CMPUT463/563
Probabilistic Graphical Models

Supervised Learning: Bayesian Networks

Lili Mou

Dept. Computing Science, University of Alberta

lmou@ualberta.ca

Outline

- Supervised learning: all variables are observed in training
- MLE of BN decomposes to every conditional probability
- For tabular parametrization
  - MLE: counting
  - MAP: add-($\alpha - 1$) smoothing with Dirichlet prior
  - Bayesian: add-α smoothing with Dirichlet prior
- Hierarchical Bayes
- Application: pLSI, LDA

Bayesian Networks

BN over variables $X_1, \cdots, X_N$ is a DAG with joint probability

$$p(X_1, \cdots, X_N) = \prod_{i=1}^{N} p(X_i \mid \mathrm{Par}(X_i))$$

Supervised training: all variables are observed in each training sample. $\mathscr{D} = \{(x_1^{(m)}, \cdots, X_N^{(m)})\}_{m=1}^{M}$ for M samples

Maximum likelihood estimation: Assuming parameters $\boldsymbol{\theta}$ are unknown constants; choosing such parameters that maximize the probability of data

$$\mathscr{L}(\boldsymbol{\theta}; \mathscr{D}) = P(\mathscr{D}; \boldsymbol{\theta})$$

Likelihood (of parameters) is just a fancy name of probability of data. It’s also common to say the likelihood of data.

MLE for BN

maximize $P(\mathscr{D};\boldsymbol{\theta}) \Leftrightarrow$ maximize $\log P(\mathscr{D};\boldsymbol{\theta})$

$$\Leftrightarrow \text{maximize} \log \prod_{m=1}^{M} \prod_{n=1}^{N} P(x_n^{(m)} \mid \mathrm{Par}(x_n^{(m)}))$$

$$\Leftrightarrow \text{maximize} \sum_{n=1}^{M} \underline{\left[\sum_{m=1}^{M} \log P(x_i^{(m)} \mid \mathrm{Par}(x_i^{(m)}; \boldsymbol{\theta}_i)) \right]}$$

The optimization decomposes for the parameters $\boldsymbol{\theta}_i$ of each variable X_i

In other words, the **training of a parameter** in BN only concerns the observations **involving the parameter**

Such decomposition works for any BN parametrization (discrete/continuous) and for weight sharing.

Categorial (multinomial) distribution:

Tabular BN

$$\hat{\pi}_k^{(\mathrm{MLE})} = \frac{N_k}{\sum_{k'} N_{k'}}$$

N_k is the count of kth category

- **Moment**
  - nth-order raw moment $\mathbb{E}[X^n]$
    - E.g., mean = 1st-order raw moment
  - nth-order central moment $\mathbb{E}[(X - \mathbb{E}[X])^n]$
    - E.g., variance = 2nd-order central moment
- **Moment matching**
  - Tune model parameters (esp. moment parameters) s.t. model moment matches empirical moment
  - E.g., π_k is the moment parameter of $\mathbb{E}[X_k]$ for $X_k = 1\{X = k\}$

    $\frac{N_k}{\sum_{k'} N_{k'}}$ is the empirical moment
- For categorical distribution (and a wider range, known as **exponential family**), **moment matching is equivalent to MLE**

Max a posteriori estimation

Bayesian interpretation of probability: everything unknown is a random variable.

- Parameters are treated as random variables with **Prior distribution** $P(\theta)$, the belief of θ before seeing data
- Under some parameter, the **likelihood** is still the probability of data $P(\mathscr{D} \mid \theta)$. Here, $P(\mathscr{D} \mid \theta)$ is computationally the same as $P(\mathscr{D}; \theta)$. When using "|", we treat θ as a random variable; when using ";", we treat θ as an unknown constants.
- **Posterior distribution:** $P(\theta \mid \mathscr{D})$, the belief of θ after seeing data
- **Max a posteriori (MAP) estimation** $\quad \operatorname{argmax}_\theta P(\theta \mid \mathscr{D})$

