Brain-like Computing & Machine Intelligence

Latent Dirichlet Allocation (LDA)

Dept. Computer Science & Engineering
Shanghai Jiao Tong University

Latent Dirichlet Allocation (LDA)

- A generative probabilistic model for collections of discrete data such text corpora.
- **A three-level hierarchical Bayesian model**
  - Each document of a collection is modeled as finite mixture over an underlying set of topics.
  - Each topic is characterized by a distribution over words.
- The topic probabilities provide an explicit representation of a document
  - It has natural advantages over unigram model and probabilistic LSI model.

History (1) – Text processing

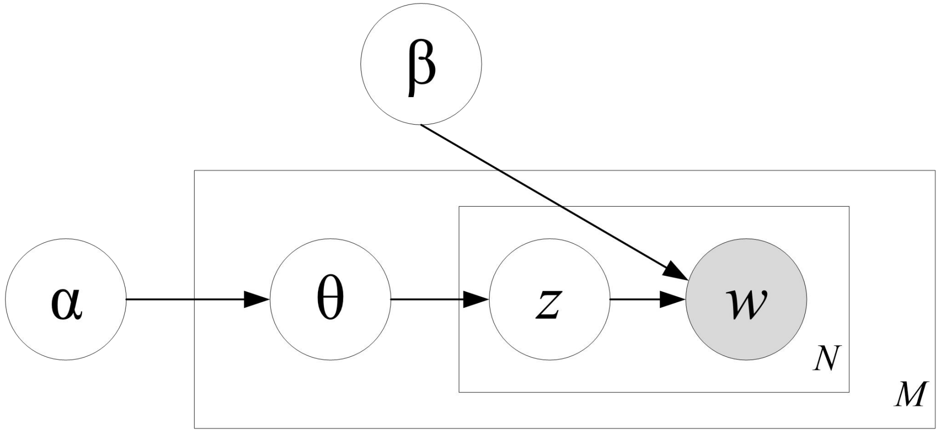
- IR – text to real number vector(Baeza-Yates and Ribeiro-Neto, 1999), tfidf (Salton and McGill, 1983)
  - tfidf – shortcoming: (1) Lengthy and (2) Cannot model inter- and intra- document statistical structure
- LSI – dimension reduction (Deerwester et al., 1990)
  - Advantages: achieve significant compression in large collections and capture synonymy and polysemy.
- Generative probabilistic model – to study the ability of LSI (Papadimitriou et al., 1998)
  - Why LSI, we can model the data directly using maximum likelihood or Bayesian methods.

History (2) – Text processing

- Probabilistic LSI – also aspect model. Milestone (Hofmann, 1999).
  - $P(w_i|\theta_j)$, d={w_1, …, w_N}, and $\boldsymbol{\theta}$={θ_1, …, θ_k}. each word is generated from a single model θ_j. Document *d* is considered to be a mixing proportions for these mixture components $\boldsymbol{\theta}$, that is a list of numbers (the mixing proportions for topics).
  - Disadvantage: **no probabilistic model at document level**.
    - The number of parameters grows linearly with the size of corpora.
    - It is not clear to assign probability to document outside of the collection. (does not make any assumptions about how the mixture weights θ are generated, making it difficult to test the generalizability of the model to new documents.)

Notation

- $D=\{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M\}$, $\boldsymbol{d}=\{w_1, \ldots, w_N\}$,
  and $\boldsymbol{\theta}=\{\theta_1, \ldots, \theta_k\}$,
  equivalently $D=\{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$. (Bold variable denotes vector.)
- Suppose we have V distinct words in the whole data set.



LDA

- **The basic idea:** Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.
- For each document d, we generate as follows:

  1. Choose $N \sim \text{Poisson}(\xi)$
  2. Choose $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$
  3. For each of the N words w_n :
     (a) Choose a topic $z_n \sim \text{Multinomial}(\boldsymbol{\theta})$
     (b) Choose a word w_n from $p(w_n \mid z_n, \boldsymbol{\beta})$,
     a multinomial probability conditioned on the topic z_n

k topics **z**

$\boldsymbol{\beta}$ is a $\mathrm{k} \times V$ matrix with $\beta_{ij} = p(w_i = 1 \mid z_j = 1)$

Dirichlet Random Variables $\boldsymbol{\theta}$

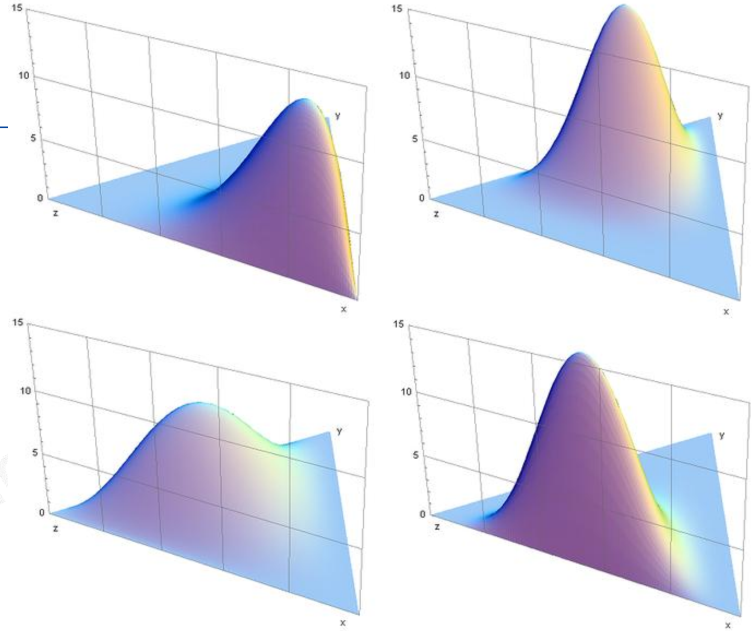
- A *k*-dimensional Dirichlet random variables $\boldsymbol{\theta}$ can take values in the (*k*-1)-simplex
  - (a *k*-vector $\boldsymbol{\theta}$ lies in the (*k*-1)-simplex if $\boldsymbol{\theta}_i \geq 0, \ \sum_{i=1}^{k} \boldsymbol{\theta}_i = 1$, and thus the probability density can be:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{\alpha}_i)} \boldsymbol{\theta}_1^{\boldsymbol{\alpha}_i - 1} \cdots \boldsymbol{\theta}_k^{\boldsymbol{\alpha}_k - 1}$$

where $\boldsymbol{\alpha}$ is a *k*-vector parameter with $\boldsymbol{\alpha}_i>0$, and $\Gamma(x)$ is a gamma function

