample from a Dirichlet

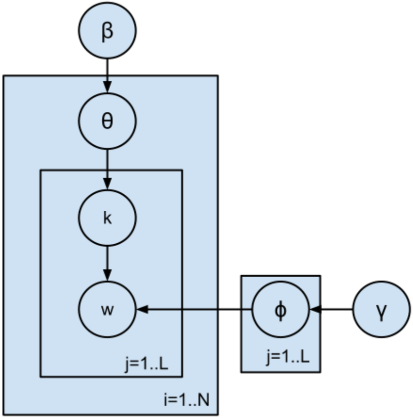
- A particular $1 \times n$ vector with sum of 1
- $[x_1, x_2, \ldots, x_n]$ such that $x_1 + x_2 + \ldots + x_n = 1$
- every $x_i > 0$

A Sample from a Dirichlet

- A particular $1 \times n$ vector with sum of 1
- $[x_1, x_2, \ldots, x_n]$ such that $x_1 + x_2 + \ldots + x_n = 1$
- every $x_i > 0$
- A multinomial distribution

Topic Modeling

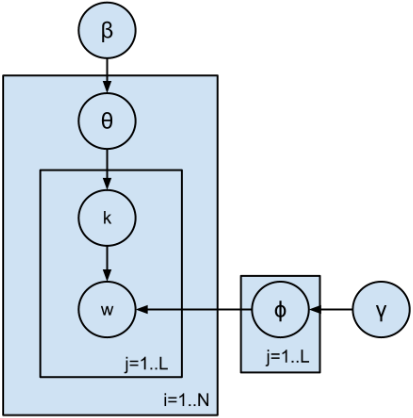
- Sample θ and ϕ from a Dirichlet distribution
- This is important for when we turn the model around:



Topic Modeling

- Sample θ and ϕ from a Dirichlet distribution
- This is important for when we turn the model around:
- Dirichlet distribution is conjugate prior of multinomial:

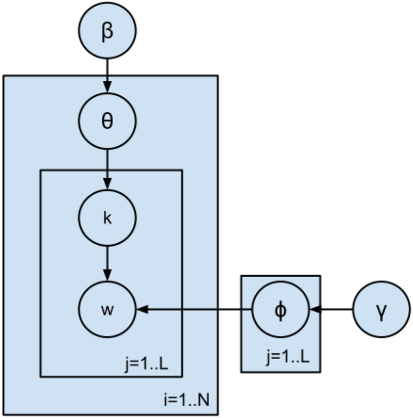
  Given a Dirichlet prior, and counts of topic assignments, the posterior is also Dirichlet



Topic Modeling

- Sample θ and ϕ from a Dirichlet distribution
- This is important for when we turn the model around:
- Dirichlet distribution is conjugate prior of multinomial:

  Given a Dirichlet prior, and counts of topic assignments, the posterior is also Dirichlet
- β and γ are smoothing parameters



Inference

- Generative model explains how the data was created

Inference

- Generative model explains how the data was created
- Inference: trying to guess model parameters

Gibbs Sampling

- Hard to determine most likely model parameters

Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters

Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can't sample from overall distribution: sample instead a single variable

Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can't sample from overall distribution: sample instead a single variable
- Take a walk through distribution

Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can't sample from overall distribution: sample instead a single variable
- Take a walk through distribution
    - One step (parameter) at a time

Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can't sample from overall distribution: sample instead a single variable
- Take a walk through distribution
  - One step (parameter) at a time
  - Spend more time walking around more likely areas

Gibbs Sampling

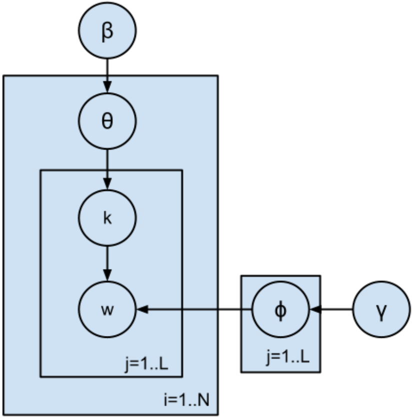
- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can't sample from overall distribution: sample instead a single variable
- Take a walk through distribution
    - One step (parameter) at a time
    - Spend more time walking around more likely areas
    - We can get to likely areas from anywhere

