Statistical Language Models

Week 7

- $P(\lambda) = \Gamma(\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$
- $P(Y) = Poisson(\lambda) = \frac{1}{y!} \lambda^{y} e^{-\lambda}$

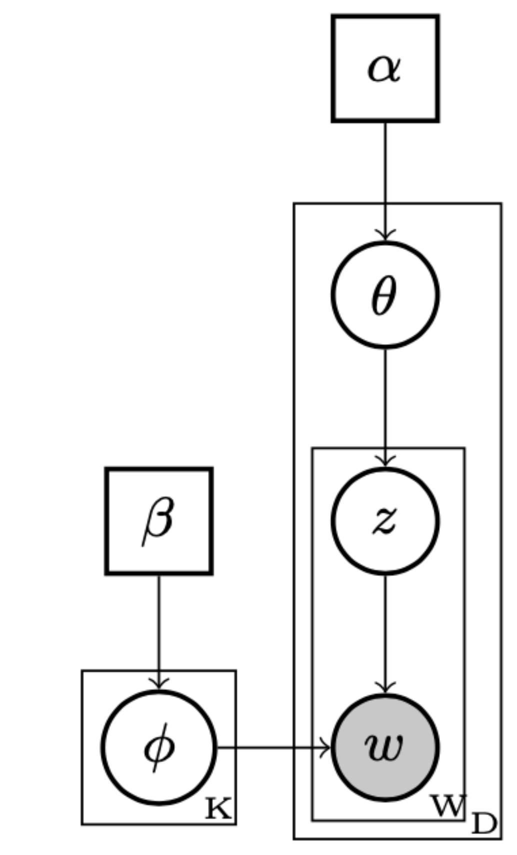


Figure 1: Graphical model for LDA.

- LDA assumes the following generative process for each document **w** in a corpus **D**:

- 1. Choose N ~ Poisson(ξ), there are N word in the document
  2. Choose a topic allocation θ~Dir(α), the Dirichlet has a prior vector α.
  3. For each of the **N** words **w**$_n$:

- (a) Choose a topic $\mathbf{z_n}$ ~ Multinomial(**θ**). Each position in the doc has a latent topic
  (b) Choose a word $\mathbf{w_n}$ from $\mathbf{p(w_n \mid z_n,\phi)}$, a multinomial probability conditioned on the topic $\mathbf{z_n}$. Each topic has it's own pdf over words, the word Dirichlet has prior vector **β**

- library(tidyverse)
- library(tidytext)
- library(topicmodels)
- library(tm)
- library(dplyr)

GETTING STARTED WITH THE BEATLES

- wordcount = uniquesongs %>%
- unnest_tokens(output = word, input = songlyric) %>%
- #anti_join(stop_words) %>%
- group_by(track_title) %>%
- count(word,sort=TRUE)%>%
- ungroup()
- DTM = wordcount %>% cast_dtm(term=word,document=track_title,value=n)

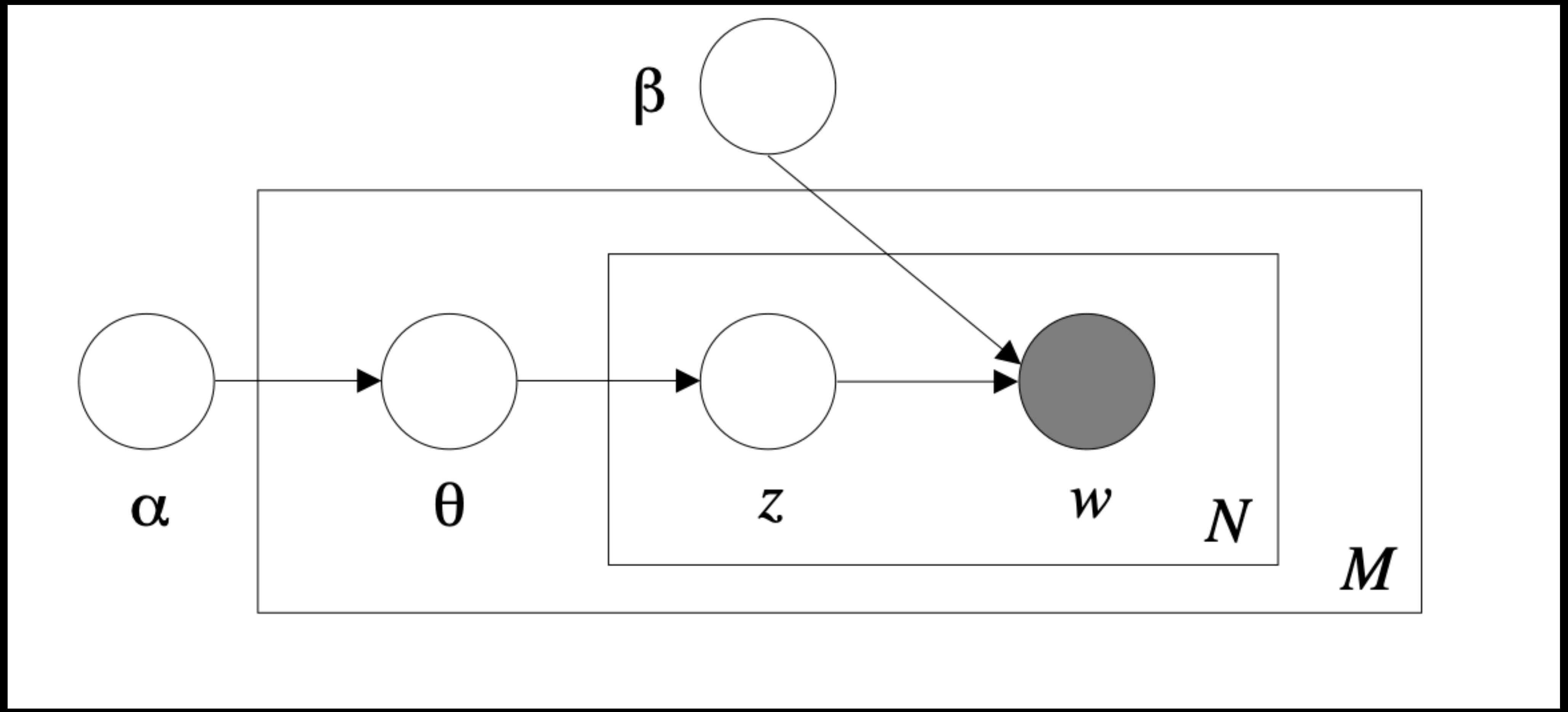
Beatles LDA

- #remove sparse documents
- DTM95 = removeSparseTerms(DTM,.95)
- DTM95matrix = as.matrix(DTM95)
- #less than 5 words or less than 5 unique words are too sparse
- Removethese = which(apply(DTM95matrix,1,function(x){sum(x>0)<5 | sum(x)<5}))
- DTM = wordcount %>% filter (!(wordcount$track_title %in% names(Removethese)))%>%