Graphical Interpretation

- The probability density of the Dirichlet distribution when K=3 for various parameter vectors α.
- Clockwise from top left: α=(6, 2, 2), (3, 7, 5), (6, 2, 6), (2, 3, 4).



$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{\alpha}_i)} \boldsymbol{\theta}_1^{\boldsymbol{\alpha}_i - 1} \cdots \boldsymbol{\theta}_k^{\boldsymbol{\alpha}_k - 1}$$

$$\boldsymbol{\theta}_i \geq 0, \ \sum_{i=1}^{k} \boldsymbol{\theta}_i = 1$$

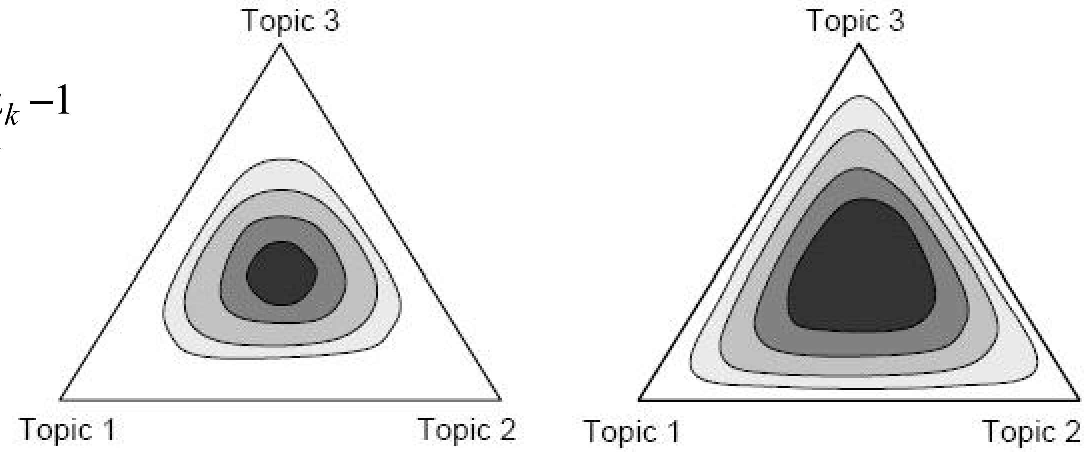


Figure 3. Illustrating the symmetric Dirichlet distribution for three topics on a two-dimensional simplex. Darker colors indicate higher probability. Left: $\alpha = 4$. Right: $\alpha = 2$.

Multinomial Distribution

- Each trial can end in exactly one of k categories

  n independent trials
- Probability a trial results in category i is p_i

$$p_1 + \ldots + p_k = 1$$

- Y_i is the number of trials resulting in category i

$$Y_1 + \ldots + Y_k = n$$

Multinomial Distribution

$$p(y_1, \ldots, y_k \mid p) = \frac{n!}{y_1! \ldots y_k!} p_1^{y_1} \ldots p_k^{y_k},$$

where $\sum_{i=1}^{k} y_i = n, \sum_{i=1}^{k} p_i = 1, y_i \geq 0, p_i \geq 0.$

When $k = 2$,

$$p_i(y_i \mid p) = \frac{n!}{y_i!(n - y_i)!} p_i^{y_i} (1 - p_i)^{n - y_i} \qquad y_i = 0, 1, .., n$$

(Y_i has a marginal binomial distribution)

$$\Rightarrow E(Y_i) = np_i \qquad V(Y_i) = np_i(1 - p_i)$$

Notations

N : The word number in document D

M : The number of documnets

K : The number of topics

V: The word number in the vocabulary

$\boldsymbol{\theta}$: *distribution* of tiopics in document d sampled from Dir($\boldsymbol{\alpha}$)

z_n : The topic variable of word w_n in document d sampled from Multinomial($\boldsymbol{\theta}$)

w_n : word variable sampled from $p(w_n \mid z_n, \boldsymbol{\beta})$, a multinomial probability conditioned on the topic z_n

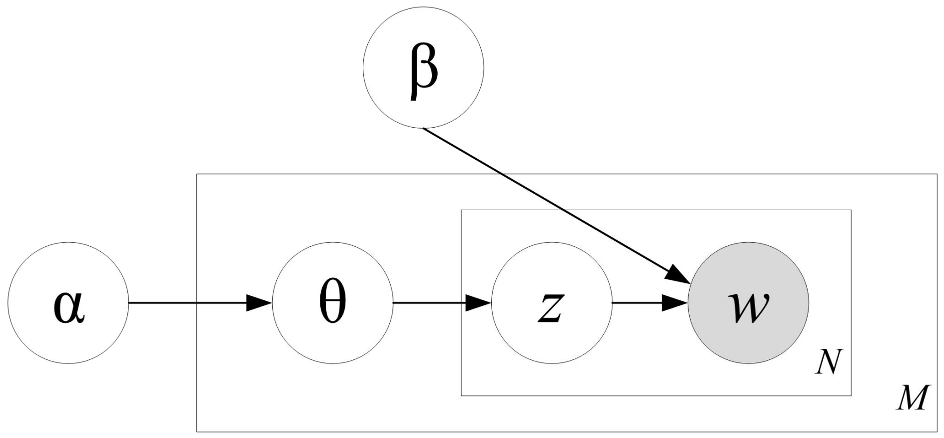
1. Choose N p Poisson(ξ)
2. Choose $\boldsymbol{\theta}$ p Dir($\boldsymbol{\alpha}$)
3. For each of the N words w_n :
   (a) Choose a topic z_n p Multinomial($\boldsymbol{\theta}$)
   (b) Choose a word w_n from $p(w_n \mid z_n, \boldsymbol{\beta})$, a multinomial pdf conditioned on the topic z_n

