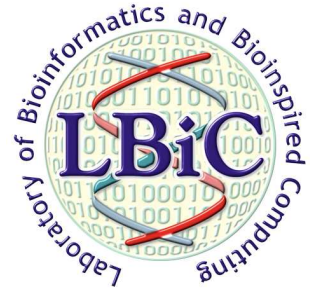
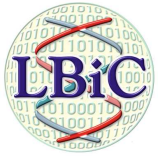


Statistics and Machine Learning

Part 2: Classification and Clustering Problems

Conrado Miranda
LBiC, FEEC, University of Campinas

Contents



Introduction

Previous part

- Lots of statistics properties;
- Regression.

Focus of this part

- How to interpret and build the problem and expand existing solutions;
- Not how to **solve** the problem for the general case, as this can be very, very hard.

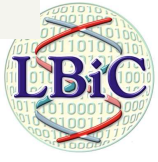


Plate Notation: Node Types

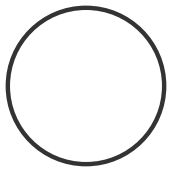
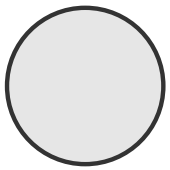
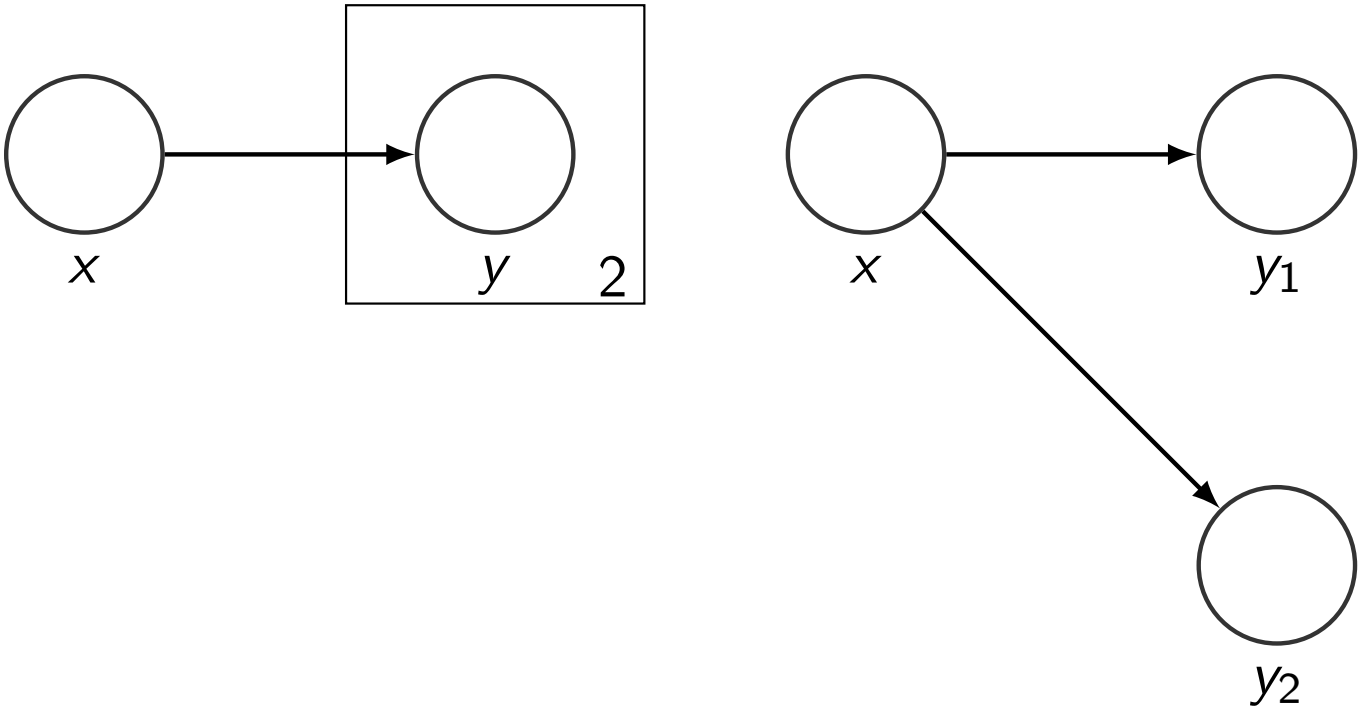
	Hidden	Known
Random Variables	 θ	 θ
Fixed Values	 θ	 θ

Plate Notation: Bayesian Network

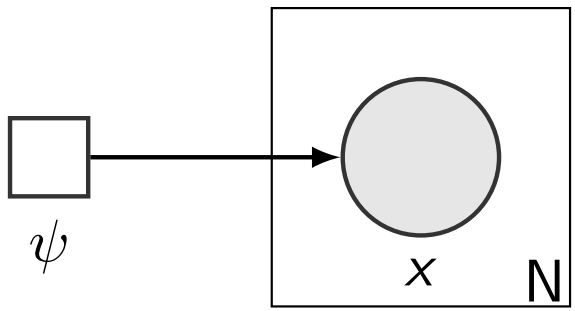


$$p_{1,2}(y_1, y_2|x) = p_1(y_1|x)p_2(y_2|x) \quad (1a)$$

$$y_1 \sim D_1(x) \quad (1b)$$

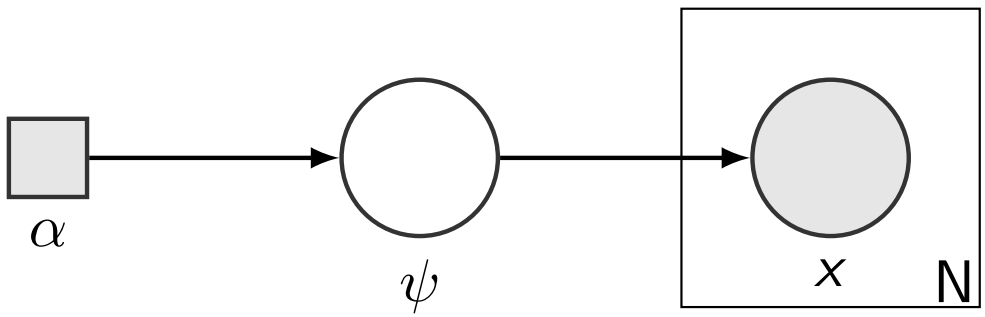
$$y_2 \sim D_2(x) \quad (1c)$$

Plate Notation: Dice Roll (non-Bayesian)