Prior Distribution of Categorical Distribution

- K-way categorical distribution is fully characterized by $\pi_k = \Pr[X = k]$ for $k = 1, \cdots, K$, such that

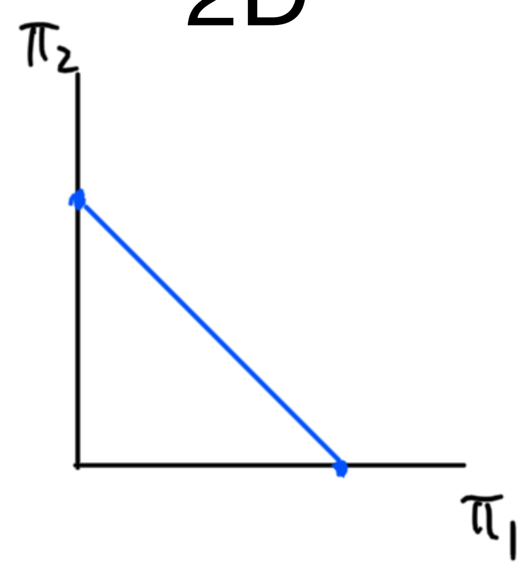
$$\pi_k \geq 0 \quad \text{and} \quad \sum_k \pi_k = 1$$

- Consider one-hot representation

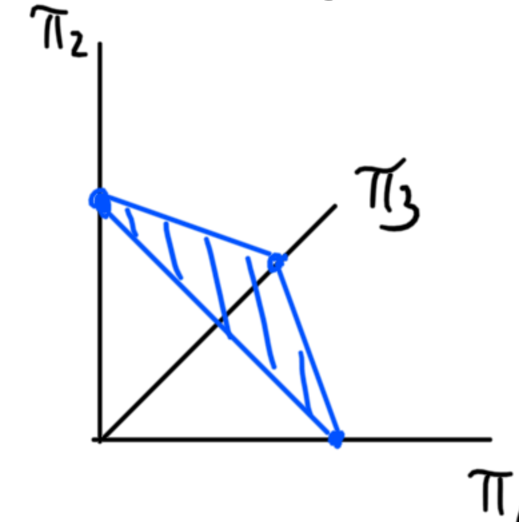
$$\boldsymbol{X} = (X_1, \cdots, X_K)$$

$$P(\boldsymbol{X}) = \prod_{k=1}^{K} \pi_k^{X_k}$$

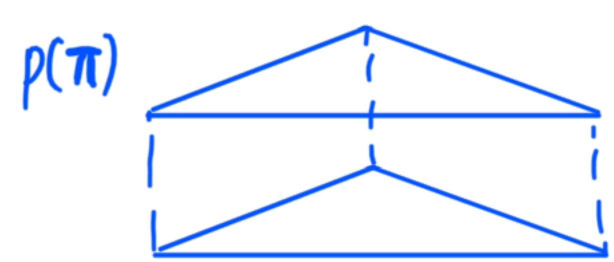
2D



3D



- A prior distribution is a distribution over the blue area (known as **simplex**). E.g., uniform over the simple