Joint Distribution

- Given the parameters **α** and **β**, the joint distribution of a topic mixture **θ**, a set of N topics **z**, and a set of N words **w** is given by:

$$p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n \mid \boldsymbol{\theta}) p(w_n \mid z_n, \boldsymbol{\beta}) \tag{1}$$

where $p(z_n|\boldsymbol{\theta})$ is simply θ_i for the unique i such that $z_n^i=1$.



Joint Distribution

- Given the parameters **α** and **β**, the joint distribution of a topic mixture **θ**, a set of N topics **z**, and a set of N words **w** is given by:

$$p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n \mid \boldsymbol{\theta}) p(w_n \mid z_n, \boldsymbol{\beta})$$

where $p(z_n|\theta)$ is simply θ_i for the unique i such that $z_n^i=1$.

Integrating marginal over **θ** and summing over **z** we obtain the distribution of a document

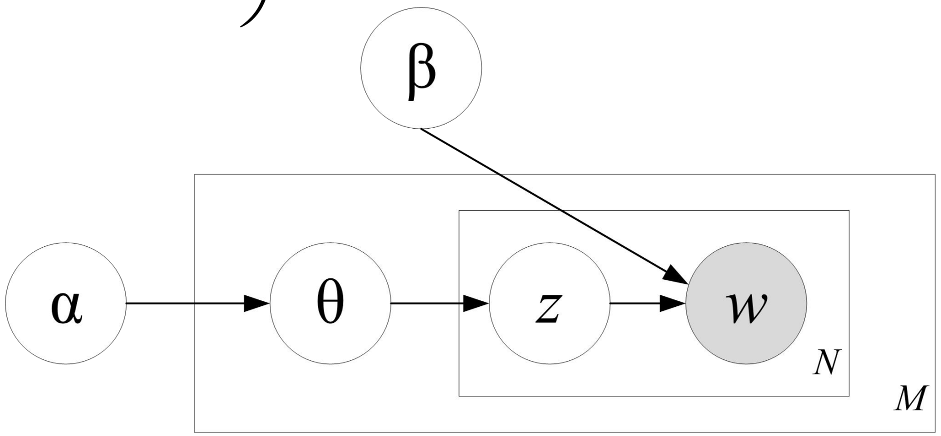
$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \boldsymbol{\theta}) p(w_n \mid z_n, \boldsymbol{\beta}) \right) d\boldsymbol{\theta} \qquad (2)$$

Joint Distribution (cont.)

- Finally, taking the product of the marginal probabilities of single documents, we can obtain the probability of a corpus:

$$p(D \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{M} \int p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \left(\prod_{n=1}^{N_d} \sum_{z_{nd}} p(z_{dn} \mid \boldsymbol{\theta}_d) p(w_{dn} \mid z_{dn}, \boldsymbol{\beta}) \right) d\boldsymbol{\theta}_d$$

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \boldsymbol{\theta}) p(w_n \mid z_n, \boldsymbol{\beta}) \right) d\boldsymbol{\theta}$$



Graphical Interpretation (cont.)

- The Dirichlet prior on the topic-word distributions can be interpreted as forces on the topic locations with higher **β** moving the topic locations away from the corners of the simplex.