$$x_i \sim Cat(\psi) \tag{2}$$

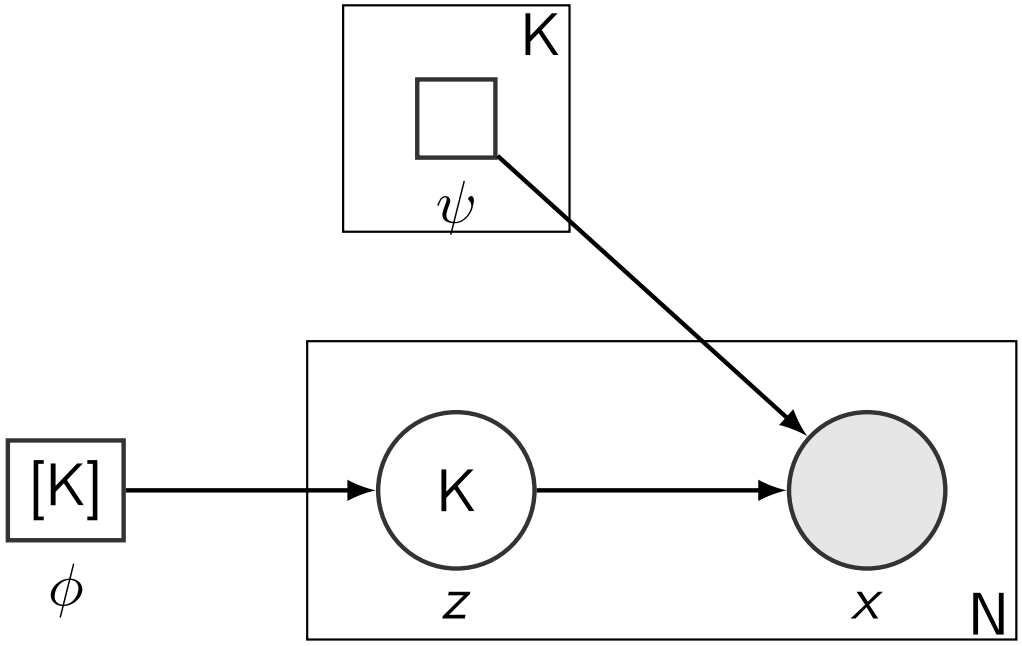
Plate Notation: Dice Roll (Bayesian)



$$\psi \sim Dir(\alpha) \tag{3a}$$

$$x_i \sim Cat(\psi) \tag{3b}$$

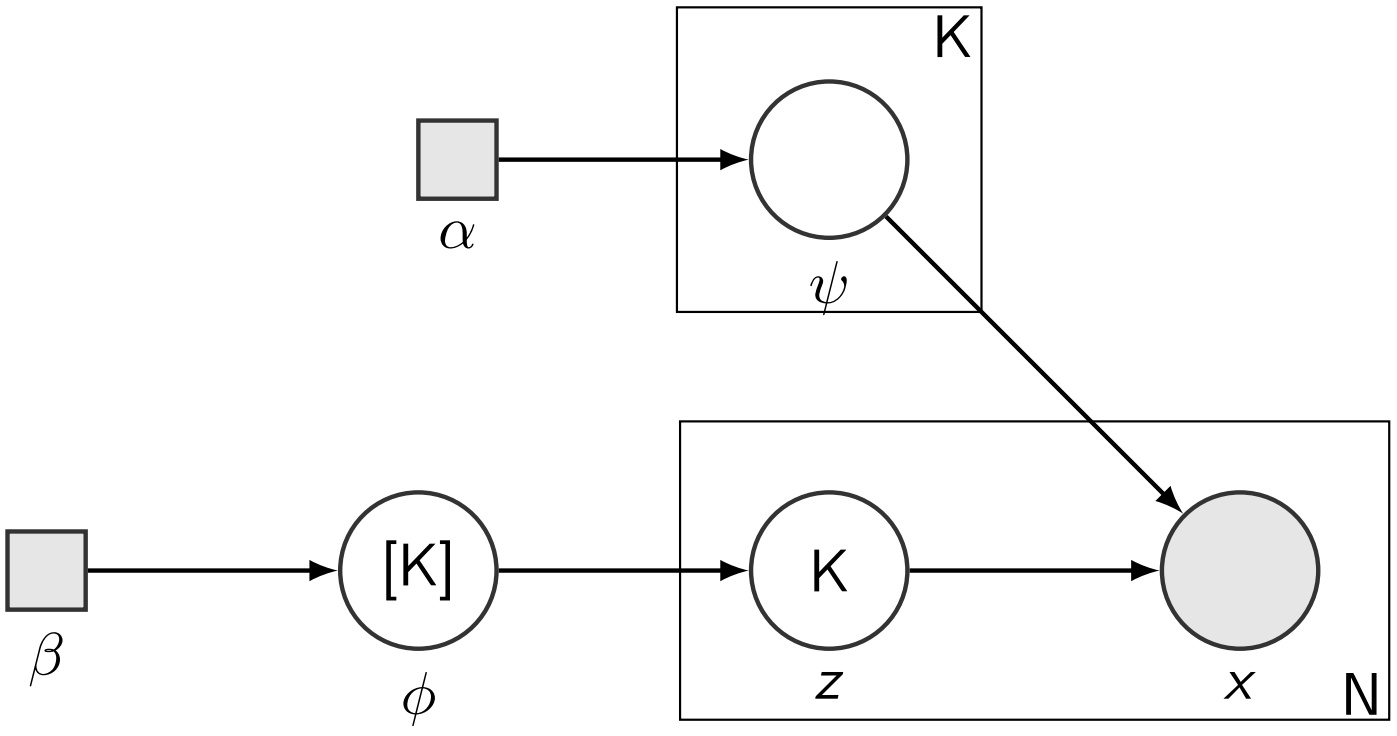
Plate Notation: Dice Roll with Multiple Dice (non-Bayesian)



$$z_i \sim Cat(\phi) \tag{4a}$$

$$x_i \sim Cat(\psi_{z_i}) \tag{4b}$$

Plate Notation: Dice Roll with Multiple Dice (Bayesian)