Dirichlet Distribution

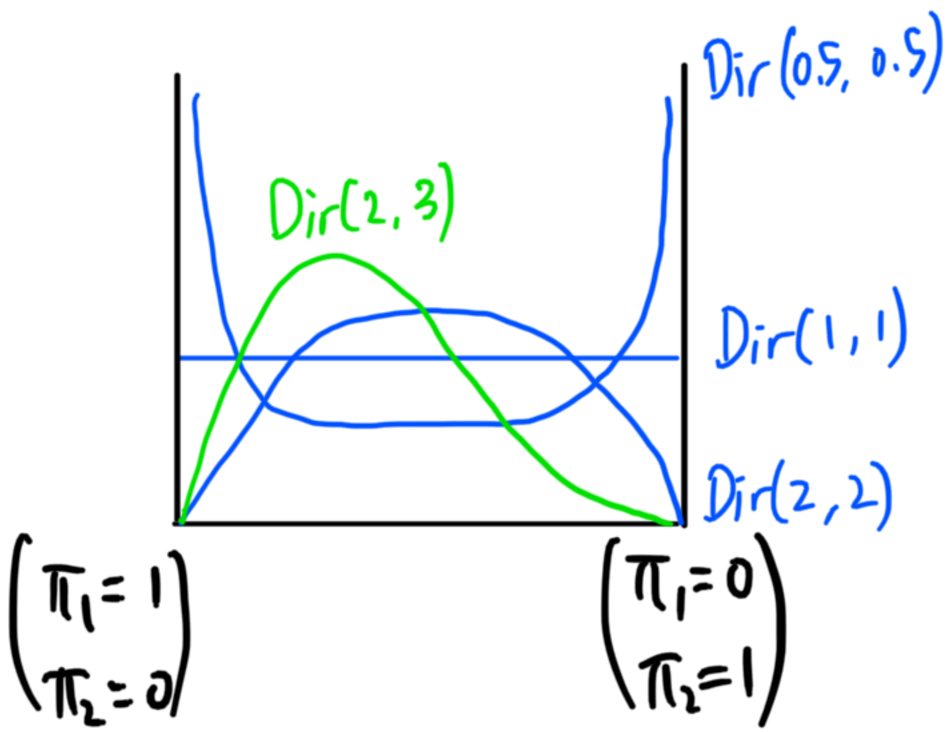
- Dirichlet distribution

$$P(\boldsymbol{\pi};\boldsymbol{\alpha}) = \underbrace{\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}}_{\text{Normalizing factor}} \underbrace{\prod_{k=1}^{K}\pi_k^{\alpha_k-1}}_{\text{Very analogous to Categorical distribution}}$$

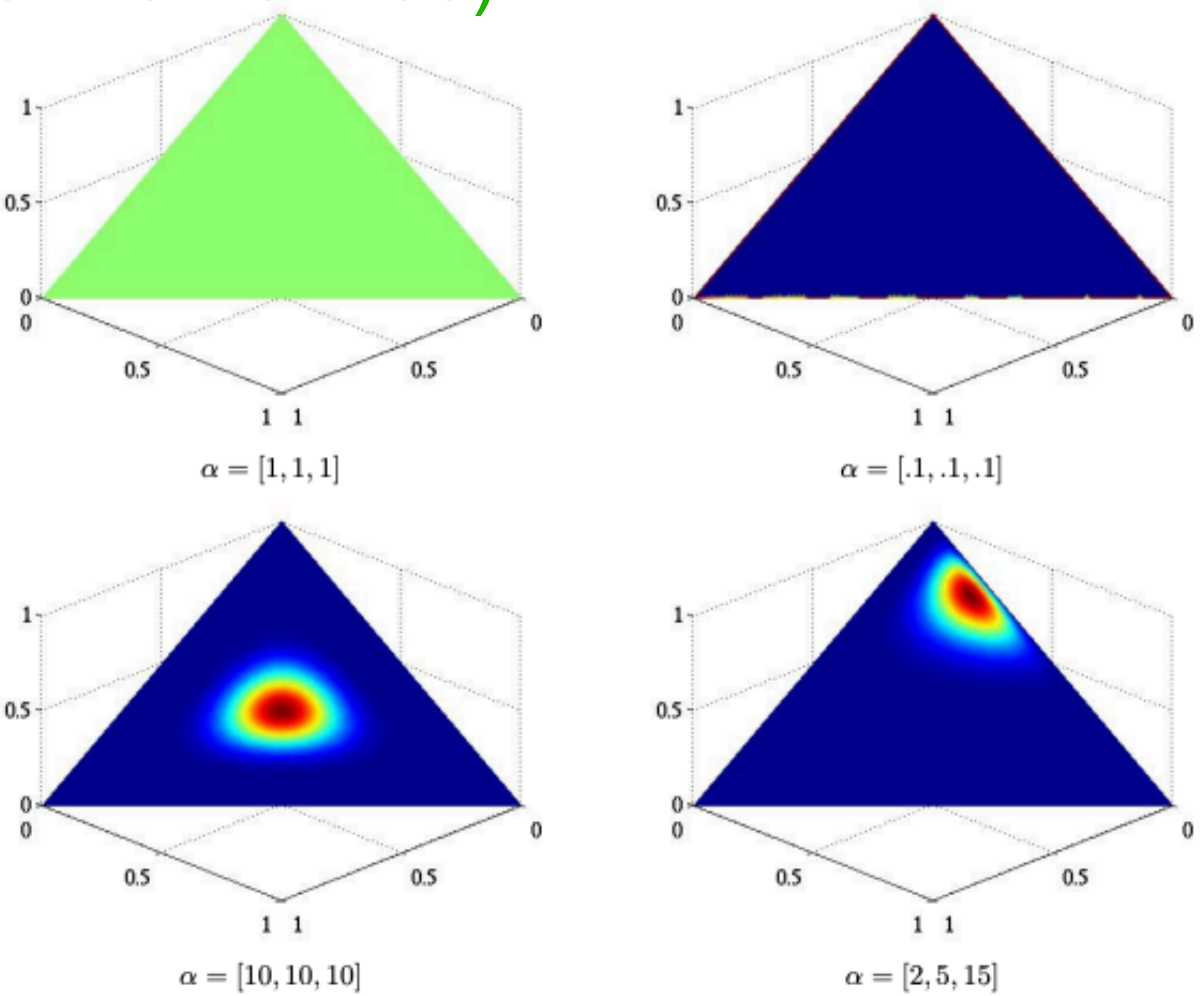
Very analogous to Categorical distribution