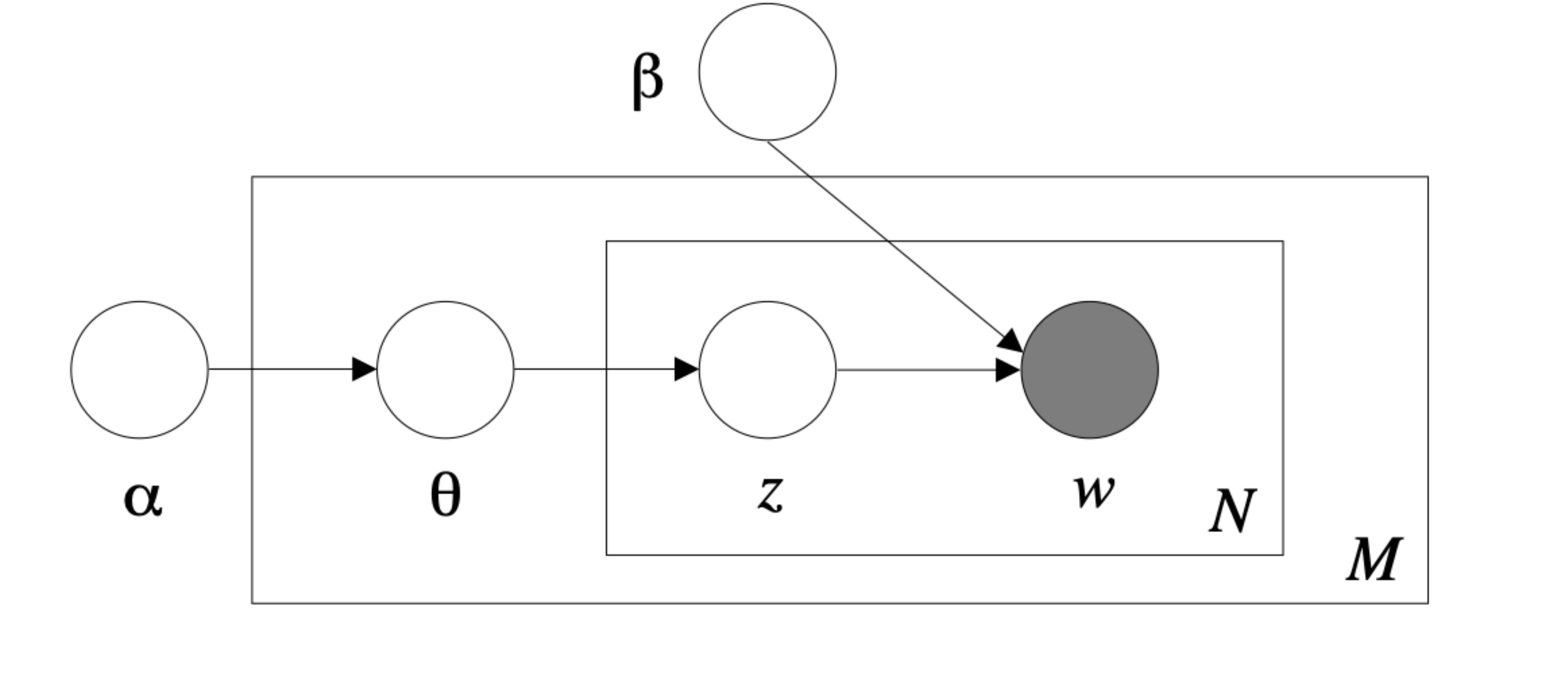
  cast_dtm(term=word,document=track_title,value=n)

Beatles LDA

- k=9
- BeatlesLDA = LDA(DTM, k,method="Gibbs")
- topics = tidy(BeatlesLDA, matrix = "beta")



- library(ggplot2)
- library(dplyr)
- TopWords = topics %>%
- group_by(topic) %>%   # take an action within topic values
- top_n(10, beta) %>%   # find the largest 10 values based on the 'beta' column
- ungroup() %>%   # stop acting within a topic
- arrange(topic, -beta)   # sort the



- TopWords %>%
- mutate(term = reorder_within(term, beta, topic)) %>% # Used for faceting (glue topic to term) basically make sure that topic 1 is my topic #1
- ggplot(aes(term, beta, fill = factor(topic))) +
- geom_col(show.legend = FALSE) +
- facet_wrap(~ topic, scales = "free") +
- coord_flip() +
- scale_x_reordered()

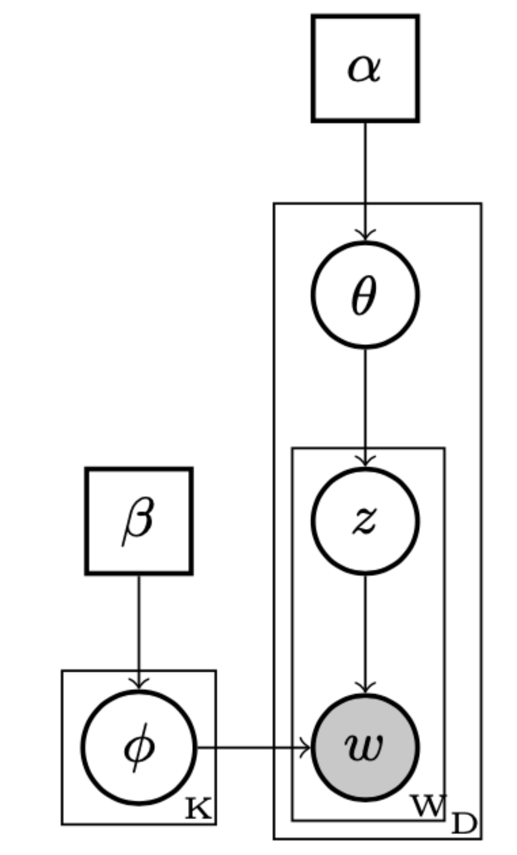


Figure 1: Graphical model for LDA.

- LDA assumes the following generative process for each document **w** in a corpus **D**:

- 1. Choose N ~ Poisson(ξ), there are N word in the document
  2. Choose a topic allocation θ~Dir(α), the Dirichlet has a prior vector α.
  3. For each of the **N** words $\mathbf{w}_n$:

- (a) Choose a topic $\mathbf{z_n}$ ~ Multinomial(**θ**). Each position in the doc has a latent topic
  (b) Choose a word $\mathbf{w_n}$ from $\mathbf{p(w_n \mid z_n, \phi)}$, a multinomial probability conditioned on the topic $\mathbf{z_n}$. Each topic has it's own pdf over words, the word Dirichlet has prior vector **β**

Full LDA Model

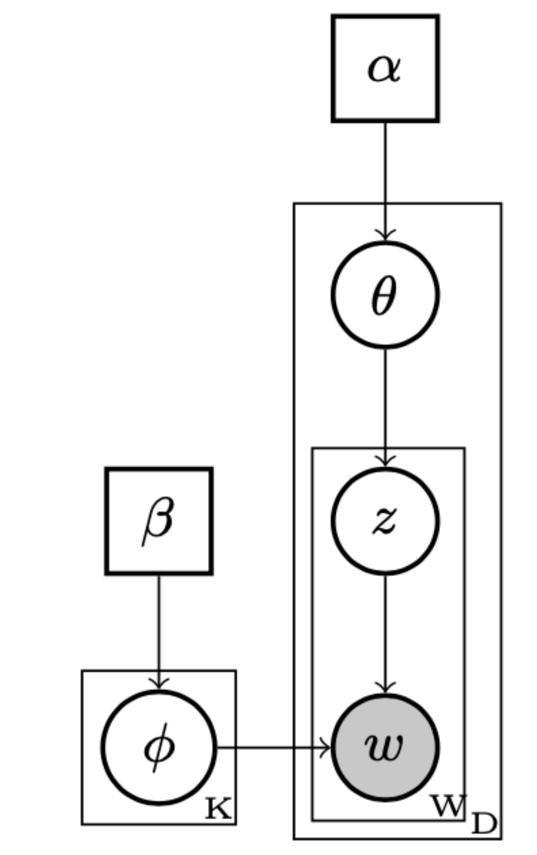


Figure 1: Graphical model for LDA.

- $$P(W, Z, \theta, \phi \mid \alpha, \beta) = \prod_{i=1}^{K} P(\phi_k \mid \beta) \prod_{j=1}^{M} P(\theta_j \mid \alpha) \prod_{t=1}^{N_j} P(z_{j,t} \mid \theta_j) P(w_{jt} \mid \phi_{z_{jt}})$$

Park, H., Park, T., & Lee, Y. S. (2010). Partially collapsed Gibbs sampling for latent Dirichlet allocation. Journal of Machine Learning Research, 13, 63–78. https://doi.org/10.1016/j.eswa.2019.04.028

Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation

Ian Porteous
Dept. of Computer Science
University of California, Irvine
Irvine, CA 92697-3425
iporteou@ics.uci.edu

David Newman
Dept. of Computer Science
University of California, Irvine
Irvine, CA 92697-3425
newman@uci.edu

Alexander Ihler
Dept. of Computer Science
University of California, Irvine
Irvine, CA 92697-3425
ihler@ics.uci.edu

Arthur Asuncion
Dept. of Computer Science
University of California, Irvine
Irvine, CA 92697-3425
asuncion@ics.uci.edu

Padhraic Smyth
Dept. of Computer Science
University of California, Irvine
Irvine, CA 92697-3425
smyth@ics.uci.edu

Max Welling
Dept. of Computer Science
University of California, Irvine
Irvine, CA 92697-3425
welling@ics.uci.edu

Gibbs Sampling

- Goal: sample from target distribution P(A,B)
- Conditional distributions are available: P(A|B) and P(B|A)
- Generate a sample of size N from P(A,B) via:
- for(1 in 1:N){
    - A_{i+1} ~ P(A|B_i)
    - B_{i+1} ~ P(B|A_{i+1})
- }