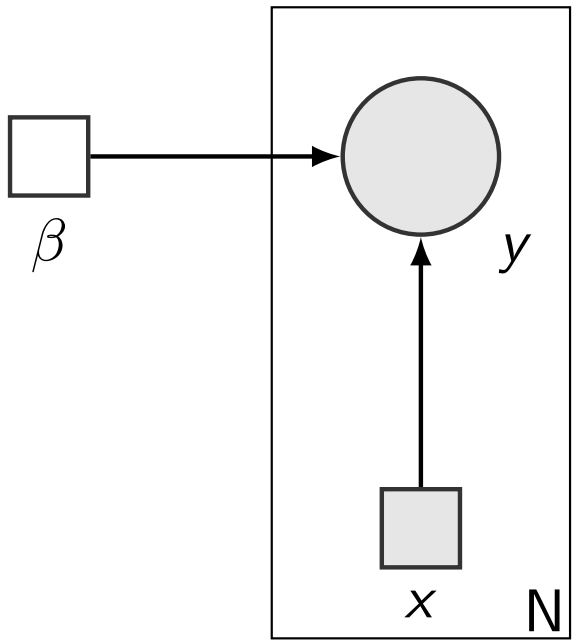
$$\psi_k \sim Dir(\alpha) \quad (5a)$$

$$\phi \sim Dir(\beta) \quad (5b)$$

$$z_i \sim Cat(\phi) \quad (5c)$$

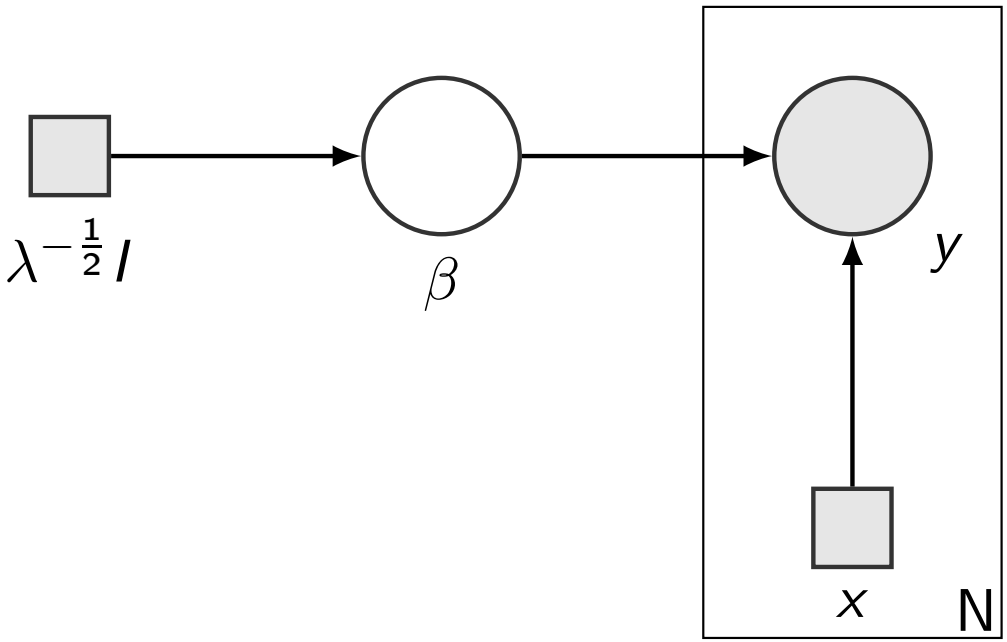
$$x_i \sim Cat(\psi_{z_i}) \quad (5d)$$

Plate Notation: Linear Regression



$$y_i \sim N(x_i\beta, I) \tag{6}$$

Plate Notation: Linear Ridge Regression



$$\beta \sim N(0, \lambda^{-\frac{1}{2}} I) \tag{7a}$$

$$y_i \sim N(x_i \beta, I) \tag{7b}$$

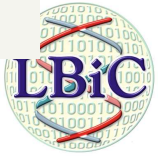
Logistic Regression: Binary Classification

Logistic Function

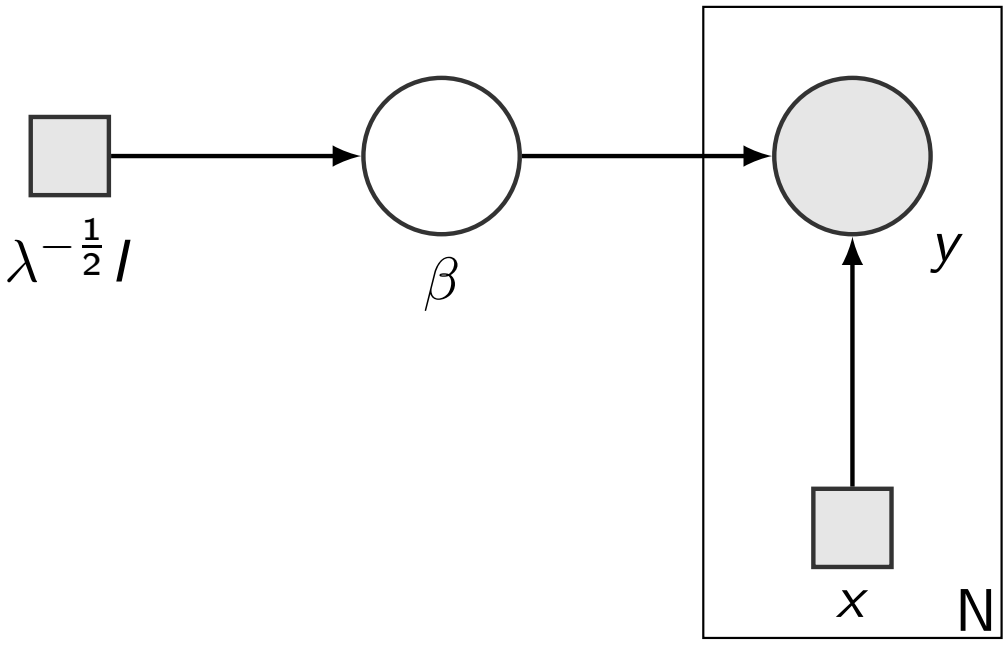
$$y = \sigma(z) = \frac{1}{1 + \exp(-z)} \tag{8}$$