Gibbs Sampling

- Hard to determine most likely model parameters
- Hard for even relatively likely parameters
- Can't sample from overall distribution: sample instead a single variable
- Take a walk through distribution
    - One step (parameter) at a time
    - Spend more time walking around more likely areas
    - We can get to likely areas from anywhere
    - It doesn't matter where we start!

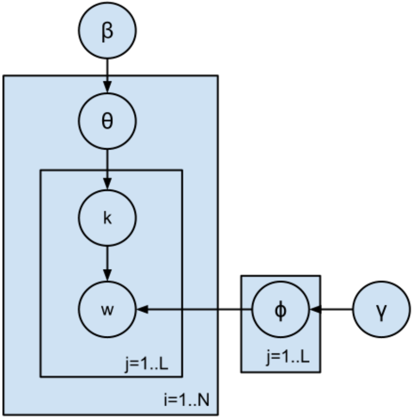
Gibbs Sampling in a Topic Model

- Start with random assignment of topics



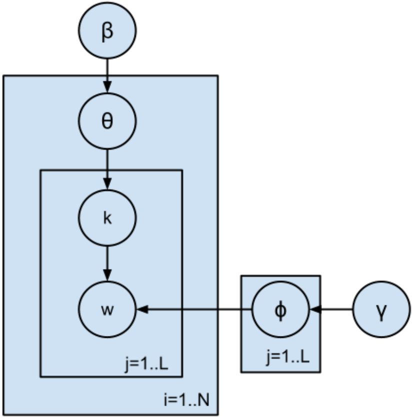
Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- For each $< word, document >$ pair:



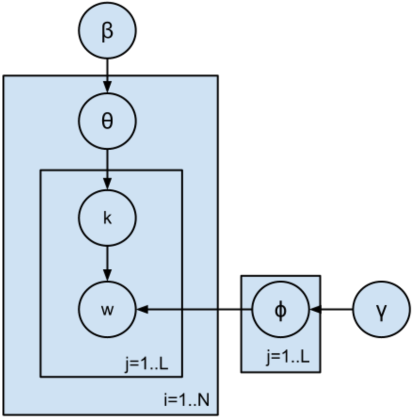
Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- For each $< word, document >$ pair:
  - Sample θ based on counts and prior



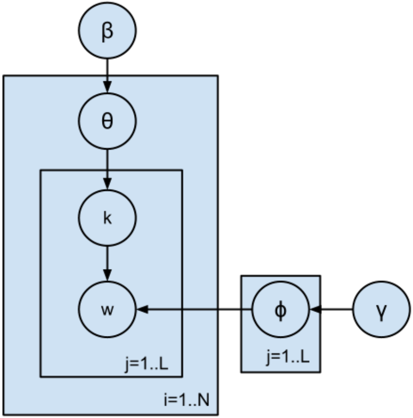
Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- For each $< word, document >$ pair:
  - Sample θ based on counts and prior
  - Sample ϕ based on counts and prior



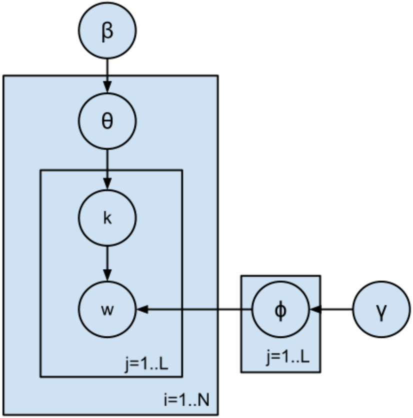
Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- For each $< word, document >$ pair:
  - Sample θ based on counts and prior
  - Sample ϕ based on counts and prior
  - Choose k based on θ, ϕ, and w



