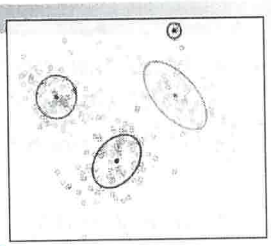
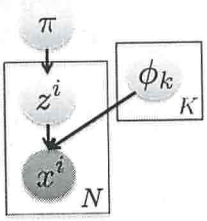
# Gaussian Mixture Model



- Most commonly used mixture model
- Observations: $x^1, \ldots, x^N$
- Parameters: $\theta = \{\pi, \phi\}$

$$\pi = [\pi_1, \ldots, \pi_K]$$ mixture weights

$$\phi = \{\phi_k\} = \{\mu_k, \Sigma_k\}$$ params for each cluster $k$




- Likelihood:

$$p(x^i \mid \theta) = \sum_k \pi_k p(x^i \mid \phi_k)$$

- Ex. $z^i$ = country of origin, $x^i$ = height of i[th] person
  - $k^{th}$ mixture component = distribution of heights in country $k$





# Motivates EM Algorithm

attempting to compute $\hat{\theta}^{ML}$

- Initial guess: $\hat{\theta}^{(0)}$
- Estimate at iteration $t$: $\hat{\theta}^{(t)}$

e.g. $y^i = \begin{bmatrix} x^i \\ z^i \end{bmatrix}$

- **E-Step**

  Compute $U(\theta, \hat{\theta}^{(t)}) = E[\log p(y \mid \theta) \mid x, \hat{\theta}^{(t)}]$

- **M-Step**

  Compute $\hat{\theta}^{(t+1)} = \arg\max_\theta U(\theta, \hat{\theta}^{(t)})$

→ $\hat{\theta}$ converges to local mode

# MAP Estimation

- Bayesian approach:
  - Place **prior** $p(\theta)$ on parameters
  - Infer **posterior** $p(\theta \mid x)$ $= \frac{p(x|\theta)\,p(\theta)}{p(x)}$
- Many, many, many motivations and implications
  - For the sake of this class, simplest motivation is to think of this as akin to regularization

$$\hat{\theta}^{MAP} = \arg\max_{\theta} \log p(\theta \mid x) = \arg\max_{\theta} \underbrace{\log p(x|\theta)}_{\text{ML term}} + \log p(\theta)$$ ← reg.

  - Saw importance of regularization in logistic regression (ML estimate can overfit data and lead to poor generalization)



# EM Algorithm – MAP Case

- Re-derive EM algorithm for $p(\theta \mid x)$
- Add $\log p(\theta)$ to $U(\theta, \hat{\theta}^{(t)})$ ← prev $E[\log p(y|\theta) \mid x, \hat{\theta}]$; now $E[\log p(y|\theta), x, \hat{\theta}] + \log p(\theta)$
  - What must be computed in E-Step remains unchanged because this term does not depend on *y*.
  - M-Step becomes:

$$\hat{\theta}^{(t+1)} = \arg\max_{\theta} U(\theta, \hat{\theta}^{(t)}) + \log p(\theta)$$

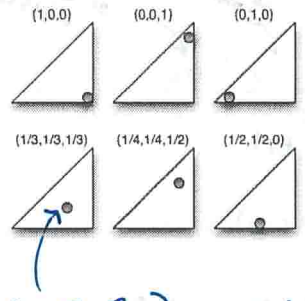
affects max w.r.t. $\theta$

# MAP EM Example – MoG

$p(\theta|x)$ in same family as $p(\theta)$

- For mixture of Gaussians, conjugate priors are:

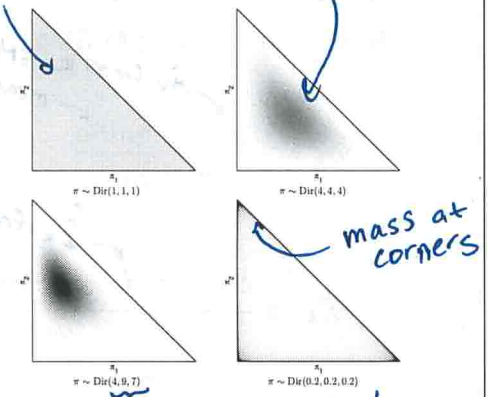
$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$$




$(\pi_1, \pi_2, \pi_3)$ is a pt on the simplex

uniform on simplex

$\mathrm{Dir}(\alpha_1, \alpha_2, \alpha_3)$ is prior on simplex




mass at corners

hyperparams $\alpha$ determine form





# MAP EM Example – MoG

- For mixture of Gaussians, conjugate priors are:

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$$

- Dirichlet posterior
  - Assume we condition on observations $z^i \sim \pi$
  - Count occurrences of $z^i = k$ : $n_k = |\{i : z^i = k\}|$
  - Then,

$$p(\pi \mid \alpha, z^1, \ldots, z^N) \propto \prod_i p(z^i \mid \pi)\, p(\pi \mid \alpha)$$

$$\propto \prod_k \Big(\prod_{i: z^i = k} \pi_k\Big) \cdot \pi_k^{\alpha_k - 1} \propto \prod_k \pi_k^{n_k + \alpha_k - 1}$$

  - Conjugacy: This posterior has same form as prior

# MAP EM Example – MoG

- For mixture of Gaussians, conjugate priors are:

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \quad \{\mu_k, \Sigma_k\} \sim \mathrm{NIW}(m_0, \kappa_0, \nu_0, S_0)$$

- Results in following M-Step:

$\hat{\bar{x}}_k$ from ML EM; mean of pseudo-obs; soft counts as in ML EM; pseudo-counts of obs. in cluster k

$$\hat{\mu}_k = \frac{r_k \bar{x}_k + \kappa_0 m_0}{r_k + \kappa_0} \qquad \hat{\pi}_k = \frac{r_k + \alpha_k - 1}{N + \sum_k \alpha_k - K}$$

$\hat{\Sigma}_k$ from ML EM

$$\hat{\Sigma}_k = \frac{S_0 + r_k S_k + \frac{\kappa_0 r_k}{\kappa_0 + r_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)'}{\nu_0 + r_k + d + 2}$$

↑ dimension





# Posterior Computations

- MAP EM focuses on point estimation:

$$\hat{\theta}^{MAP} = \arg\max_{\theta} p(\theta \mid x)$$

- What if we want a full characterization of the posterior?
  - Maintain a measure of uncertainty
  - Estimators other than posterior mode (different loss functions)
  - Predictive distributions for future observations

$p(x^{N+1} \mid x^1, \ldots, x^N)$ ← int. over uncertain model params

$= \int p(x^{N+1} \mid \theta)\, p(\theta \mid x^1, \ldots, x^N)\, d\theta$

- Often no closed-form characterization (e.g., mixture models)
- Alternatives:
  - Monte Carlo based estimates using samples from posterior
  - Variational approximations to posterior (more next time)

$\hat{F}(\theta) = \frac{1}{M} \sum_{i=1}^{M} f(\theta^{(i)})$

$\theta^{(i)} \sim \pi(\theta)$

# Gibb Sampling

- Want draws: $(\theta_1, \ldots, \theta_n) \sim \pi(\underline{\theta})$ ($n$ parameters or latent vars; can't sample directly)
- Construct Markov chain whose steady state distribution is $\pi(\underline{\theta})$
- Simplest case:

For $t = 1, \ldots,$ Niter

For $i = 1, \ldots, n$ (can use random ordering)

$$\theta_i^{(t)} \sim p(\theta_i \mid \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_n^{(t-1)})$$