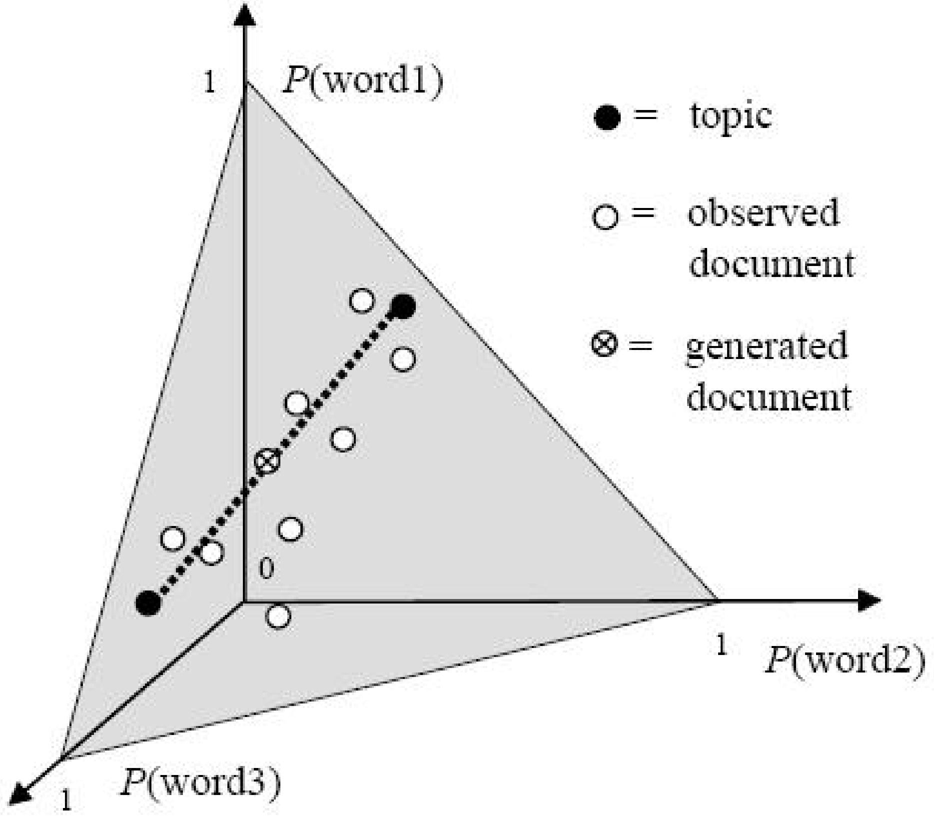


Figure 5. A geometric interpretation of the topic model.

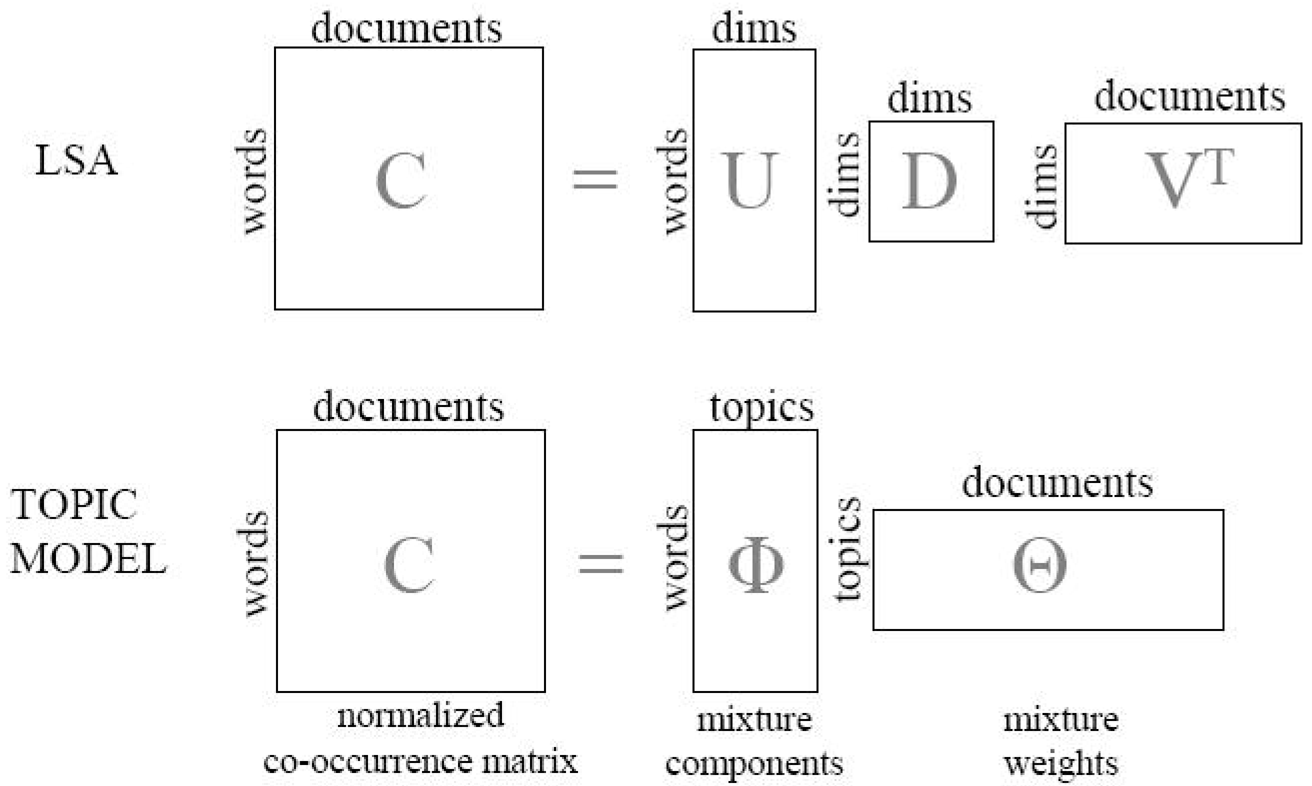
Matrix Interpretation

- In the topic model, the word-document co-occurrence matrix is split into two parts:

  a topic matrix Φ

  a document matrix Θ

- Note that the diagonal matrix D in LSA can be absorbed in the matrix U or V, making the similarity between the two representations even clearer.



Inference and Parameter Estimation

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \boldsymbol{\theta}) p(w_n \mid z_n, \boldsymbol{\beta}) \right) d\boldsymbol{\theta}$$

- Dirichlet random variables

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{\alpha}_i)} \boldsymbol{\theta}_1^{\boldsymbol{\alpha}_1 - 1} \cdots \boldsymbol{\theta}_k^{\boldsymbol{\alpha}_k - 1}$$

$\boldsymbol{\beta}$ is a $k \times V$ matrix
with $\beta_{ij} = p(w_j = 1 \mid z_i = 1)$

- Polynomial distribution

$$p\left(z_1, \ldots, z_k \mid \theta\right) = \frac{1}{z_1! \ldots z_k!} \theta_1^{z_1} \ldots \theta_k^{z_k}$$

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\Gamma\left(\sum_i \boldsymbol{\alpha}_i\right)}{\prod_i \Gamma(\boldsymbol{\alpha}_i)} \int \left(\prod_{i=1}^{k} \boldsymbol{\theta}_i^{\boldsymbol{\alpha}_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\boldsymbol{\theta}_i \boldsymbol{\beta}_{ij})^{w_n^j} \right) d\boldsymbol{\theta} \qquad (3)$$