Logistic Regression

$$\mathcal{L}(\beta|y_1, \ldots, y_N) = \prod_{i=1}^{N} p(y_i|\beta) = \prod_{i=1}^{N} \underbrace{\sigma(x_i\beta)}_{p_i}^{y_i} (1 - \sigma(x_i\beta))^{1-y_i} \tag{9}$$



Logistic Regression: Plate Model



$$\beta \sim N(0, \lambda^{-\frac{1}{2}} I) \tag{10a}$$

$$y_i \sim B(1, \sigma(x_i \beta)) \tag{10b}$$

Logistic Regression: Generalization for Multiple Classes

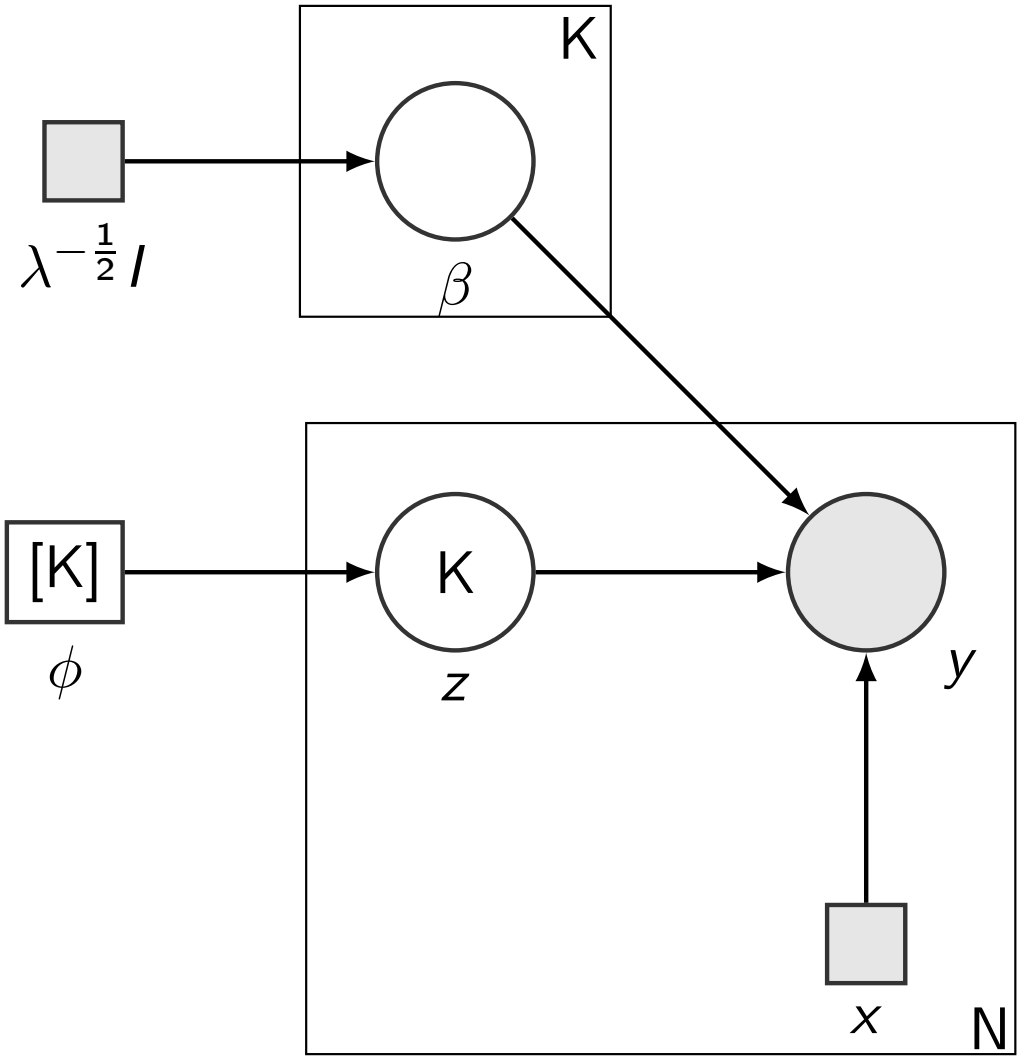
Softmax Function

$$y_k = \Sigma_k(x; \{\beta_j\}) = \frac{\exp(-x\beta_k)}{\sum_{j=1}^{K} \exp(-x\beta_j)} \tag{11}$$

Softmax Regression

$$\mathcal{L}(\beta_1, \ldots, \beta_K | y_1, \ldots, y_N) = \prod_{i=1}^{N} \prod_{k=1}^{K} \underbrace{\Sigma_k(y_i; \{\beta_j\})}_{p_{i,k}}^{y_{i,k}} \tag{12}$$