Gibbs Sampling in LDA

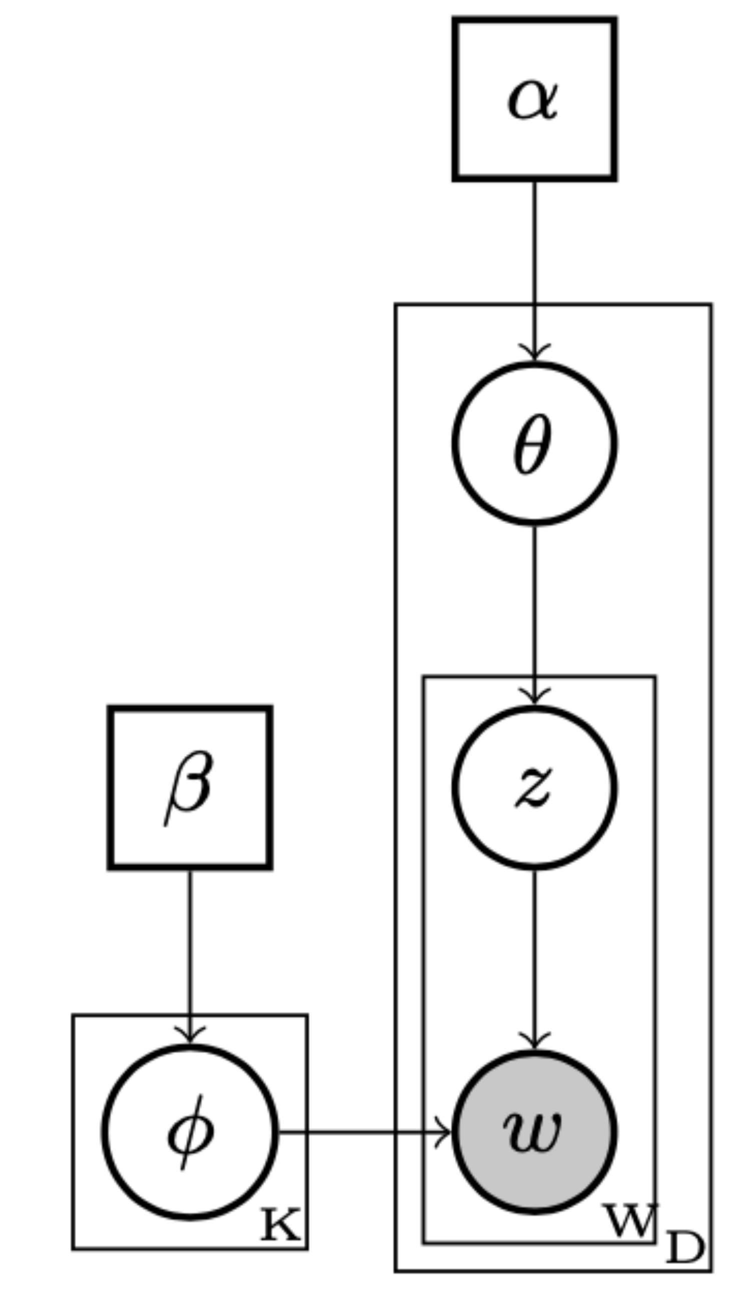


Figure 1: Graphical model for LDA.

- Within LDA, the generative process:
    - Sample topic z_{ij} ~ Multinomial(θ_j)
    - Sample word w_{ij} ~ Multinomial($\phi_{z_{ij}}$)
- Multinomial Parameter $\phi_{z_{ij}}$ indicates which words are important in topic z_{ij}
- Multinomial Parameter θ_j indicates which topics are important in document j

Gibbs Sampling in LDA

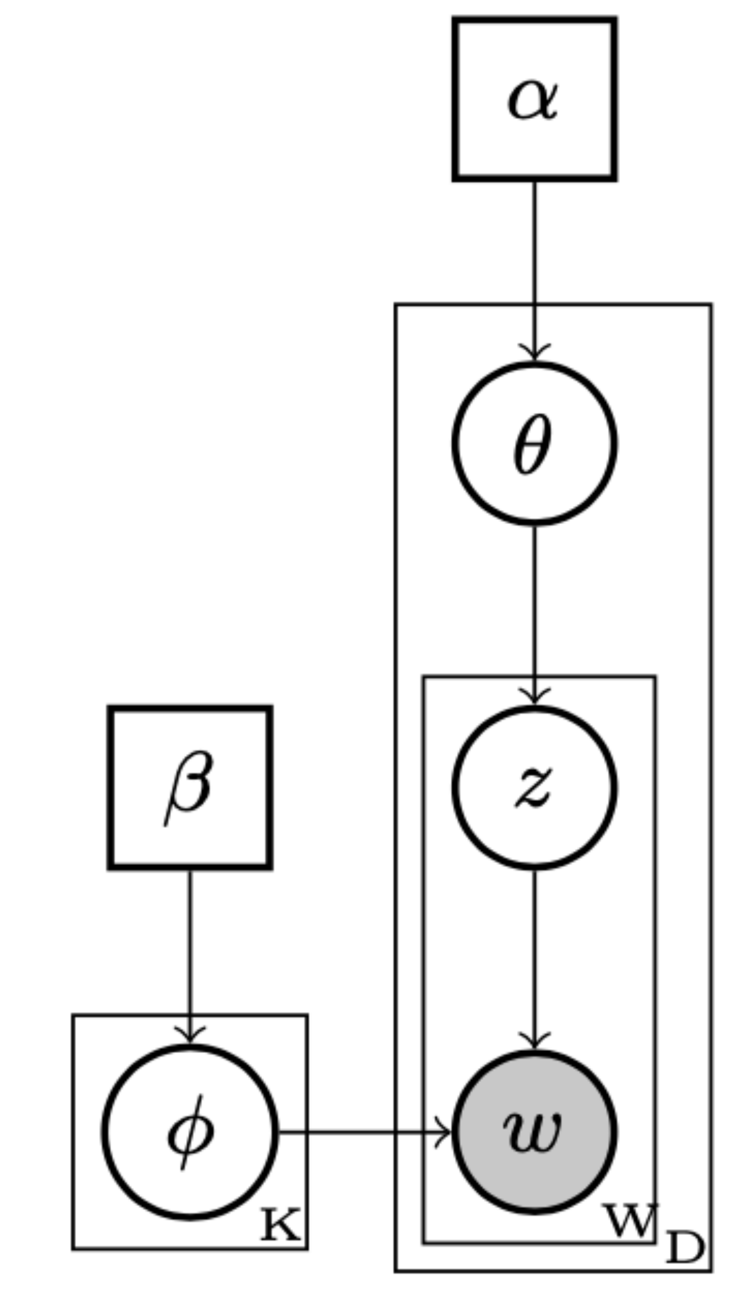


Figure 1: Graphical model for LDA.

- Within LDA, the generative process:
  - Sample topic z_{ij} ~ Multinomial(θ_j)
  - Sample word w_{ij} ~ Multinomial($\phi_{z_{ij}}$)
- Multinomial Parameter $\phi_{z_{ij}}$ indicates which words are important in topic z_{ij}
- Multinomial Parameter θ_j indicates which topics are important in document j
- Want posterior P(z_{ij}, $\phi_{z_{ij}}$, θ_j | w_{ij})
- (yes I'm being sloppy by hiding both α and β, but it's visually cleaner)

Collapsed Gibbs Sampling in LDA

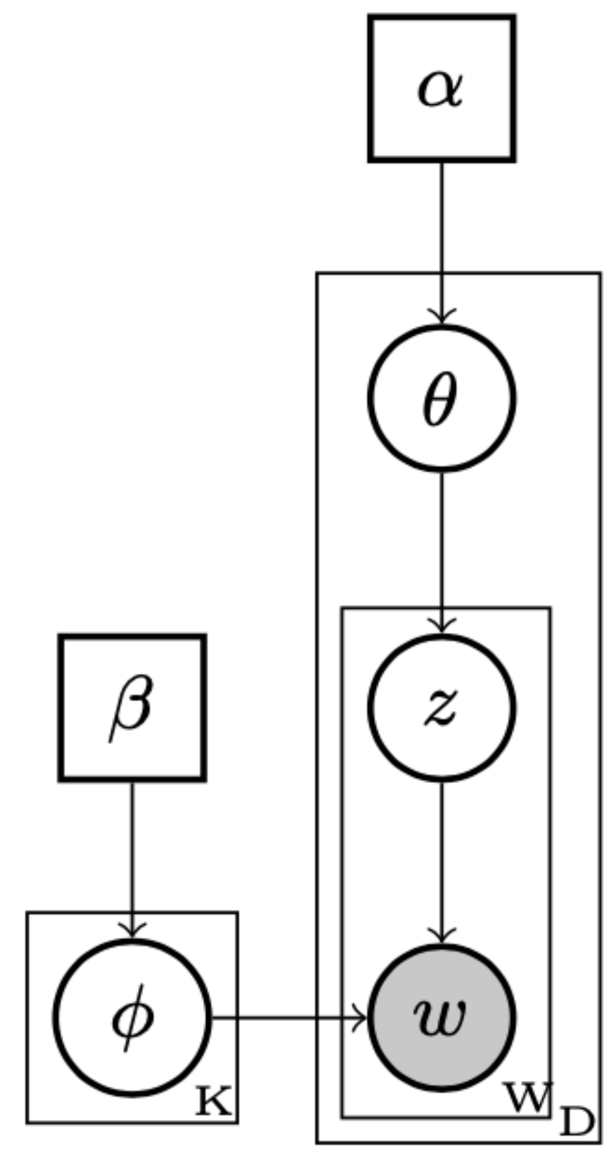


Figure 1: Graphical model for LDA.