Inference and Parameter Estimation

- The key inferential problem: To compute the posterior distribution of the hidden variables given a document:

$$p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})} \quad (4)$$

- Such distribution is intractable to compute in general.
  - For normalization in the above distribution, we have to marginalize over the hidden variables and write the Equation (a) in terms of the model parameters:

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\Gamma(\sum_i \boldsymbol{\alpha}_i)}{\prod_i \Gamma(\boldsymbol{\alpha}_i)} \int \left(\prod_{i=1}^{k} \boldsymbol{\theta}_i^{\boldsymbol{\alpha}_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\boldsymbol{\theta}_i \boldsymbol{\beta}_{ij})^{w_n^j} \right) d\boldsymbol{\theta}$$

$$p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n \mid \boldsymbol{\theta}) p(w_n \mid z_n, \boldsymbol{\beta})$$

Inference and Parameter Estimation

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\Gamma(\sum_i \boldsymbol{\alpha}_i)}{\prod_i \Gamma(\boldsymbol{\alpha}_i)} \int \left(\prod_{i=1}^{k} \boldsymbol{\theta}_i^{\boldsymbol{\alpha}_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\boldsymbol{\theta}_i \boldsymbol{\beta}_{ij})^{w_n^j} \right) d\boldsymbol{\theta}$$

- This function is intractable due to the coupling between **θ** and **β** in the summation over latent topics (Dickey, 1983).
- How to deal with the intractable exact inference:
  - Approximate inference algorithms, e.g., Laplace approximation, variational approximation, and Markov chain Monte Carlo (Jordan, 1999).

Variational Inference

- A simple convexity-based **variational algorithm** for inference in LDA.
- The basic idea: to obtain an adjustable lower bound on the log likelihood (Jordan, 1999).

$$\log p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \log \int \sum_{z} \frac{p(\theta, z, w \mid \alpha, \beta) q(\theta, z)}{q(\theta, z)} d\theta \tag{5}$$

$$\geq \int \sum_{z} q(\theta, z) \log p(\theta, z, w \mid \alpha, \beta) d\theta - \int \sum_{z} q(\theta, z) \log q(\theta, z) d\theta$$

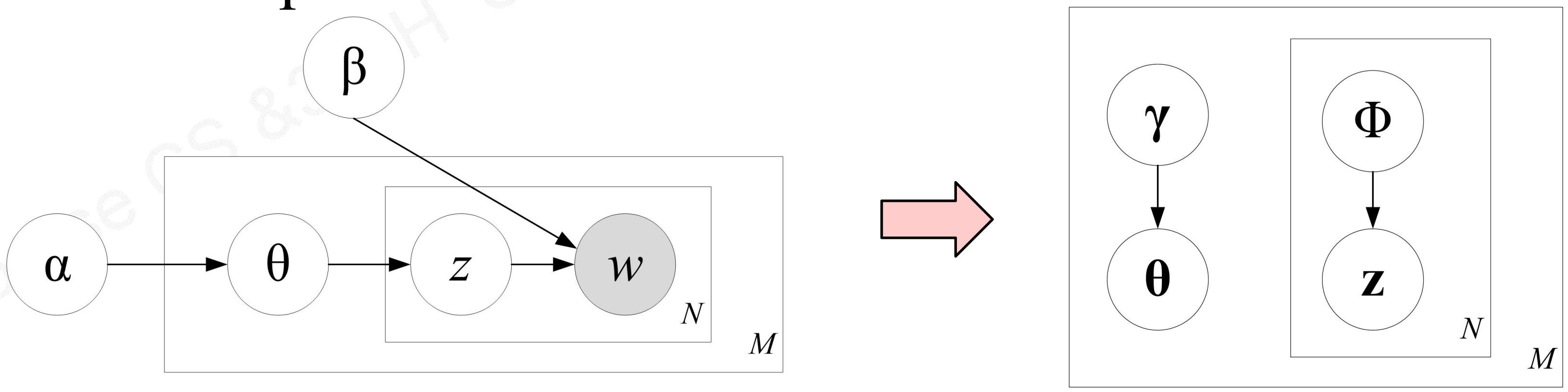
- A simple way to obtain a tractable family of lower bounds is to consider simple modifications of the original graphical model in which some of the edges and nodes are removed.

Variational Inference (cont.)

- By dropping edges between θ, z, and w, and w nodes, and also endow the resulting simplified graphical model with free variational parameters, we obtain a family of distributions on the latent variables:

$$q(\boldsymbol{\theta}, \mathbf{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) \prod_{n=1}^{N} q(z_n \mid \phi_n) \qquad (6)$$

- where the Dirichlet parameter $\boldsymbol{\gamma}$ and the multinomial parameters $(\Phi_1, \ldots, \Phi_N)$ are the free variational parameters.

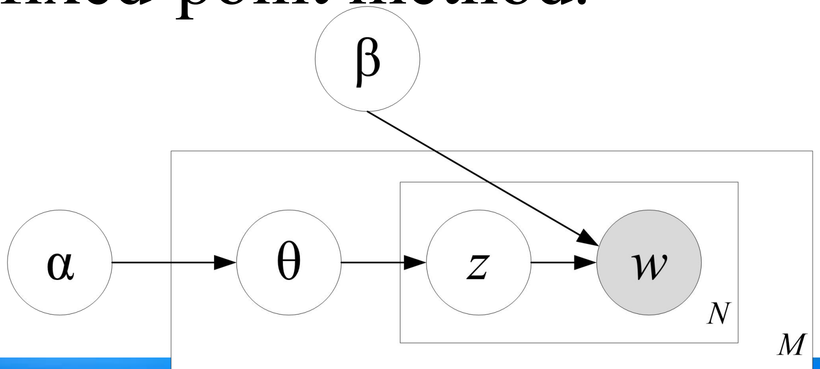


How to determine the parameters

- Set up an optimization problem to determine the values of the variational parameters γ and $\mathbf{\Phi}$.
- We can define the optimization function as minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior $p(\boldsymbol{\theta},\mathbf{z}|\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\beta})$:

$$(\gamma^*,\phi^*) = \arg\min_{(\gamma,\phi)} \mathrm{D}(q(\boldsymbol{\theta},\mathbf{z}\mid\boldsymbol{\gamma},\boldsymbol{\phi})\,\|\,p(\boldsymbol{\theta},\mathbf{z}\mid\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\beta}) \qquad (7)$$

- This minimization can be achieved by an iterative fixed-point method.