Logistic Regression: Plate Model for Softmax Regression

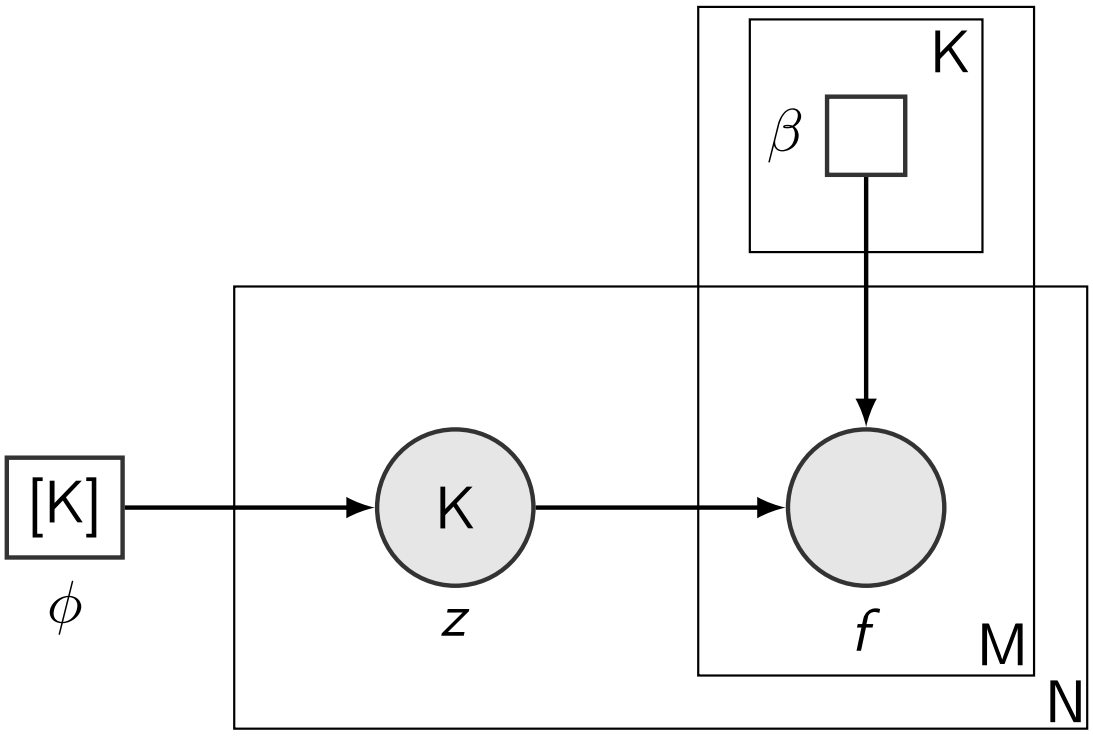


$$\beta \sim N(0, \lambda^{-\frac{1}{2}} I) \quad (13a)$$

$$z_i \sim Cat(\phi) \quad (13b)$$

$$y_i \sim Cat(\Sigma_{z_i}(x_i; \{\beta_j\})) \quad (13c)$$

Naive Bayes: Plate Model



$$z_i \sim Cat(\phi) \tag{14a}$$

$$f_{i,j} \sim D_j(\beta_{j,z_i}) \tag{14b}$$

Naive Bayes: Formulation

Definition

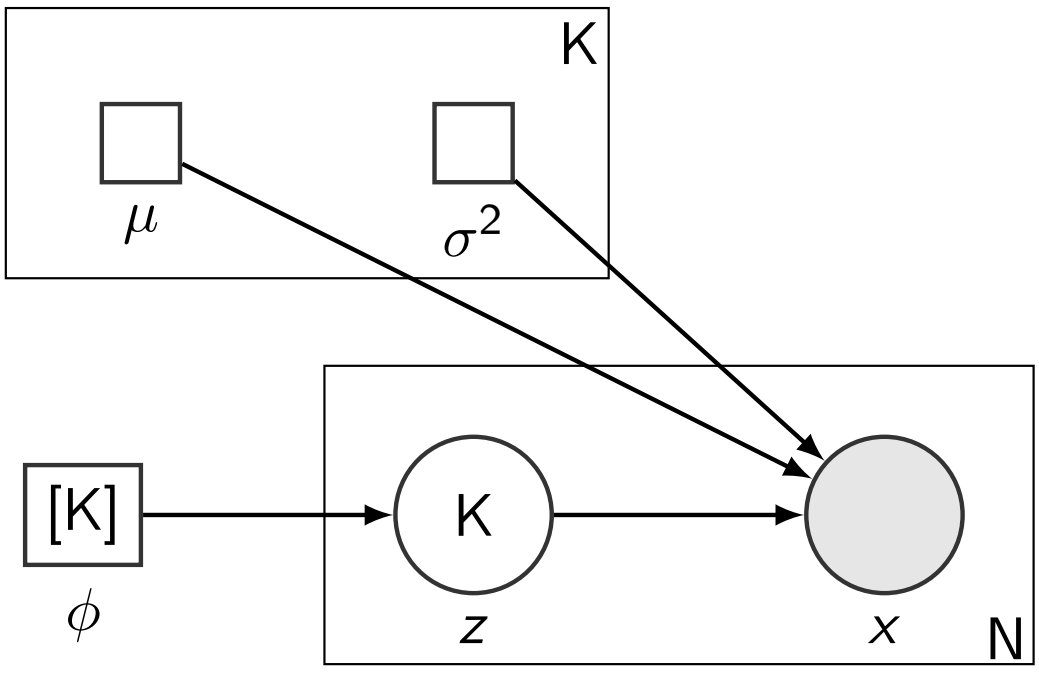
Given a set of features for a sample, find to which class it belongs. The features are built so that, given the class of the sample, the values of the features are independent.