- Want posterior P(z_{ij}, ϕ_{zij}, θ_j | w_{ij})

- Sample instead from (collapsed) posterior for latent topics:

$$P(z_{ij}, w_{ij}) = \int\int P\left(z_{ij}, \phi_{z_{ij}}, \theta_j, w_{ij}\right) d\theta d\beta$$

- After the sampler has burned-in, estimate: $\hat{\phi}_{z_{ij}}, \hat{\theta}_j \mid z_{ij}$

Collapsed Gibbs Sampling in LDA

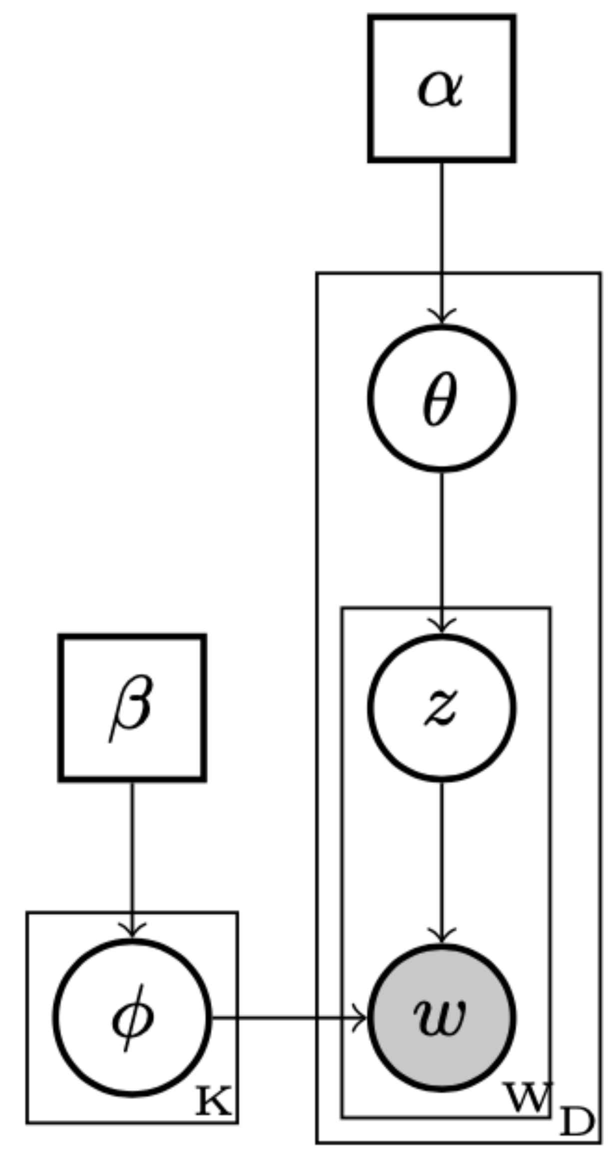


Figure 1: Graphical model for LDA.

- But first, the integrals:
- $P(z_{ij}, w_{ij}) = \int\int P\left(z_{ij}, \phi_{z_{ij}}, \theta_j, w_{ij}\right) d\theta d\phi$
- Now α and β matter so we'll explicitly mention them and note that:
- $P(W, Z \mid \alpha, \beta) = P(Z \mid \alpha) P(W \mid Z, \beta)$

$P(z \mid \alpha)$

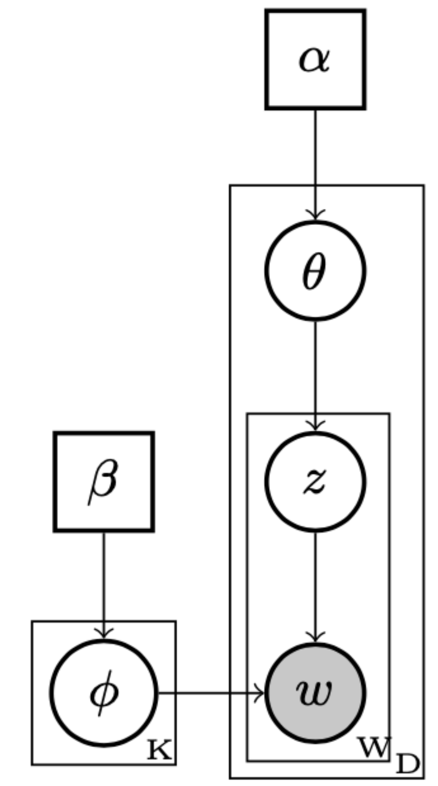


Figure 1: Graphical model for LDA.

- $P(Z \mid \alpha) = \int \left[P(\theta \mid \alpha)\right] P(Z \mid \theta) d\theta$

- $P(Z \mid \alpha) = \int \left[\prod_{j=1}^{M} \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_{ji}^{\alpha_i - 1}\right] \left[\prod_{j=1}^{M} \prod_{i=1}^{N_j} \theta_{jz_{jt}}\right] d\theta$

- $P(Z \mid \alpha) = \int \left[\prod_{j=1}^{M} \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_{ji}^{\alpha_i - 1}\right] \left[\prod_{j=1}^{M} \prod_{i=1}^{K} \theta_{jz_{jt}}^{\sum_t^{N_j} \mathbb{I}(z_{jt}=k)}\right] d\theta$

$$P(z \mid \alpha)$$

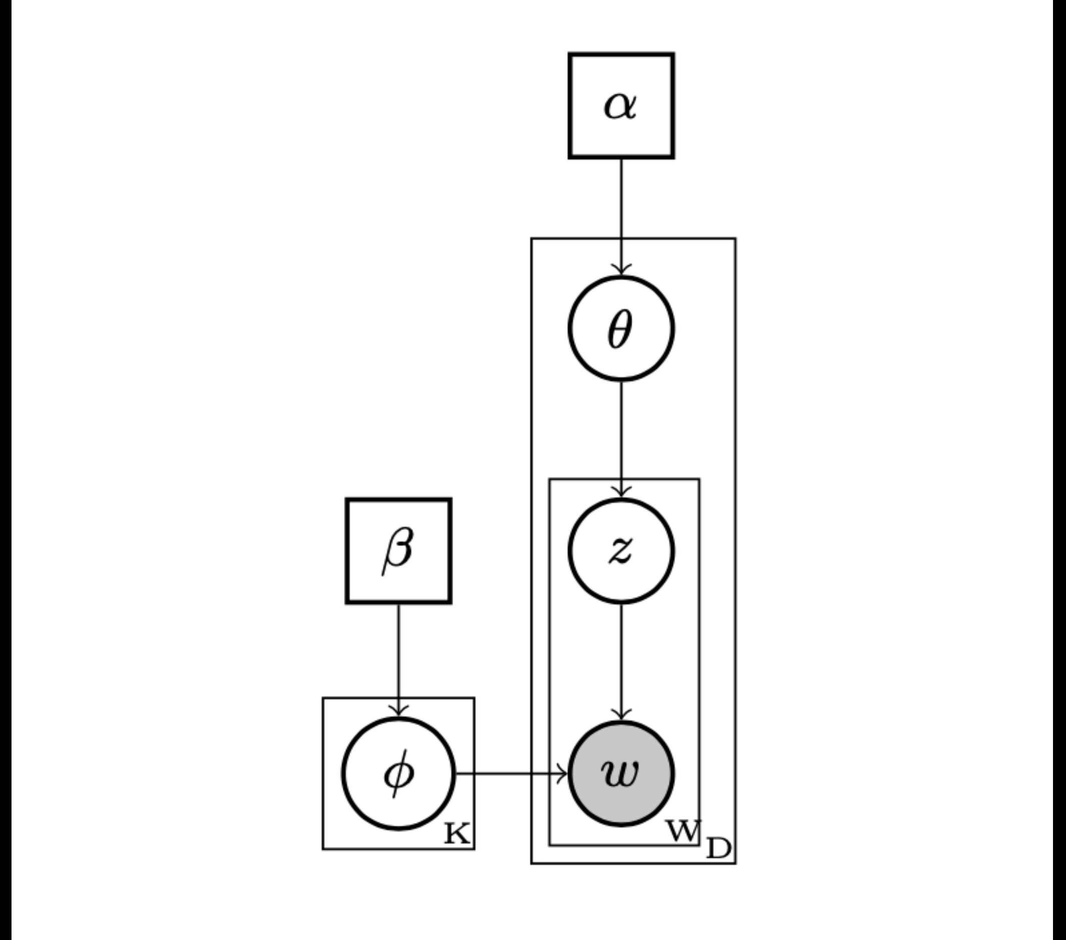


Figure 1: Graphical model for LDA.