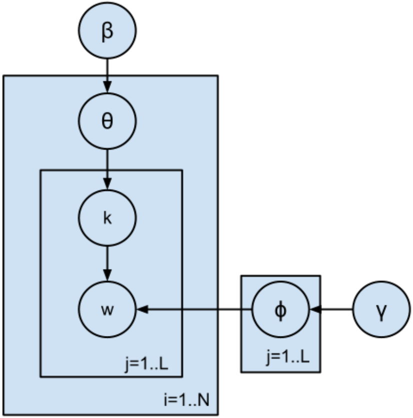
Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- For each $< word, document >$ pair:
  - Sample θ based on counts and prior
  - Sample ϕ based on counts and prior
  - Choose k based on θ, ϕ, and w
- Repeat the above many times



Gibbs Sampling in a Topic Model

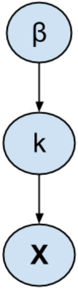
- Start with random assignment of topics
- For each $< word, document >$ pair:
  - Sample θ based on counts and prior
  - Sample ϕ based on counts and prior
  - Choose k based on θ, ϕ, and w
- Repeat the above many times
- Smoothing (β and γ) very important



Bayes Rule

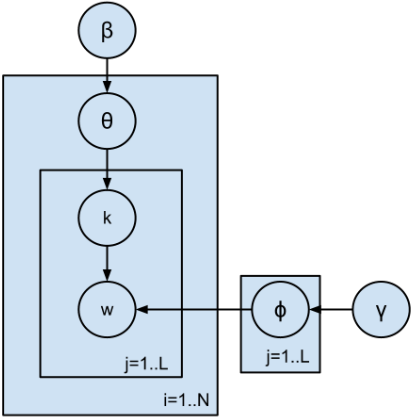
$$P(k|\beta, \mathbf{X}) \propto P(\mathbf{k}|\beta)P(\mathbf{X}|k)$$

Sampling from a conditional distribution can be broken down into sampling based on the parent nodes (prior, β) and the children (likelihood, $\mathbf{X}$)



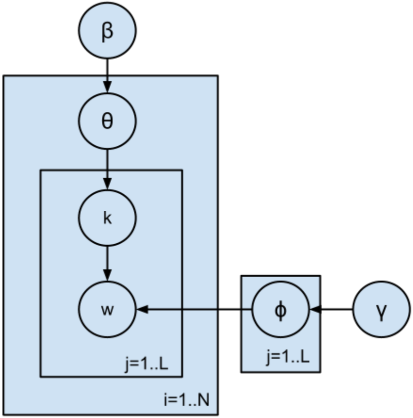
Blocked Gibbs Sampling in a Topic Model

- Start with random assignment of topics



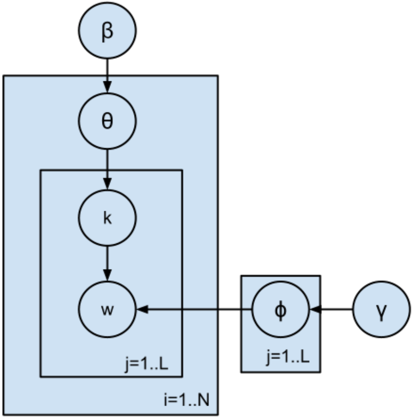
Blocked Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- Repeat many times:



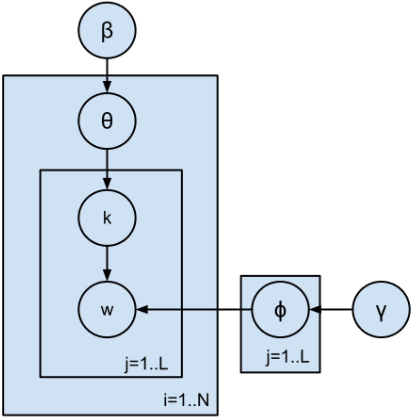
Blocked Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- Repeat many times:
    - Sample all θ and ϕ from counts and prior