Normalizing factor
(Least important, need not be memorized)

2D



3D

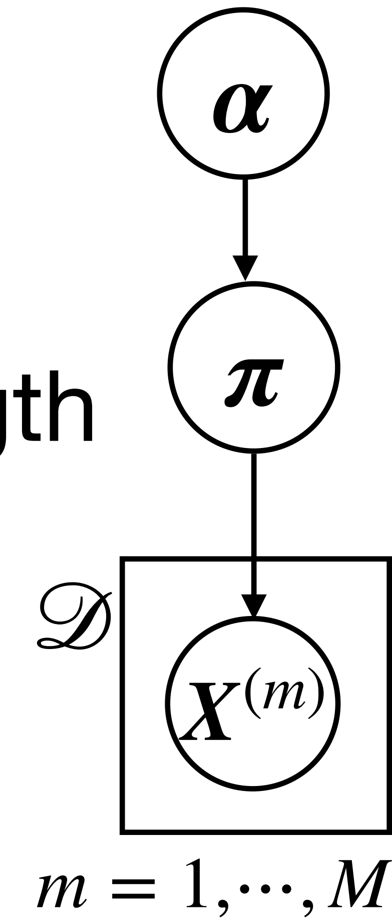


$\alpha = [1, 1, 1]$ $\alpha = [.1, .1, .1]$

$\alpha = [10, 10, 10]$ $\alpha = [2, 5, 15]$

Ref: http://mayagupta.org/publications/FrigyikKapilaGuptaIntroToDirichlet.pdf

Max a posterior inference



- Why Dirichlet? Equipped with key controls
  - Preference to certain categories; preference strength
- Computationally convenient