- $$P(Z \mid \alpha) = \int \left[\prod_{j=1}^{M} \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_{ji}^{\alpha_i - 1} \right] \left[\prod_{j=1}^{M} \prod_{i=1}^{K} \theta_{j z_{jt}}^{\sum_t^{N_j} \mathbb{I}(z_{jt}=k)} \right] d\theta$$
- $$P(Z \mid \alpha) = \left[\prod_{j=1}^{M} \frac{B\left(\alpha + \sum_k \sum_t^{N_j} \mathbb{I}(z_{jt} = k)\right)}{B(\alpha)} \right]$$

$$P(W \mid Z, \beta)$$

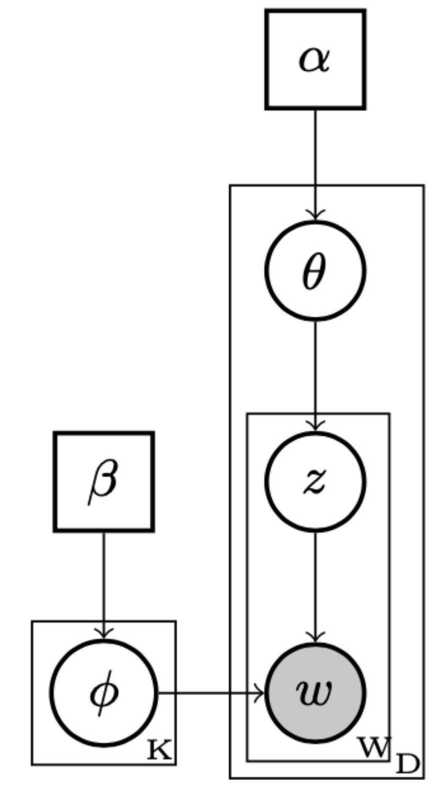


Figure 1: Graphical model for LDA.

- $P(W \mid Z, \beta) = \int \left[P(\phi \mid \beta)\right] \left[P(W \mid Z, \phi)\right] d\phi$
- $P(W \mid Z, \beta) = \int \left[\prod_{i=1}^{K} \frac{1}{B(\beta)} \prod_{r=1}^{V} \phi_{ir}^{\beta_r - 1}\right] \left[\prod_{i=1}^{K} \prod_{t=1}^{N_j} P(w_{jt} \mid \phi_{z_j t})\right] d\phi$

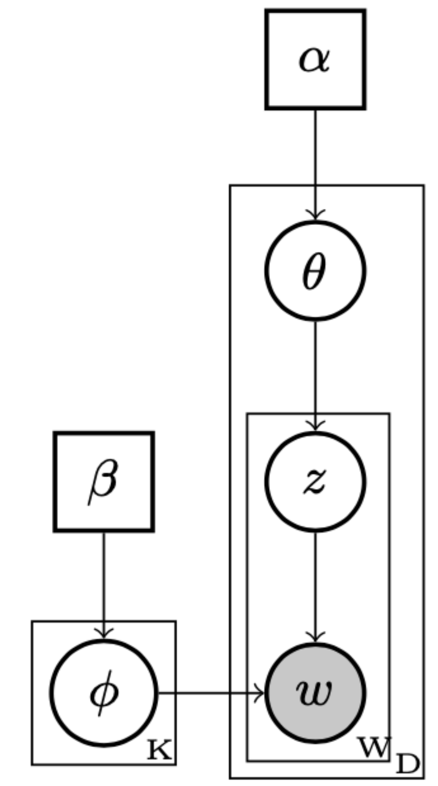


Figure 1: Graphical model for LDA.

$P(W \mid Z, \beta)$

- $P(W \mid Z, \beta) = \int \left[P(\phi \mid \beta)\right] \left[P(W \mid Z, \phi)\right] d\phi$
- $P(W \mid Z, \beta) = \int \left[\prod_{i=1}^{K} \frac{1}{B(\beta)} \prod_{r=1}^{V} \phi_{ir}^{\beta_r - 1}\right] \left[\prod_{i=1}^{K} \prod_{t=1}^{N_j} P(w_{jt} \mid \phi_{z_j t})\right] d\phi$
- $P(W \mid Z, \beta) = \int \left[\prod_{i=1}^{K} \frac{1}{B(\beta)} \prod_{r=1}^{V} \phi_{ir}^{\beta_r - 1}\right] \left[\prod_{i=1}^{K} \prod_{r=1}^{V} \phi_{ir}^{\sum_{j=1}^{M} \sum_{t}^{N_j} \mathbb{I}(w_{jt}=r \& z_{ij}=i)}\right] d\phi$
- # times word r appears in topic i

$P(W \mid Z, \beta)$

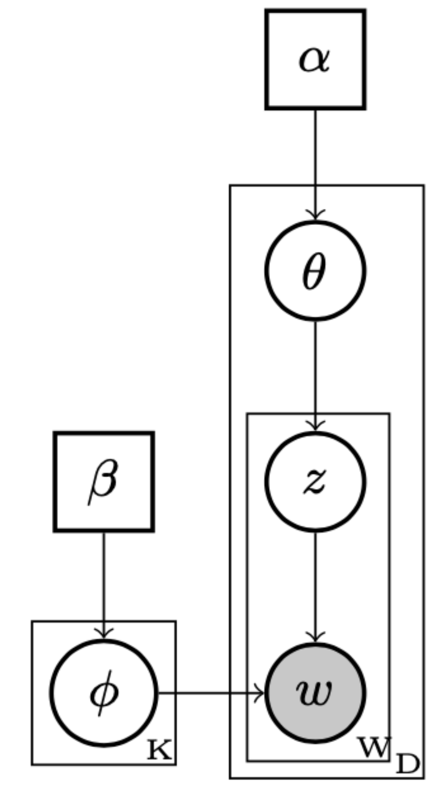


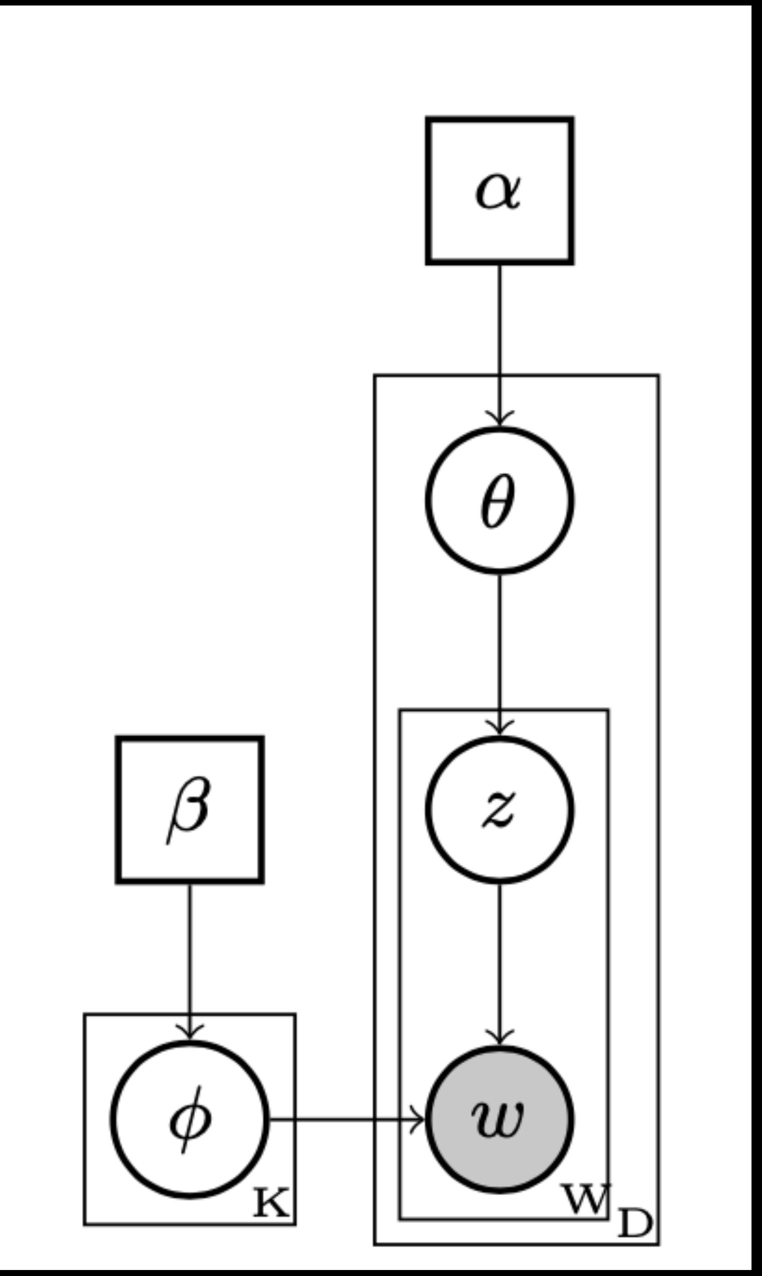
Figure 1: Graphical model for LDA.