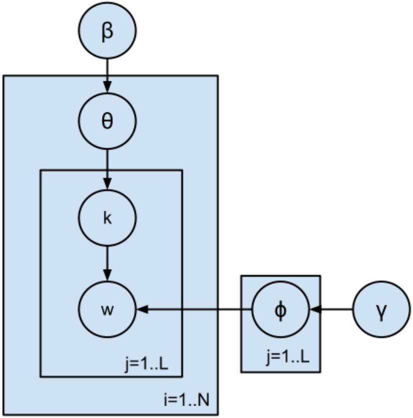
Blocked Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- Repeat many times:
  - Sample all θ and ϕ from counts and prior
  - Choose k for a number of $<$ *word*, *document* $>$ pairs



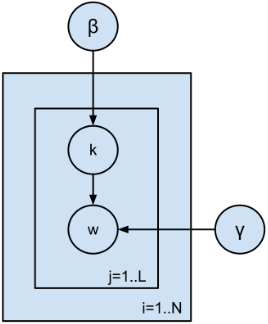
Blocked Gibbs Sampling in a Topic Model

- Start with random assignment of topics
- Repeat many times:
  - Sample all θ and ϕ from counts and prior
  - Choose k for a number of $< \textit{word}, \textit{document} >$ pairs
- More sampling, less counting



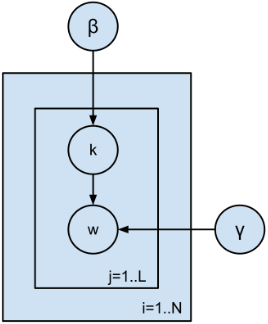
Collapsed Gibbs Sampling in a Topic Model

- Integrate out θ and ϕ



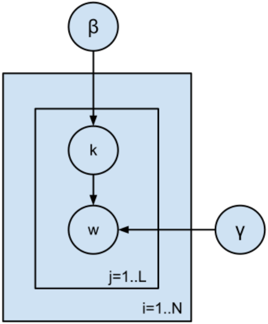
Collapsed Gibbs Sampling in a Topic Model

- Integrate out θ and ϕ
- Start with random assignment of topics



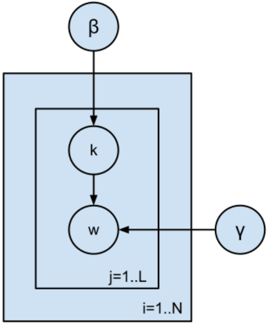
Collapsed Gibbs Sampling in a Topic Model

- Integrate out θ and ϕ
- Start with random assignment of topics
- For each $<$ *word*, *document* $>$ pair:



Collapsed Gibbs Sampling in a Topic Model

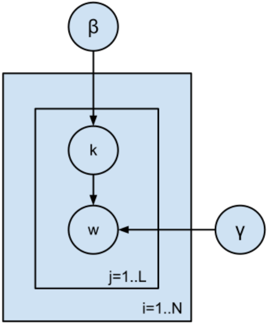
- Integrate out θ and ϕ
- Start with random assignment of topics
- For each $< word, document >$ pair:
  - Sample k directly from counts



$$P(z_i = k | z_{-i}, w) \propto \frac{n_{-i,k}^{(w_i)} + \gamma}{n_{-i,k}^{(\cdot)} + W\gamma} \frac{n_{-i,k}^{(d_i)} + \beta}{n_{-i,\cdot}^{(d_i)} + K\beta}$$

Collapsed Gibbs Sampling in a Topic Model

- Integrate out θ and ϕ
- Start with random assignment of topics
- For each $< word, document >$ pair:
  - Sample k directly from counts
- Repeat many times



$$P(z_i = k | z_{-i}, w) \propto \frac{n_{-i,k}^{(w_i)} + \gamma}{n_{-i,k}^{(\cdot)} + W\gamma} \frac{n_{-i,k}^{(d_i)} + \beta}{n_{-i,\cdot}^{(d_i)} + K\beta}$$