Variational Inference

- How to set the parameter $\boldsymbol{\gamma}$ and $\boldsymbol{\Phi}$ via an optimization procedure.
- A lower bound of the log likelihood of a document using Jensen's inequality:

$$
\begin{aligned}
\log q(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \log \int \sum_{\mathbf{z}} p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) d\theta \\
&= \log \int \sum_{\mathbf{z}} \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) q(\boldsymbol{\theta}, \mathbf{z})}{q(\boldsymbol{\theta}, \mathbf{z})} d\theta \\
&> \int \sum_{\mathbf{z}} q(\boldsymbol{\theta}, \mathbf{z}) \log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) d\theta - \int \sum_{\mathbf{z}} q(\boldsymbol{\theta}, \mathbf{z}) \log q(\boldsymbol{\theta}, \mathbf{z}) d\theta \\
&= \mathrm{E}_q[\log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})] - \mathrm{E}_q[\log q(\boldsymbol{\theta}, \mathbf{z})]
\end{aligned}
$$

Variational Inference (cont.)

- The Jensen's inequality provides us with a lower bound on the log likelihood for an arbitrary variational distribution $q(\boldsymbol{\theta},\mathbf{z}|\boldsymbol{\gamma},\Phi)$.
- The difference between the left-hand side and the right-hand side of the above equation is the KL divergence between the variational posterior probability and the true posterior probability.

$$\log p(\mathbf{w} \mid \boldsymbol{\alpha},\boldsymbol{\beta}) = L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta}) + \mathrm{D}(q(\boldsymbol{\theta},\mathbf{z} \mid \boldsymbol{\gamma},\boldsymbol{\phi}) \,\|\, p(\boldsymbol{\theta},\mathbf{z} \mid \mathbf{w},\boldsymbol{\alpha},\boldsymbol{\beta})) \quad (8)$$

$$L(\boldsymbol{\gamma},\boldsymbol{\phi};\boldsymbol{\alpha},\boldsymbol{\beta}) = \int \sum_{z} q(\theta,z) \log p(\theta,z,w \mid \alpha,\beta) d\theta - \int \sum_{z} q(\theta,z) \log q(\theta,z) d\theta$$

Variational Inference (cont.)

- Letting $L(\phi, \boldsymbol{\gamma}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ denote the right-hand side of the above equation we have:

$$\log p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathrm{D}(q(\boldsymbol{\theta}, \mathbf{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}) \,\|\, p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}))$$

- This means that maximizing the lower bound $L(\phi, \boldsymbol{\gamma}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ w.r.t. γ and Φ is equivalent to minimizing the KL divergence between the variational posterior probability and the true posterior probability, the optimization problem in equation (5).

Variational Inference (cont.)

- We can expand the above equation

$$L(\phi,\gamma;\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathrm{E}_q[\log p(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})] + \mathrm{E}_q[\log p(\mathbf{z}\,|\,\boldsymbol{\theta})] + \mathrm{E}_q[\log p(\mathbf{w}\,|\,\mathbf{z},\boldsymbol{\beta})]$$
$$- \mathrm{E}_q[\log q(\boldsymbol{\theta})] - \mathrm{E}_q[\log q(\mathbf{z})]$$

- By extending it again, we can have

$$\begin{aligned} L(\phi,\gamma;\boldsymbol{\alpha},\boldsymbol{\beta}) = {} & \log\Gamma\left(\textstyle\sum_{j=1}^{k}\boldsymbol{\alpha}_j\right) - \sum_{i=1}^{k}\log\Gamma(\boldsymbol{\alpha}_i) + \sum_{i=1}^{k}(\boldsymbol{\alpha}_i - 1)\left(\Psi(\boldsymbol{\gamma}_i) - \Psi\left(\textstyle\sum_{j=1}^{k}\boldsymbol{\gamma}_j\right)\right) \\ & + \sum_{n=1}^{N}\sum_{i=1}^{k}\boldsymbol{\phi}_{ni}\left(\Psi(\boldsymbol{\gamma}_i) - \Psi\left(\textstyle\sum_{j=1}^{k}\boldsymbol{\gamma}_j\right)\right) \quad + \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V}\boldsymbol{\phi}_{ni} w_n^j \log\boldsymbol{\beta}_{ij} \\ & - \log\Gamma\left(\textstyle\sum_{j=1}^{k}\boldsymbol{\gamma}_j\right) + \sum_{i=1}^{k}\log\Gamma(\boldsymbol{\gamma}_i) - \sum_{i=1}^{k}(\boldsymbol{\gamma}_i - 1)\left(\Psi(\boldsymbol{\gamma}_i) - \Psi\left(\textstyle\sum_{j=1}^{k}\boldsymbol{\gamma}_j\right)\right) \\ & - \sum_{n=1}^{N}\sum_{i=1}^{k}\boldsymbol{\phi}_{ni}\log\boldsymbol{\phi}_{ni} \end{aligned} \tag{9}$$

Variaitonal Multinomial

- We first maximize Eq. (15) w.r.t. Φ_{ni}, the probability that the n-th word is generated by latent topic i.