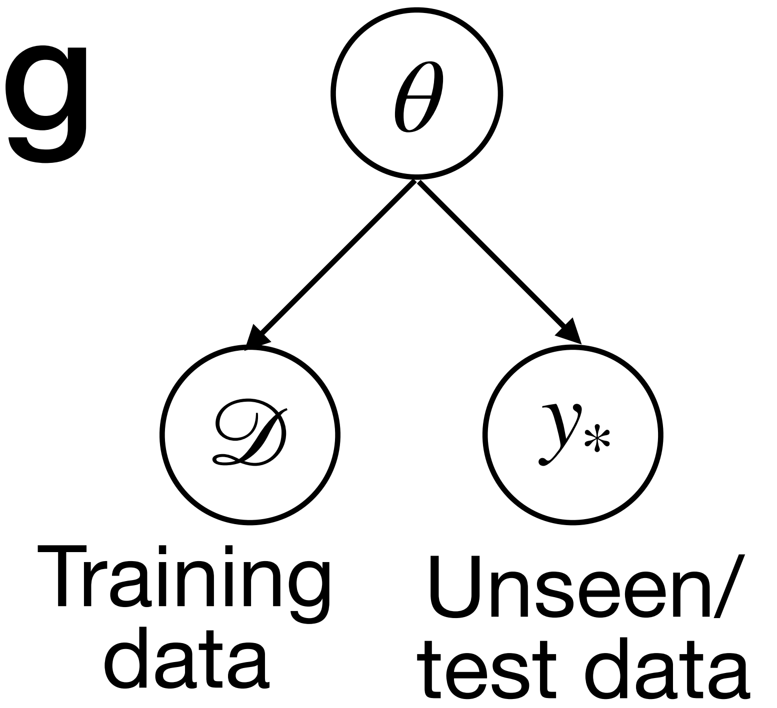
$$P(\boldsymbol{\pi} \mid \mathscr{D}) \propto P(\boldsymbol{\pi})P(\mathscr{D} \mid \boldsymbol{\pi}) \propto \pi_k^{\alpha_k - 1}\pi_k^{N_k} = \pi_k^{\alpha_k + N_k - 1}$$

N_k is the count of $X = k$ in the training data

- In other words, $\boldsymbol{\pi} \mid \mathscr{D} \sim \mathrm{Dir}(\alpha_1 + N_1, \cdots, \alpha_K + N_K)$
  - The posterior has the same form as the prior, known as **conjugate prior**.
  - It also has the same form as MLE. Thus, MAP w/ Dirichlet prior is **equivalent to add-$(\boldsymbol{\alpha} - 1)$ smoothing** for MLE

$$\hat{\pi}_k^{\mathrm{MAP}} = \mathrm{argmax}\, P(\boldsymbol{\pi} \mid \mathscr{D}) = \frac{N_k + \alpha_k - 1}{\sum_{k'}(N_{k'} + \alpha_{k'} - 1)}$$

Bayesian Learning



- Prior $P(\theta)$
- Likelihood $P(\mathscr{D} \mid \theta)$
- Posteriori $P(\theta \mid \mathscr{D}) \propto P(\theta)P(\mathscr{D} \mid \theta)$

- Recall: **Max a posteriori (MAP) inference**

$$\hat{\theta}^{(\mathrm{MAP})} = \mathrm{argmax}_\theta \, P(\theta \mid \mathscr{D}) \qquad \hat{y}_*^{(\mathrm{MAP})} = P(y \mid x_*, \hat{\theta}^{(\mathrm{MAP})})$$

- **Bayesian learning:**
  - No specific θ is chosen
  - Everything unrelated to the final prediction should be marginalized out

$$\hat{y}_*^{(\mathrm{Bayesian})} = \int P(y \mid x_*, \theta) P(\theta \mid \mathscr{D}) d\theta \propto \int P(y \mid x_*, \theta) P(\theta) P(\mathscr{D} \mid \theta) d\theta$$

Bayesian Learning for Dirichlet-Categorical

$$p(y^* \mid \mathcal{D}) = \int_\Delta p(y^* \mid \boldsymbol{\pi})\, p(\boldsymbol{\pi} \mid \mathcal{D})\, d\boldsymbol{\pi}$$

$$= \int_\Delta \pi_{y^*} \frac{\Gamma\left(\sum_{k=1}^K \alpha_k'\right)}{\prod_{k=1}^K \Gamma(\alpha_k')} \prod_{k=1}^K \pi_k^{\alpha_k' - 1}\, d\boldsymbol{\pi}$$

$$\left[\alpha_k' = \alpha_k + N_k, \quad \text{prior + count}\right]$$

$$= \frac{\Gamma\left(\sum_{k=1}^K \alpha_k'\right)}{\prod_{k=1}^K \Gamma(\alpha_k')} \int_\Delta \underbrace{\prod_{k=1}^K \pi_k^{\mathbb{1}\{y^*=k\} + \alpha_k' - 1}}_{\propto \mathrm{Dir}(\mathbb{1}\{y^*=k\} + \alpha_k' - 1)}\, d\boldsymbol{\pi}$$