Class Probability

$$p(k|f_1,\ldots,f_M) = \frac{p(k)p(f_1,\ldots,f_M|k)}{p(f_1,\ldots,f_M)} = \frac{p(k)\prod_{j=1}^{M} p(f_j|k)}{p(f_1,\ldots,f_M)} \quad (15a)$$

$$= \frac{p(k)\prod_{j=1}^{M} p(f_j|k)}{\sum_{l=1}^{K} p(l)\prod_{j=1}^{M} p(f_j|l)} \quad (15b)$$

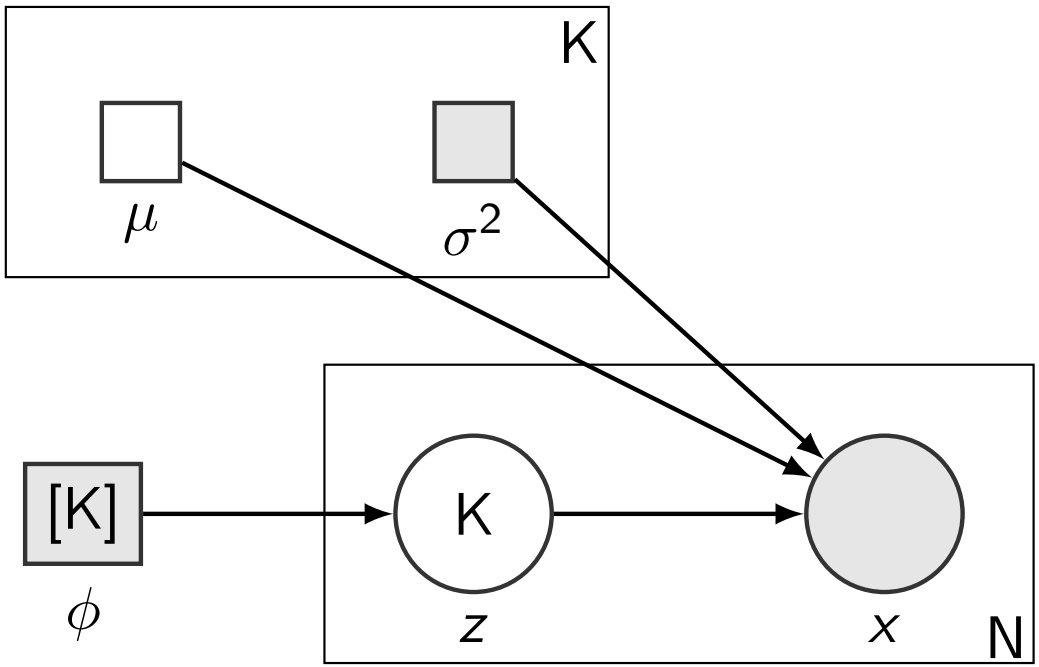
Mixture of Gaussians: Plate Model



$$z_i \sim Cat(\phi) \tag{16a}$$

$$x_i \sim N(\mu_{z_i}, \sigma^2_{z_i}) \tag{16b}$$

Mixture of Gaussians: K-Means



$$\phi_k = \frac{1}{K}, \quad \sigma_k \to 0 \tag{17}$$

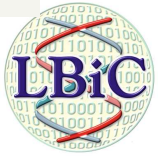
LDA: Problem Definition

Problem

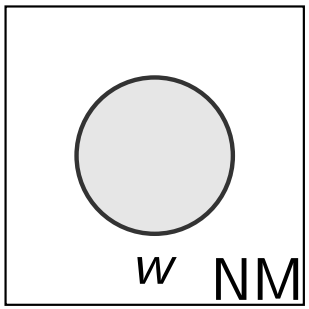
Given a set of documents, one desires to classify them according to the words that occur in each document. This classification doesn't have to be exclusive, with each document covering one or more topics.

Data provided

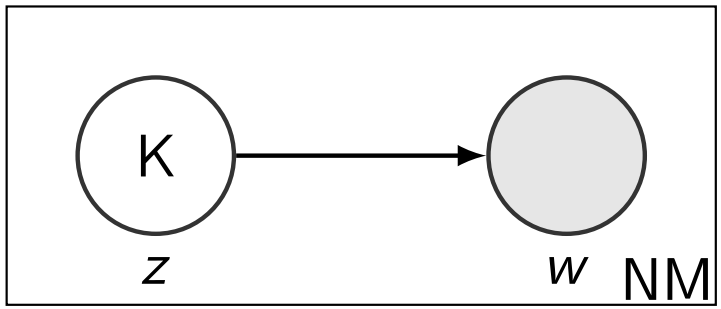
Let M be the number of documents provided, and N the number of words in each document. The documents must be classified in K different classes. There are a total of V different words in the documents.