- We form the Lagrangian by isolating the terms which contain Φ_{ni} and adding the appropriate Lagrange multipliers. Let β_{iv} be $p(w^v{}_n=1|z^i=1)$ for the appropriate v. (recall that each w_n is a vector of size V with exactly one component equal to one; we can select the unique v such that $w^v{}_n=1$):

$$L_{[\phi_{ni}]} = \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) + \phi_{ni}\log\boldsymbol{\beta}_{iv} - \phi_{ni}\log\phi_{ni} + \lambda_n(\sum_{j=1}^{k}\phi_{ni} - 1)$$

Variaitonal Multinomial (cont.)

- Taking derivatives w.r.t. Φ_{ni}, we obtain:

$$\frac{\partial L}{\partial \phi_{ni}} = \Psi(\boldsymbol{\gamma}_i) - \Psi\left(\sum_{j=1}^{k} \boldsymbol{\gamma}_j\right) + \log \boldsymbol{\beta}_{iv} - \log \phi_{ni} - 1 + \lambda$$

- Setting this to zero yields the maximizing value of the variational parameter Φ_{ni} :

$$\phi_{ni} \propto \boldsymbol{\beta}_{iv} \exp\left(\Psi(\boldsymbol{\gamma}_i) - \Psi\left(\sum_{j=1}^{k} \boldsymbol{\gamma}_j\right)\right)$$

Variational Dirichlet

- Next we maximize equation (15) w.r.t. γ_i, the i-th component of the posterior Dirichlet parameters, the terms containing γ_i are:

$$L_{[\gamma]} = \sum_{i=1}^{k} (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) + \sum_{n=1}^{N} \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j))$$

$$- \log\Gamma(\sum_{j=1}^{k} \gamma_j) + \log\Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j))$$

- By simplifying

$$L_{[\gamma]} = \sum_{i=1}^{k} (\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)) - \log\Gamma(\sum_{j=1}^{k} \gamma_j) + \log\Gamma(\gamma_i)$$

Variational Dirichlet (cont.)

- Taking the derivative w.r.t. γ_i:

$$\frac{\partial L}{\partial \boldsymbol{\gamma}_i} = \Psi'(\boldsymbol{\gamma}_i)(\boldsymbol{\alpha}_i + \sum_{n=1}^{N} \boldsymbol{\phi}_{ni} - \boldsymbol{\gamma}_i) - \Psi'(\sum_{j=1}^{k} \boldsymbol{\gamma}_j) \sum_{i=1}^{k} (\boldsymbol{\alpha}_i + \sum_{n=1}^{N} \boldsymbol{\phi}_{ni} - \boldsymbol{\gamma}_i)$$

- Setting this equation to zero yields a maximum at:

$$\boldsymbol{\gamma}_i = \boldsymbol{\alpha}_i + \sum_{n=1}^{N} \boldsymbol{\phi}_{ni}$$

Solve the Optimization Problem

- Derivate the KL divergence and setting them equal to zero, we obtain the following update equations:

$$\boldsymbol{\phi}_{ni} \propto \boldsymbol{\beta}_{iw_n} \exp\{\mathrm{E}_q[\log(\boldsymbol{\theta}_i) \mid \boldsymbol{\gamma}]\} \tag{10}$$

$$\boldsymbol{\gamma}_i = \boldsymbol{\alpha}_i + \sum_{n=1}^{N} \boldsymbol{\phi}_{ni} \tag{11}$$

where the expectation in the multinomial update can be computed as follows:

$$\mathrm{E}_q[\log(\boldsymbol{\theta}_i) \mid \boldsymbol{\gamma}] = \Psi(\boldsymbol{\gamma}_i) - \Psi(\sum_{j=1}^{k} \boldsymbol{\gamma}_j) \tag{12}$$

where ψ is the first derivative of the $\log\Gamma$ function which is computable via Taylor approximations

Computing E[log $p(\theta_i|\alpha)$]

- Recall that a distribution is in the exponential family if it can be written in the form:

$$p(x \mid \eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\}$$

where η is the natural parameter, T(x) is the sufficient statistic, and A(η) is the log of the normalization factor.

- Rewrite the Dirichlet in this form by exponentiating the log of Eq.: $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \exp\{(\sum_{i=1}^{k}(\boldsymbol{\alpha}_i - 1)\log\boldsymbol{\theta}_i) + \log\Gamma(\sum_{i=1}^{k}\boldsymbol{\alpha}_i) - \sum_{i=1}^{k}\log\Gamma(\boldsymbol{\alpha}_i)\}$$

Computing E[log(θ_i|α)] (cont.)

- From this form we see that the natural parameter of the Dirichlet is $\eta_i=\alpha_i-1$ and the sufficient statistic is $T(\theta_i)=\log\theta_i$. Moreover, based on the general fact that the derivative of the log normalization factor w.r.t. the natural parameter is equal to the expectation of the sufficient statistic, we obtain:

$$\mathrm{E}[\log(\boldsymbol{\theta}_i) \mid \boldsymbol{\alpha}] = \Psi(\boldsymbol{\alpha}_i) - \Psi(\textstyle\sum_{j=1}^{k} \boldsymbol{\alpha}_j)$$

where ψ is the digamma function, the first derivative of the log Gamma function.