- $P(W \mid Z, \beta) = \int \left[P(\phi \mid \beta) \right] \left[P(W \mid Z, \phi) \right] d\phi$
- $P(W \mid Z, \beta) = \int \left[\prod_{i=1}^{K} \frac{1}{B(\beta)} \prod_{r=1}^{V} \phi_{ir}^{\beta_r - 1} \right] \left[\prod_{i=1}^{K} \prod_{r=1}^{V} \phi_{ir}^{\sum_{j=1}^{M} \sum_{t}^{N_j} \mathbb{I}(w_{jt}=r \& z_{it}=i)} \right] d\phi$
- $P(W \mid Z, \beta) = \prod_{i=1}^{K} \frac{B\left(\beta + \sum_{j=1}^{M} \sum_{t}^{N_j} \mathbb{I}(w_{jt} = r \,\&\, z_{it} = i)\right)}{B(\beta)}$

- Document j, topic k, word index i
- Observed word w_{ij} and generative random variable word x_{ij}
- At each iteration we need $$P(z_{ij} = k \mid z^{\neg ij}, x, \alpha, \beta) = \frac{1}{Z}\alpha_{kj} b_{wk}$$
- Where $\alpha_{kj} = N_{kj}^{\neg ij} + \alpha$, (# words in a topic + α)
- Where $b_{wk} = \dfrac{N_{wk}^{\neg ij} + \beta}{N_k^{\neg ij} + W\beta}$, (# documents with word topic+ β) ÷

(# documents and words in a topic + vocal size(W) * β)

- $N_{wk} = \sum_j N_{wkj}$, $N_k = \sum_j \sum_w N_{wkj}$, $N_{wkj} = \#\{i : x_{ij} = w, z_{ij} = k\}$



- Document j, topic k, word index i

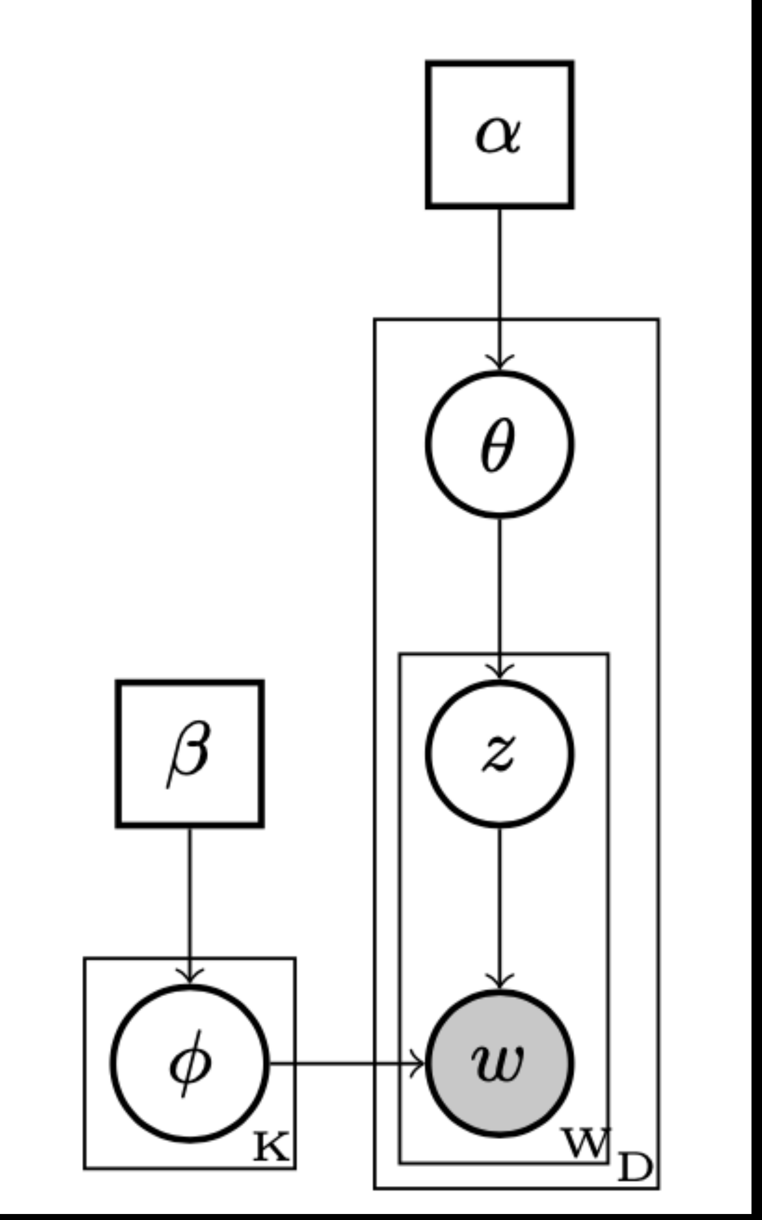
- Observed word w_{ij} and generative random variable word x_{ij}

- At each iteration we need $$P(z_{ij} = k \mid z^{\neg ij}, x, \alpha, \beta) = \frac{1}{Z}\alpha_{kj} b_{wk}$$

- Sample z_{ij} then update the counts N_{kj}, N_k, N_{wk}.

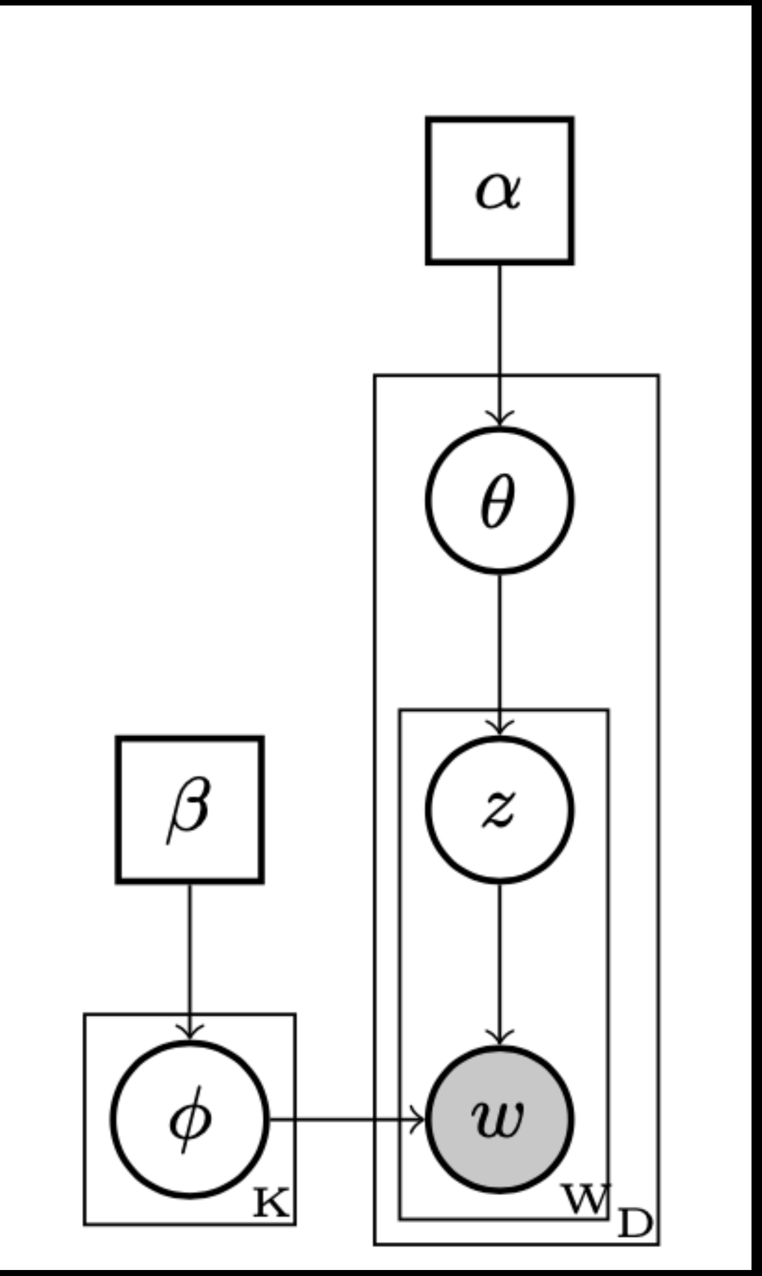
- Finally estimate

- $$\hat{\phi}_{wk} = \frac{N_{wk} + \beta}{N_k + W\beta}, \quad \hat{\phi}_{wk} = \frac{N_{kj} + \alpha}{N_j + K\alpha}$$