$$\left[\text{Note: } \int \underbrace{\frac{\Gamma\left(\sum_{k=1}^K \mathbb{1}\{y^*=k\} + \alpha_k'\right)}{\prod_{k=1}^K \Gamma(\mathbb{1}\{y^*=k\} + \alpha_k')} \prod_{k=1}^K \pi_k^{\mathbb{1}\{y^*=k\} + \alpha_k' - 1}}_{\mathrm{Dir}(\mathbb{1}\{y^*=k\} + \alpha_{k'} - 1)}\, d\pi = 1\right]$$

$$= \frac{\Gamma\left(\sum_{k=1}^K \alpha_k'\right)}{\prod_{k=1}^K \Gamma(\alpha_k')} \cdot \frac{\prod_{k=1}^K \Gamma(\mathbb{1}\{y^*=k\} + \alpha_k')}{\Gamma\left(\sum_{k=1}^K (\underbrace{\mathbb{1}\{y^*=k\}}_{=1 \text{ for exactly one term}} + \alpha_k')\right)}$$

$$= \frac{\Gamma\left(\sum_{k=1}^K \alpha_k'\right)}{\prod_{k=1}^K \Gamma(\alpha_k')} \cdot \frac{\prod_{k=1}^K \Gamma(\mathbb{1}\{y^*=k\} + \alpha_k')}{\Gamma\left(1 + \sum_{k=1}^K \alpha_k'\right)}$$

$$\Gamma(x+1) = x\,\Gamma(x)$$

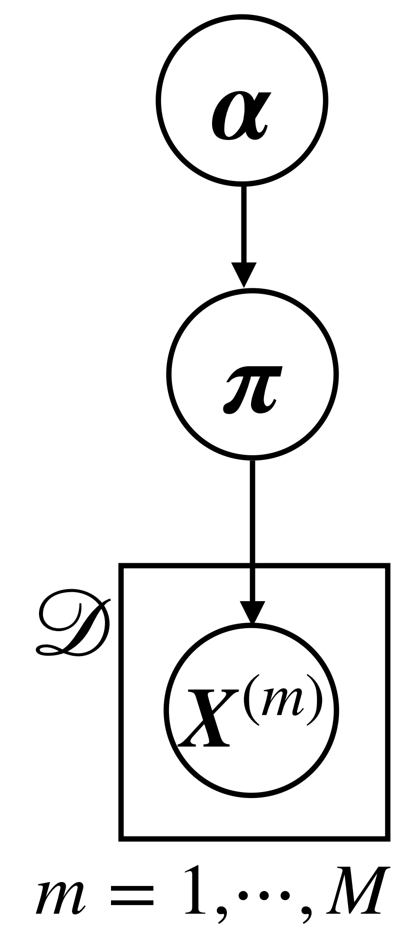
$$= \frac{\cancel{\Gamma\left(\sum_{k=1}^K \alpha_k'\right)}}{\cancel{\prod_{k=1}^K \Gamma(\alpha_k')}} \cdot \frac{\alpha_{y^*}' \cdot \cancel{\prod_{k=1}^K \Gamma(\alpha_k')}}{\sum_{k=1}^K \alpha_k' \;\cancel{\Gamma\left(\sum_{k=1}^K \alpha_k'\right)}}$$

$$= \frac{\alpha_{y^*}'}{\sum_{k=1}^K \alpha_k'} = \frac{\alpha_{y^*} + N_{y^*}}{\sum_{k=1}^K (\alpha_k + N_k)}$$

Hierarchical Bayes

- Sometimes the prior distribution has hyper-parameters Prior $P(\theta | \alpha)$