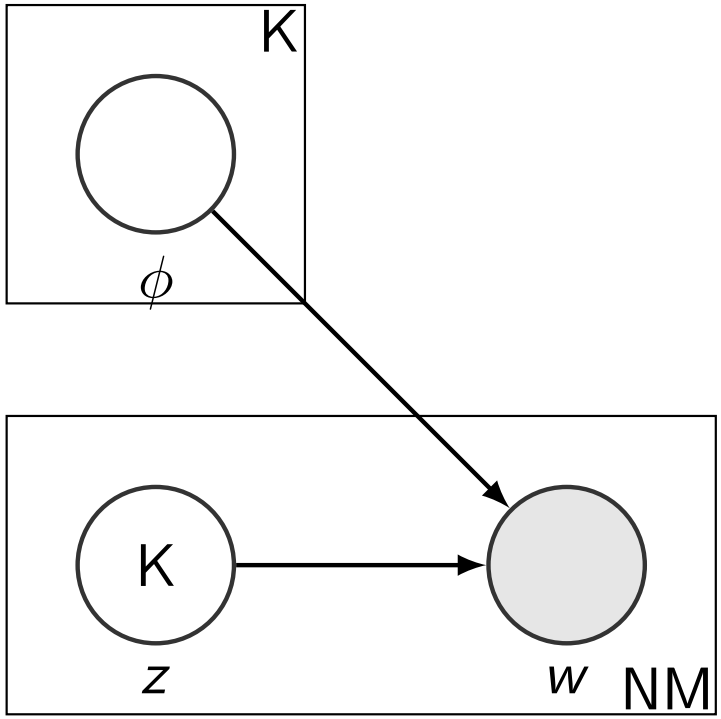
LDA: Building Model



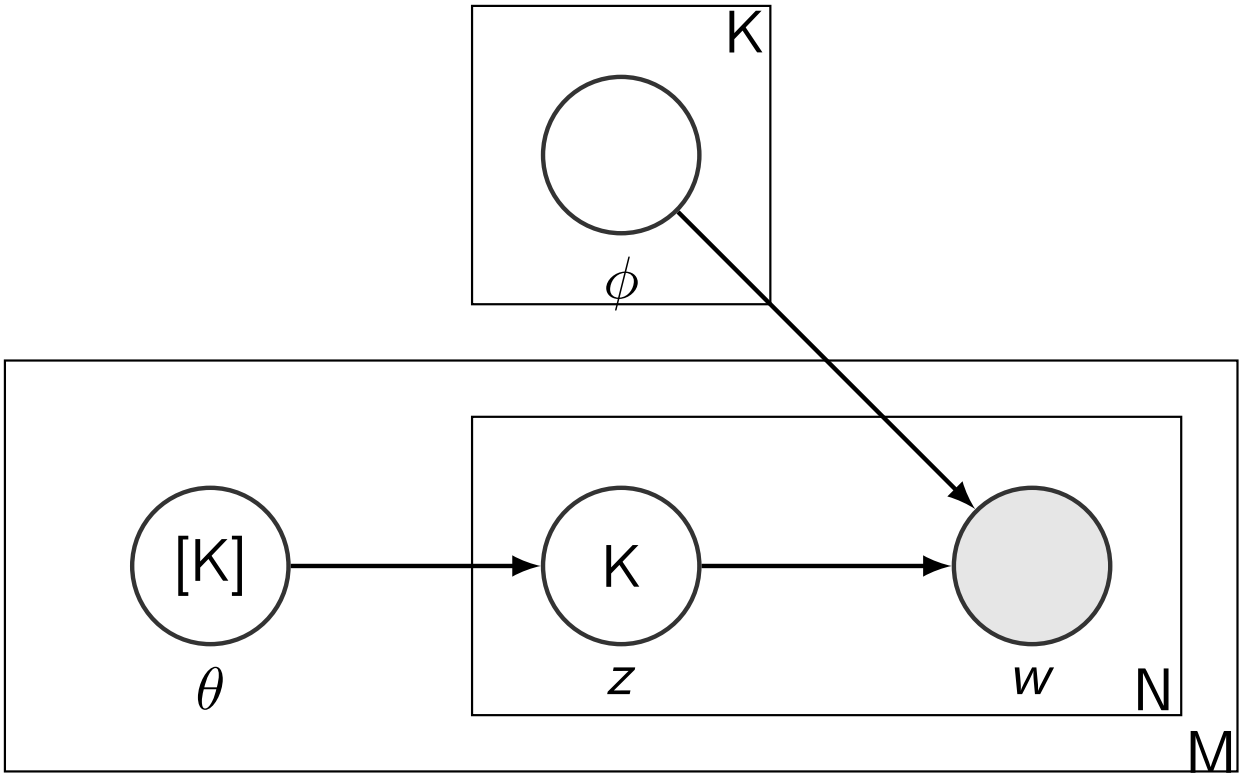
LDA: Building Model



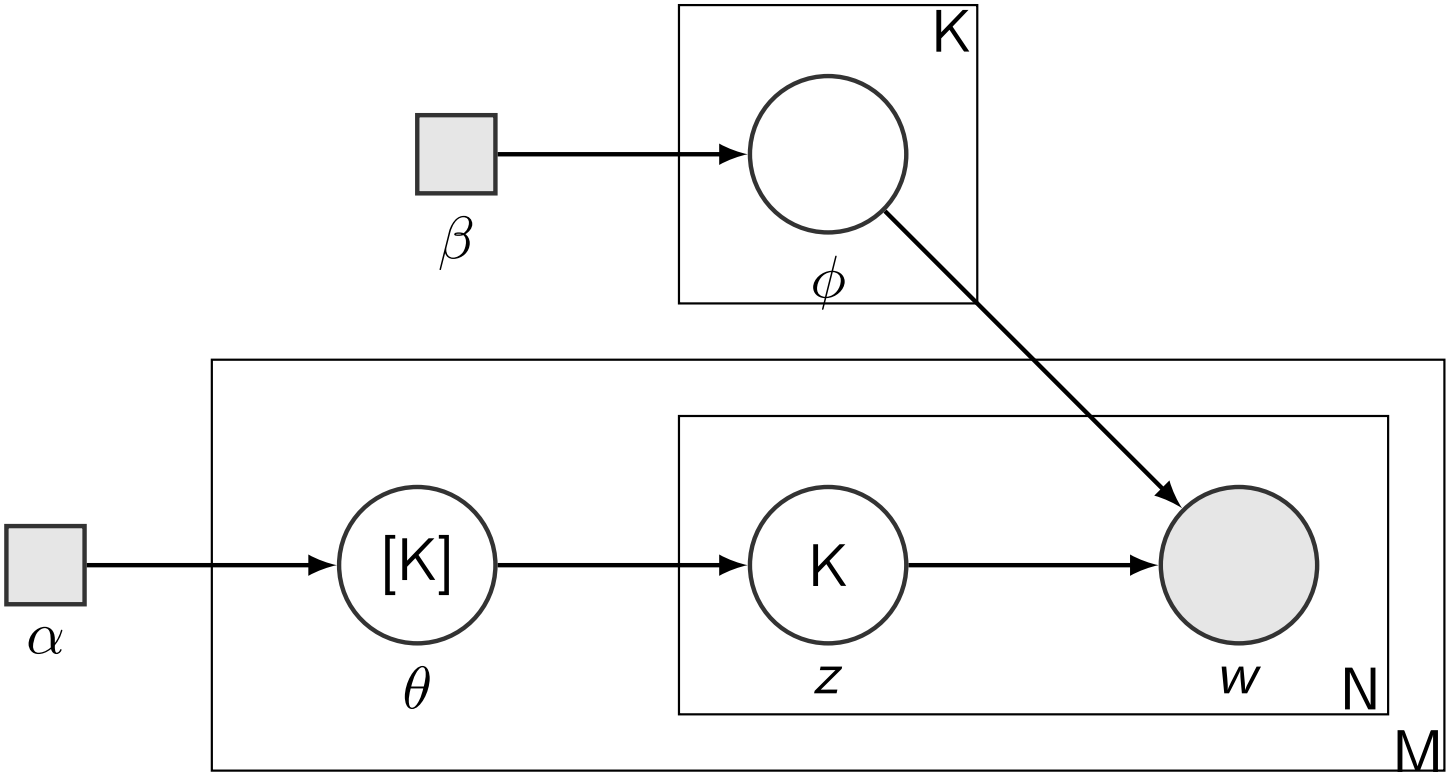
LDA: Building Model



LDA: Building Model



LDA: Building Model



$$\phi_k \sim Dir(\beta) \quad (18a)$$

$$\theta_i \sim Dir(\alpha) \quad (18b)$$

$$z_{i,j} \sim Cat(\theta_i) \quad (18c)$$

$$w_{i,j} \sim Cat(\phi_{z_{i,j}}) \quad (18d)$$

Expectation-Maximization

Likelihood for mixtures

$$\mathcal{L}(\theta|x) = p(x|\theta) = \sum_z p(x, z|\theta) \tag{19}$$

Expectation

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{z|x,\theta^{(t)}}[\log \mathcal{L}(\theta|x, Z)] \tag{20}$$

Maximization

$$\theta^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)}) \tag{21}$$

Expectation-Maximization

Alternative formulation

$$F(q, \theta) = \mathbb{E}_q[\log \mathcal{L}(\theta|x, Z)] + H(q) \tag{22}$$

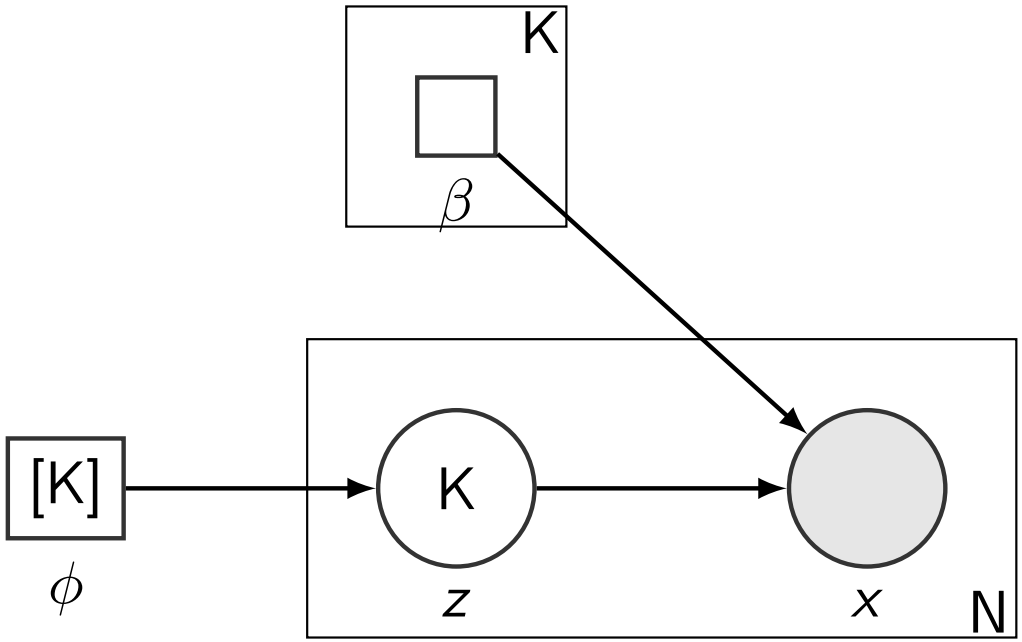
Expectation

$$q^{(t+1)} = \arg\max_q F(q, \theta^{(t)}) \tag{23}$$

Maximization

$$\theta^{(t+1)} = \arg\max_\theta F(q^{(t+1)}, \theta) \tag{24}$$

Example: Mixture of Distributions



$$z_i \sim Cat(\phi) \tag{25a}$$

$$x_i \sim D(\beta_{z_i}) \tag{25b}$$

Example: Mixture of Distributions

Definitions

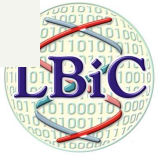
$$\theta = \{\beta_1, \ldots, \beta_K, \phi\} \tag{26a}$$

$$p(x_i|Z_i = k) = p(x_i; \beta_k) \tag{26b}$$