Gibbs sampling assumes that this has a closed-form that we can sample from

# Example – Mixture of Gaussians

- Recall model
  - Observations: $x^1, \ldots, x^N$
  - Cluster indicators: $z^1, \ldots, z^N$
  - Parameters: $\theta = \{\pi, \phi\}$ $\quad \pi = [\pi_1, \ldots, \pi_K]$ $\quad \phi = \{\phi_k\} = \{\mu_k, \Sigma_k\}$
  - Generative model:

Priors:

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \qquad z^i \sim \pi \qquad i = 1, \ldots, N$$

$$\{\mu_k, \Sigma_k\} \sim F(\phi) \quad \text{e.g. NIW} \qquad x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i})$$

- Want to draw posterior samples of model parameters

$$\pi^{(t)} \sim p(\pi \mid \phi^{(t-1)}, x^1, \ldots, x^N)$$
$$\phi^{(t)} \sim p(\phi \mid \pi^{(t)}, x^1, \ldots, x^N)$$

NO CLOSED FORM

only cond. on obs.





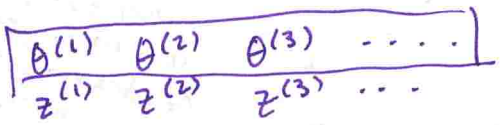
# Auxiliary Variable Samplers

- Augment variables of interest $\theta$ with variables $z$ to allow closed-form for sampling, just like in EM

Ex. Assume just one var of interest $\theta$

Sample $\theta^{(t)} \sim p(\theta \mid z^{(t-1)})$ ← each has

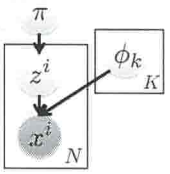
$z^{(t)} \sim p(z \mid \theta^{(t)})$ ← closed form




- In both cases, simply looking at subchain $\{\theta^{(t)}\}$ converges to draws from marginal distribution $\pi(\theta)$

# Example – Mixture of Gaussians

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \qquad z^i \sim \pi$$
$$\{\mu_k, \Sigma_k\} \sim F(\lambda) \qquad x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i})$$




- Try auxiliary variable sampler
  - Introduce cluster indicators into sampler

$\pi^{(t)} \sim p(\pi \mid z_{1:N}^{(t-1)}, \cancel{\phi}, \cancel{x_{1:N}}, \alpha) = \mathrm{Dir}(n_1 + \alpha_1, \ldots, n_K + \alpha_K)$

$n_k = |\{z^i : z^i = k\}|$

For $k = 1, \ldots, K$

$\{\mu_k^{(t)}, \Sigma_k^{(t)}\} \sim p(\phi_k \mid z_{1:N}^{(t-1)}, \cancel{\pi^{(t)}}, x_{1:N}, \lambda)$

what if high dim?

$= p(\phi_k \mid \{x^i\}_{z^{i(t-1)} = k}, \lambda) = \mathrm{NIW}(\text{post})$

For $i = 1, \ldots, N$

$z^{i(t)} \sim p(z^i \mid x^i, \pi^{(t)}, \phi^{(t)}) \propto \pi_{z^i}^{(t)} N(x^i \mid \phi_{z^i}^{(t)})$ as in EM



# Example – Clustering Results I

one random init.

another

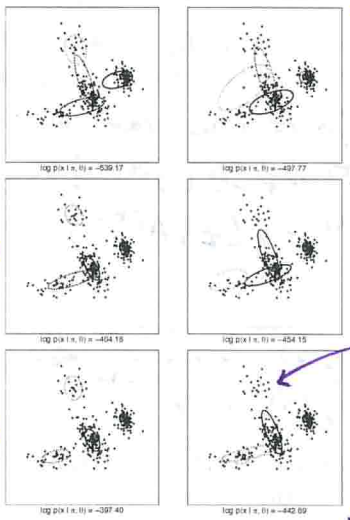



Figure courtesy of Erik Sudderth

mixing rate = time to explore support of posterior

can be SLOW

burn-in = time to stationary dist.