$$
\begin{array}{l}
\textbf{for } i \leftarrow 1 \textbf{ to } N \\
\quad \textbf{do} \\
\left\{\begin{array}{l}
u \leftarrow \text{draw from Uniform}[0,1] \\
\textbf{for } k \leftarrow 1 \textbf{ to } K \\
\quad \textbf{do} \\
\left\{ P[k] \leftarrow P[k-1] + \frac{(N_{kj}^{\neg ij} + \alpha)(N_{x_{ij}k}^{\neg ij} + \beta)}{(N_k^{\neg ij} + W\beta)} \right. \\
\textbf{for } k \leftarrow 1 \textbf{ to } K \\
\quad \textbf{do} \\
\left\{\begin{array}{l} \textbf{if } u < P[k]/P[K] \\ \quad \textbf{then } z_{ij} = k, stop \end{array}\right.
\end{array}\right.
\end{array}
$$

- Document j, topic k, word index i
- Observed word w_{ij} and generative random variable word x_{ij}
- At each iteration we need $$P(z_{ij} = k \mid z^{\neg ij}, x, \alpha, \beta) = \frac{1}{Z} a_{kj} b_{wk}$$
- Sample z_{ij} then update the counts N_{kj}, N_k, N_{wk}.

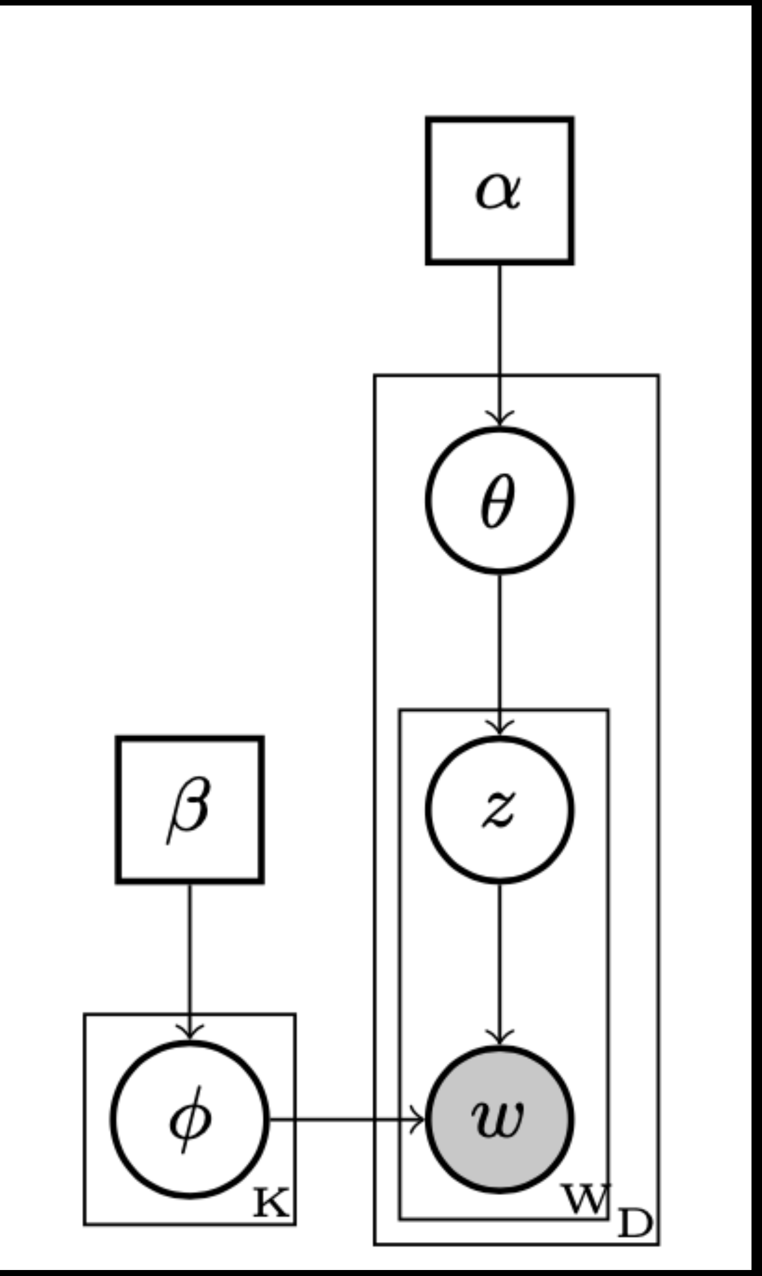


for $i \leftarrow 1$ **to** N
do
$u \leftarrow$ draw from Uniform$[0,1]$
for $k \leftarrow 1$ **to** K
do
$P[k] \leftarrow P[k-1] + \frac{(N_{kj}^{\neg ij} + \alpha)(N_{x_{ij}k}^{\neg ij} + \beta)}{(N_k^{\neg ij} + W\beta)}$
for $k \leftarrow 1$ **to** K
do
if $u < P[k]/P[K]$
then $z_{ij} = k, stop$

- Now estimate topic and word allocations

- $$\theta_{jk} \approx \frac{\sum_t^{N_j} \mathbb{I}(z_{jt} = k) + \alpha_k}{\sum_{i=1}^{K} \sum_t^{N_j} \mathbb{I}(z_{jt} = k) + \alpha_i}$$

- $$\phi_{kv} \approx \frac{\sum_{j=1}^{M} \sum_t^{N_j} \mathbb{I}(w_{jt} = v \,\&\, z_{it} = k) + \beta_v}{\sum_{r=1}^{V} \sum_{j=1}^{M} \sum_t^{N_j} \mathbb{I}(w_{jt} = r \,\&\, z_{it} = k) + \beta_r}$$



Variational Inference for LDA

- Gibbs & Variational EM Algorithm
- http://times.cs.uiuc.edu/course/598f16/notes/lda-survey.pdf