$$p(Z_i = k) = \phi_k \tag{26c}$$

Likelihood

$$\mathcal{L}(\theta|x, z) = p(x, z|\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}(z_i = k)\phi_k p(x_i; \beta_k) \tag{27}$$



Example: Mixture of Distributions

Expectation

$$T_{i,k}^{(t)} = p(Z_i = k|x_i, \theta^{(t)}) = \frac{p(x_i|Z_i = k, \theta^{(t)})p(Z_i = k|\theta^{(t)})}{p(x_i|\theta^{(t)})} \tag{28a}$$

$$= \frac{\phi_k p(x_i; \beta_k)}{\sum_{j=1}^{K} \phi_j p(x_i; \beta_j)} \tag{28b}$$

Likelihood

$$Q(\theta|\theta^{(t)}) = \mathbb{E}[\log \mathcal{L}(\theta|x, Z)] \tag{29a}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} T_{i,k}^{(t)} (\log \phi_k + \log p(x_i; \beta_k)) \tag{29b}$$

Example: Mixture of Distributions

Maximization: ϕ_k

$$\phi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} T_{i,k}^{(t)} \tag{30}$$

Maximization: β_k for normal distribution

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} T_{i,k}^{(t)} x_i}{\sum_{i=1}^{n} T_{i,k}^{(t)}} \tag{31a}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} T_{i,k}^{(t)} (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^\top}{\sum_{i=1}^{n} T_{i,k}^{(t)}} \tag{31b}$$

Conclusion

Mixture models are powerful

- For non-Bayesian mixtures, the big problem may be estimating the parameters of each mixture;
- Expectation-maximization solves the problem partially;
- Easier to learn for exponential family!
- May become intractable very easily.