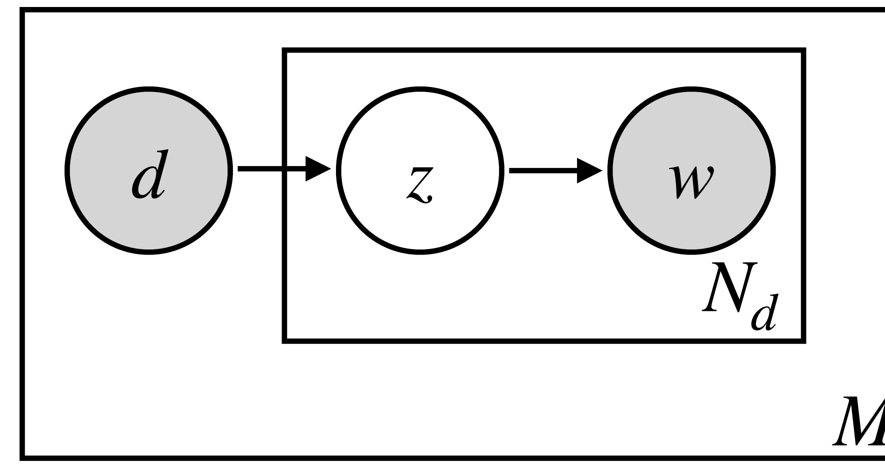
- How can we handle them?
  - Maximum likelihood estimation (Type-II MLE, empirical Bayes)
  - Max a posteriori estimation (Type-II MAP)
  - Full Bayesian



Probabilistic Latent Semantic Indexing (pLSA)

[Hofmann, SIGIR'99]

- Documents: $d \in \{d_1, \cdots, d_M\}$
- Latent topics: $z = \{z_1, \cdots, z_K\}$
  - Think of sport, politics, etc.
- Words $w = \{w_1, \cdots, w_{|V|}\}$



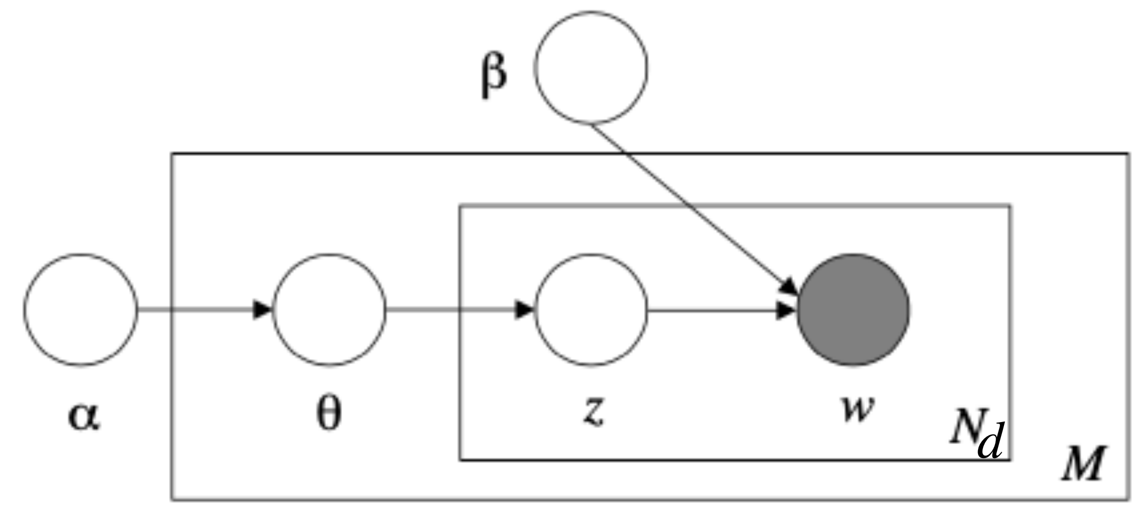
- Select a document d
- Pick a latent class z with probability $P(z \mid d)$
- Generate a word $P(w \mid z)$

Latent Dirichlet Allocation

[Blei, Ng, Jordan, JMLR'03]

- For each document of length N_d
  - Choose $\theta \sim Dir(\alpha)$
  - For each word w_n
    - $z_n \sim cat(\theta)$
    - $w_n \sim P(w_n | z_n, \beta)$

      where $\beta_{ij} = \Pr[w = j | z = i]$



- A more complicate variant