Can be SLOW

given $z^1, \ldots, z^N$ low post. prob of drawing $\phi_k$ here

will eventually happen maybe not in lifetime

# Collapsed Gibbs Samplers

- Marginalize a set of latent variables or parameters
  - Sometimes marginalized variables are nuisance parameters ← what you don't care about
  - Other times what gets marginalized are the variables
    - Make post-facto inferences on variables of interest based on sampled variables

$\theta$ = param of interest, but high dim

$z$ = enables sampling

What about just sampling $z$'s? Feasible?

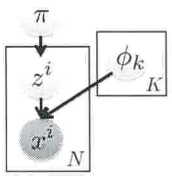
If so, can for $\hat{\theta}^{(t)}$ from sampled $z_{1:N}^{(t)}$

- Can improve efficiency if marginalized variables are high-dim
  - Reduced dimension of search space
  - But, often introduces dependences!




# Example – Collapsed MoG Sampling

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \qquad z^i \sim \pi$$
$$\{\mu_k, \Sigma_k\} \sim F(\phi) \qquad x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i})$$

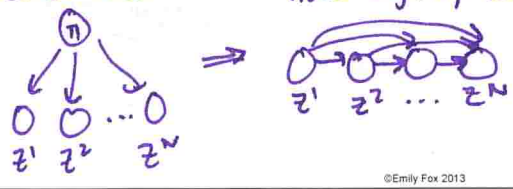



- Collapsed sampler

For $i = 1, \ldots, N$

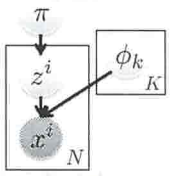
$$z^{i\,(t)} \sim p(z^i \mid \underbrace{z^{1\,(t)}, \ldots, z^{i-1\,(t)}, z^{i+1\,(t)}, \ldots, z^{N\,(t)}}_{z_{\setminus i}^{(t)}}, x_{1:N}, \alpha, \lambda)$$

cond. ind. ⇒ now tightly coupled



# Example – Collapsed MoG Sampling

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \qquad z^i \sim \pi$$
$$\{\mu_k, \Sigma_k\} \sim F(\phi) \qquad x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i})$$




- Derivation

$p(z^i \mid z_{\setminus i}, X_{1:N}, \alpha, \lambda) \propto p(z^i \mid z_{\setminus i}, \alpha)\, P(x^i \mid z^i, z_{\setminus i}, X_{\setminus i}, \lambda)$

$p(z^i = k \mid z_{\setminus i}, \alpha) = \int p(z^i = k \mid \pi)\, p(\pi \mid z_{\setminus i})\, d\pi = \dfrac{n_k^{\setminus i} + \alpha_k}{N - 1 + \sum \alpha_k}$ (counts not including $z^i = k$)

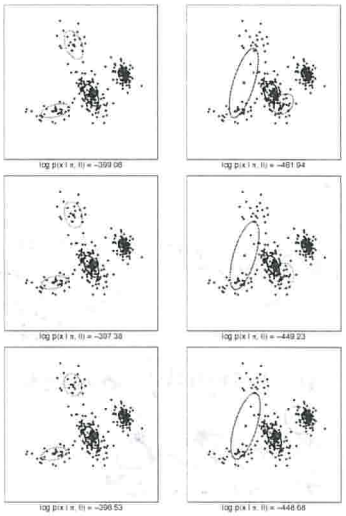
$p(x^i \mid z_{1:N}, X_{\setminus i}, \lambda) \propto \int \prod_{i: z^i = k} p(x^i \mid \phi_k)\, p(\phi_k \mid \lambda)\, d\phi_k$ = student-t pred. like.