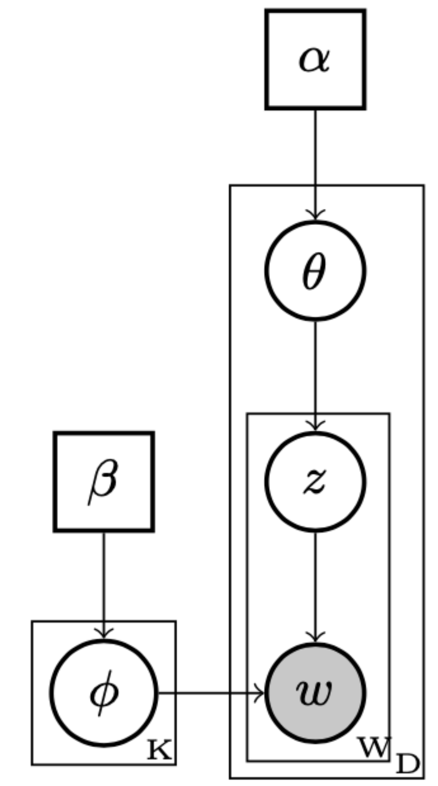
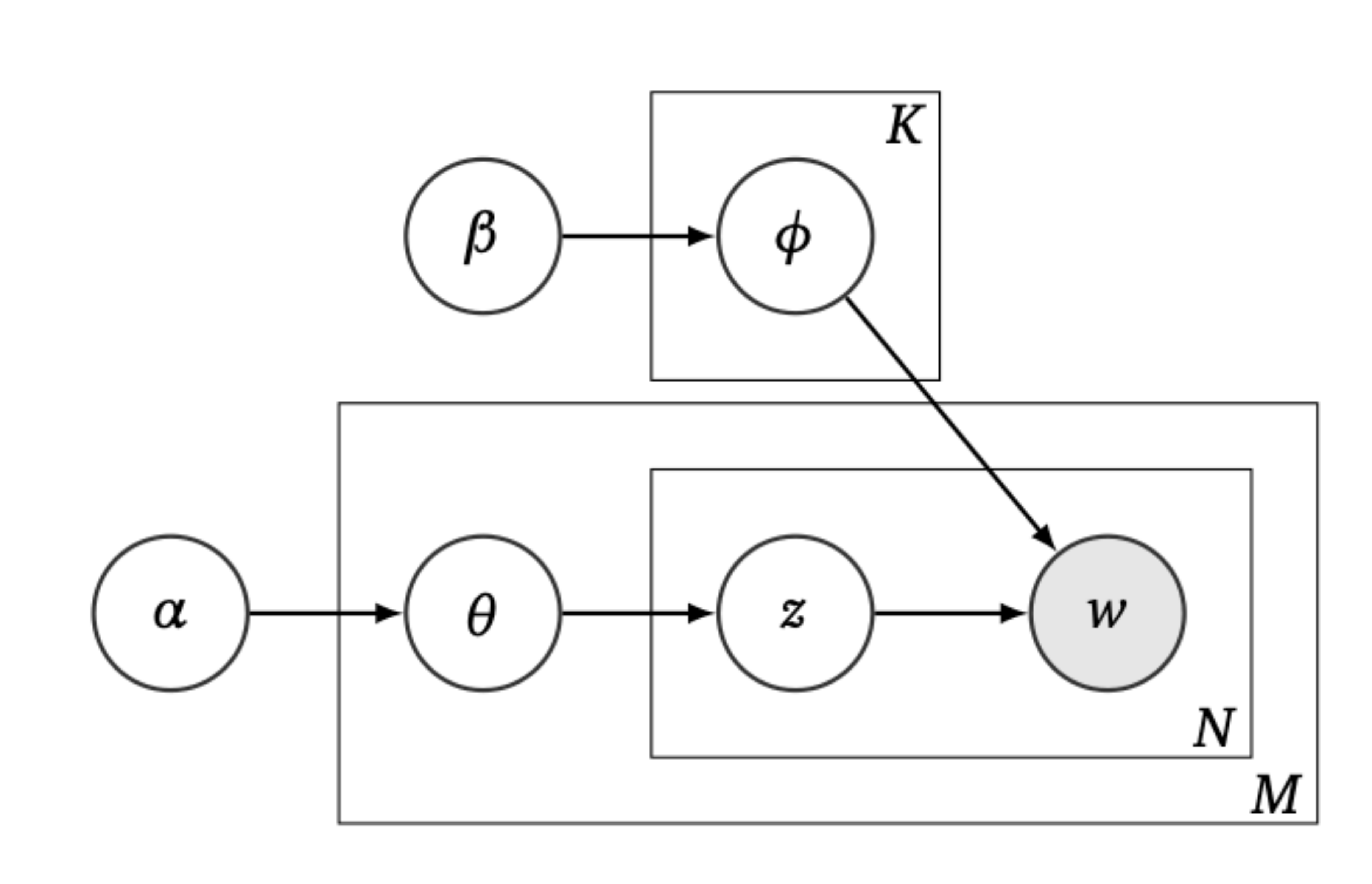


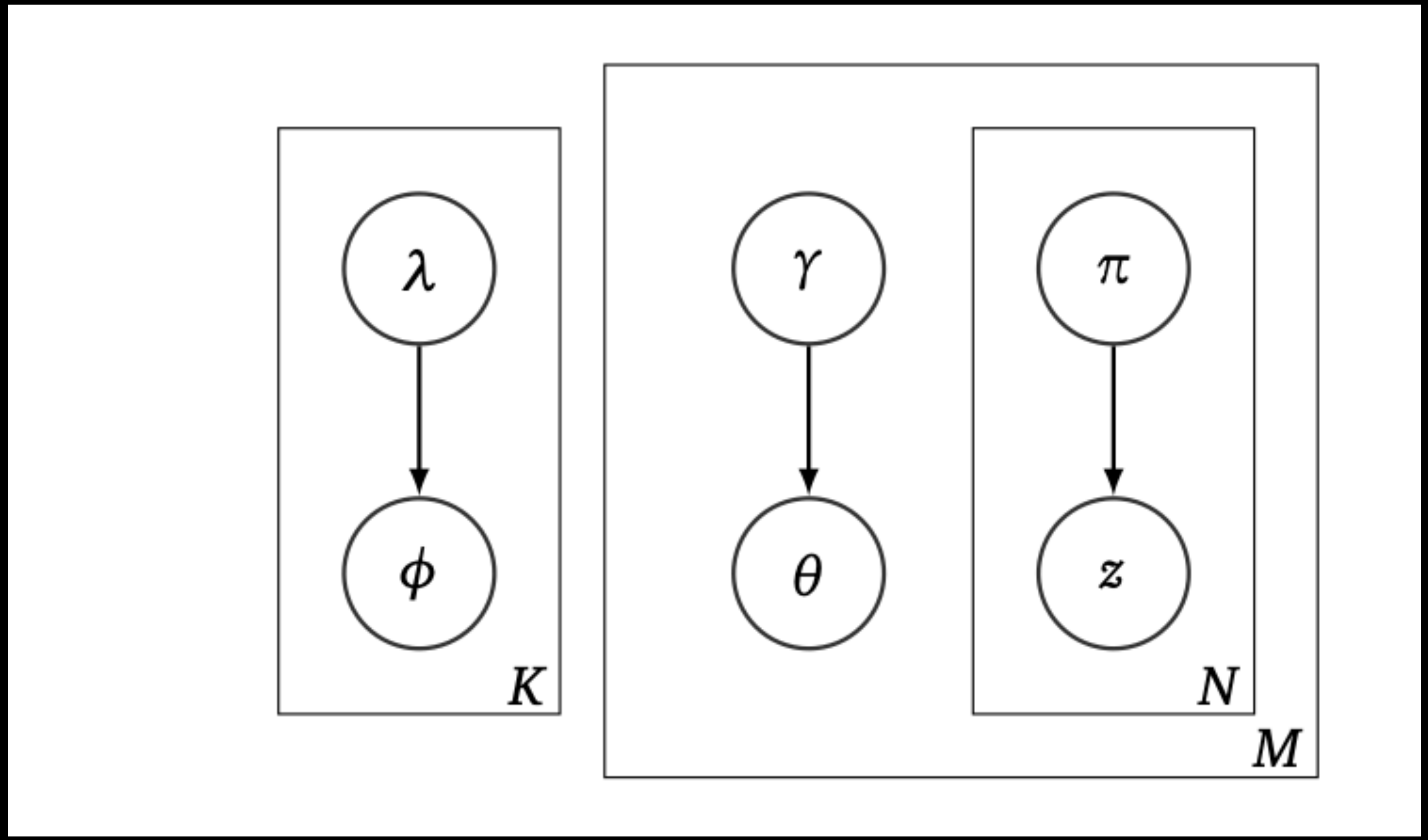
Figure 1: Graphical model for LDA.

- $$P(W, Z, \theta, \phi \mid \alpha, \beta) = \prod_{i=1}^{K} P(\phi_k \mid \beta) \prod_{j=1}^{M} P(\theta_j \mid \alpha) \prod_{t=1}^{N_j} P(z_{j,t} \mid \theta_j) P(w_{jt} \mid \phi_{z_{jt}})$$
- Target distribution:
- $$P(Z, \theta \mid \alpha, \beta, W) \propto \prod_{j=1}^{K} P(\theta_j \mid \alpha,) \prod_{t=1}^{N_j} P(z_{jt} \mid \theta_j) P(w_{jt} \mid \beta_{zt})$$

Variational Methods

- Approximate a distribution using an easy to use distribution. Typically fit minimizing Kullback-Leibler (KL) divergence between the variational distributions **q** and the true posteriors **p**


β
ϕ
K
α
θ
z
w
N
M


λ
ϕ
K
γ
θ
π
z
N
M

- Select a variational distribution q with parameters γ, π
- $P(Z, \theta \mid \alpha, \beta) \approx q(z, \theta \mid \gamma, \pi)$
- KL divergence of p to q is
- $D(q \mid \mid p) = \int_{\theta} \sum_{z} q(z_j, \theta_j \mid \gamma_j, \pi_j) log \left(\frac{q(z_j, \theta_j \mid \gamma_j, \pi_j)}{p(z_j, \theta_j \mid w_j, \alpha, \beta)} \right) d\theta$

- Variational Inference is done on each document (variational assumption of factorizability)