- Important facts:

$$p(z_{1:N} \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(n_k + \alpha_k)}{\Gamma(\sum_k n_k + \alpha_k)} \qquad \frac{\Gamma(m+1)}{\Gamma(m)} = m$$



# Example – Clustering Results II



← worst case is still bad

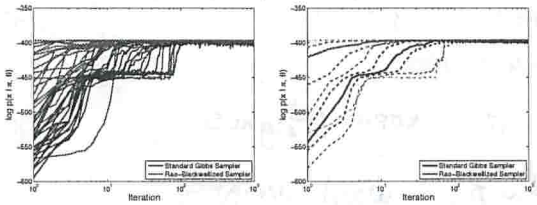
Seq. sampling $z^i$ can make it hard to make global assign. changes (strong dependencies)

Trade-off

Figure courtesy of Erik Sudderth

# Comparing Collapsed vs. Regular

**Log Likelihood vs. Gibbs Iteration**
(multiple chains)




Overall, in this case, collapsed has faster burn-in typically, but worst case still the same

Figure courtesy of Erik Sudderth



# Task 2: Cluster Documents

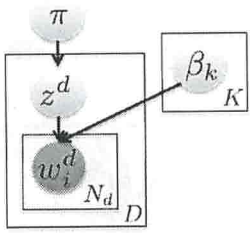
- **Previously:**
  - Cluster documents based on topic

sports


world news

# A Generative Model

- Documents: $x^1, \ldots, x^D$
- Associated topics: $z^1, \ldots, z^D$
- Parameters: $\theta = \{\pi, \beta\}$
- Generative model:




$z^d \sim \pi$ topic weights

$w_i^d \mid z^d \sim \beta_{z_d}$ word weights for topic $z^d$

Bayesian approach:

priors $\pi \sim Dir(\alpha_1, \ldots, \alpha_K)$

$\beta_k \sim Dir(\lambda_1, \ldots, \lambda_V)$ size of vocab

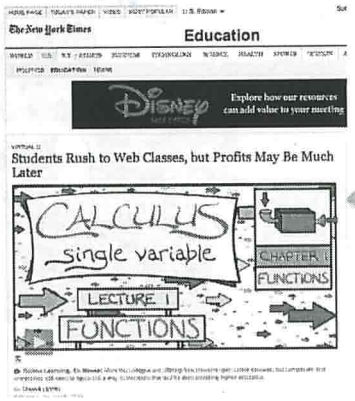
$k = 1, \ldots, K$





# Task 2: Cluster Documents

- **Now:** Document may belong to multiple clusters




EDUCATION

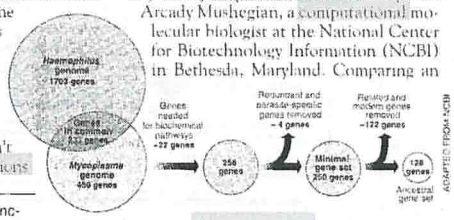
FINANCE

TECHNOLOGY

# Latent Dirichlet Allocation (LDA)

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



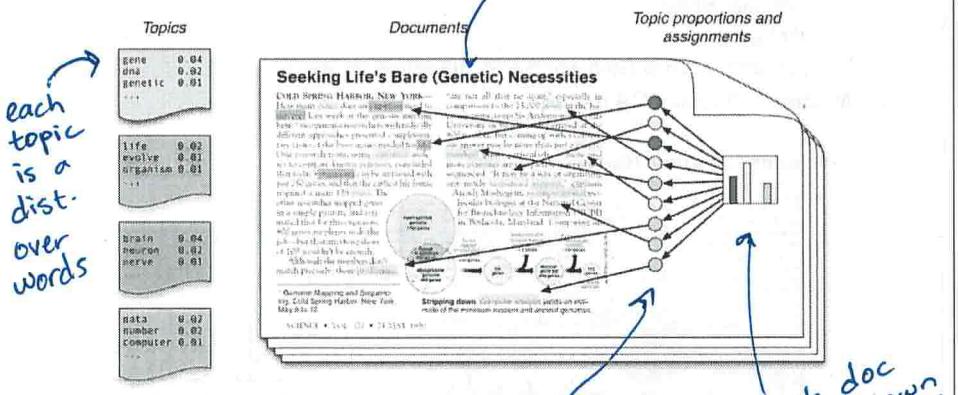

**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

mixed membership model

# Latent Dirichlet Allocation (LDA)

each doc is a mixture of corpus-wide topics

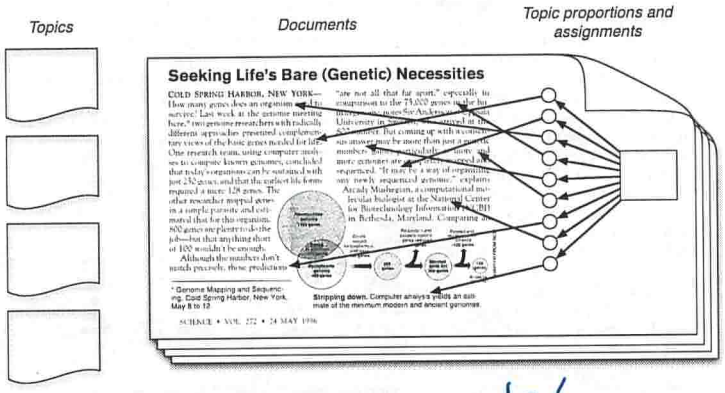



each topic is a dist. over words

every word is assigned to one of these topics

each doc has its own prevalence of topics

# Latent Dirichlet Allocation (LDA)

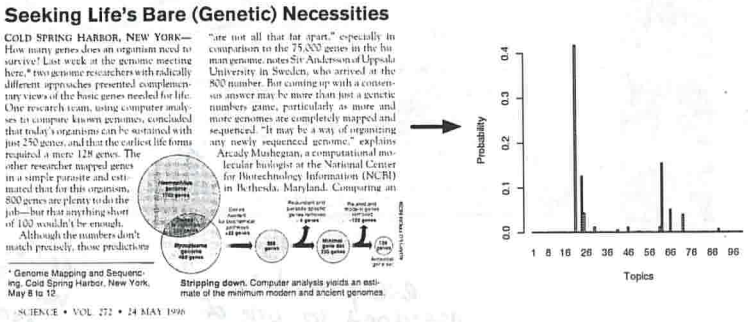



Only observe the words!

Want posterior p(topics, proportions, assignments | docs)

# Example Inference – Topic Weights

- **Data:** The OCR'ed collection of *Science* from 1990-2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model:** 100-topic LDA model



# Example Inference – Topic Words

highest prob. words in topic

| topic 1 | 2 | 3 | 4 | ... |
|---|---|---|---|---|
| human | evolution | disease | computer | |
| genome | evolutionary | host | models | |
| dna | species | bacteria | information | |
| genetic | organisms | diseases | data | |
| genes | life | resistance | computers | |
| sequence | origin | bacterial | system | |
| gene | biology | new | network | |
| molecular | groups | strains | systems | |
| sequencing | phylogenetic | control | model | |
| map | living | infectious | parallel | |
| information | diversity | malaria | methods | |
| genetics | group | parasite | networks | |
| mapping | new | parasites | software | |
| project | two | united | new | |
| sequences | common | tuberculosis | simulations | |





# LDA Generative Model

- Observations: $w_1^d, \ldots, w_{N_d}^d$ $d = 1, \ldots, D$
- Associated topics: $z_1^d, \ldots, z_{N_d}^d$ — corpus-wide topics that define word probabilities
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model: ↑ doc-specific topic weights

$z_i^d \sim \pi^d$

$w_i^d | z_i^d \sim \beta_{z_i^d}$

Priors:

$\pi^d \sim Dir(\alpha_1, \ldots, \alpha_K)$

$\beta_k \sim Dir(\lambda_1, \ldots, \lambda_V)$