Variational Inference Algorithm

(1) initialize $\phi_{ni}^{0} = \frac{1}{k}$ for all i and n

(2) initialize $\gamma_{\text{i}} = \alpha_i + \frac{N}{k}$ for all i

(3) repeat

(4) for n=1 to N

(5) for i=1 to k

(6) $\phi_{ni}^{t+1} = \beta_{iw_n} \exp(\Psi(\gamma_i^t))$

(7) normalize ϕ_{ni}^{t+1} to sum to 1

(8) $\gamma_i^t = \alpha + \sum_{n=1}^{N} \phi_{ni}^{t+1}$

(9) until convergence

Each iteration requires $O((N+1)k)$ operations

For a single document the iteration number is on the order of the number of words in it

Thus, the total number of operations roughly on the order of N^2k

Parameter Estimation

- We can use a empirical Bayes method for parameter estimation. In particular, we wish to find parameters **α** and **β** that maximize the marginal log likelihood:

$$L(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{d=1}^{M} \log p(\mathbf{w}_d \mid \boldsymbol{\alpha},\boldsymbol{\beta})$$

- The quantity $p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ can be computed by the variational inference as described above.
- An alternating variational EM procedure that maximizes a lower bound w.r.t. the variational parameters **γ** and **Φ**, and then fixed values of the variational parameters, maximizes the lower bound w.r.t. the model parameter **α** and **β**.

Variational EM

1. (E-step) For each document, find the optimizing values of the variational parameters $\{\boldsymbol{\gamma}_d^*, \boldsymbol{\phi}_d^* : d \in D\}$. This is done as described in the previous section.
2. (M-step) Maximize the resulting lower bound on the log likelihood w.r.t. the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This corresponds to finding maximum likelihood estimates with expected sufficient statistics for each document under the approximate posterior which is computed in the E-step. Actually, the update for the conditional multinomial parameter $\boldsymbol{\beta}$ can be written out analytically:

$$\boldsymbol{\beta}_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \boldsymbol{\phi}_{dni}^* w_{dn}^j \qquad (13)$$

The update for $\boldsymbol{\alpha}$ can be implemented using an efficient Newton-Raphson method. These two steps are repeated until converges.

Conditional Multinomials

- To maximize w.r.t. β, we isolate terms and add Lagrange multipliers:

$$L_{[\boldsymbol{\beta}]} = \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{k}\sum_{j=1}^{V} \boldsymbol{\phi}_{dni} w_{dn}^{j} \log \boldsymbol{\beta}_{ij} + \sum_{i=1}^{k} \lambda_i \left(\sum_{j=1}^{V} \boldsymbol{\beta}_{ij} - 1\right)$$

- Taking the derivative w.r.t. β_{ij} and set it to zero, we have

$$\boldsymbol{\beta}_{ij} = \sum_{d=1}^{M}\sum_{n=1}^{N_d} \boldsymbol{\phi}_{dni} w_{dn}^{j}$$

Dirichlet

- First, we have

$$L_{[\boldsymbol{\alpha}]} = \sum_{d=1}^{M}\left(\log\Gamma(\sum_{j=1}^{k}\boldsymbol{\alpha}_j) - \sum_{i=1}^{k}\log\Gamma(\boldsymbol{\alpha}_i) + \sum_{i=1}^{k}(\boldsymbol{\alpha}_i - 1)(\Psi(\boldsymbol{\gamma}_{di}) - \Psi(\sum_{j=1}^{k}\boldsymbol{\gamma}_{dj}))\right)$$

- Taking derivative w.r.t. α_i, we obtain:

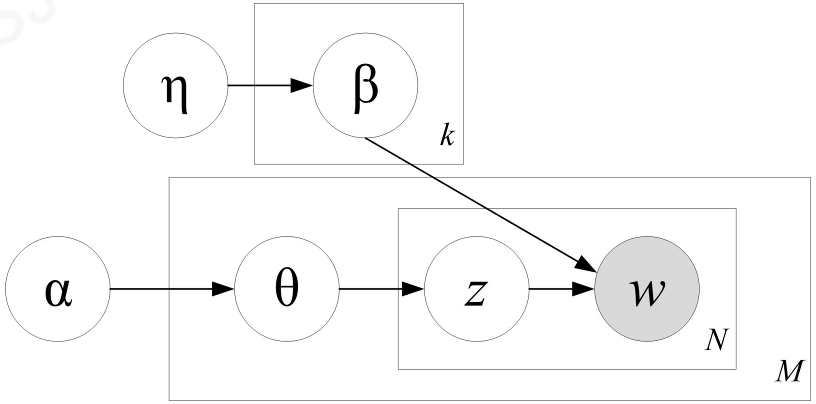
$$\frac{\partial L}{\partial \boldsymbol{\alpha}_i} = M(\Psi(\sum_{j=1}^{k}\boldsymbol{\alpha}_j) - \Psi(\boldsymbol{\alpha}_i)) + \sum_{d=1}^{M}\left(\Psi(\boldsymbol{\gamma}_{di}) - \Psi(\sum_{j=1}^{k}\boldsymbol{\gamma}_{dj})\right)$$

- This derivative depends on α_j, and we therefore must use an iterative method to find the maximal α. In particular, the Hessian is in the form found in equation:

$$\frac{\partial^2 L}{\partial \boldsymbol{\alpha}_i \partial \boldsymbol{\alpha}_j} = \delta(i,j) M \Psi'(\boldsymbol{\alpha}_i) - \Psi'(\sum_{j=1}^{k}\boldsymbol{\alpha}_j)$$

Smoothing

- Simple Laplace smoothing is no longer justified as a maximum a posteriori method in LDA setting.
- We can then assume that each row in $\boldsymbol{\beta}_{k \times V}$ is independently drawn from an exchangeable Dirichlet distribution. That is to treat $\boldsymbol{\beta}_i$ as random variables that are endowed with a posterior distribution, conditioned on the data.



Smoothing Model

- Thus we obtain a variational approach to Bayesian inference:

$$q(\boldsymbol{\beta}_{1:k}, \boldsymbol{\theta}_{1:M}, \mathbf{z}_{1:M} \mid \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\phi}) = \prod_{i=1}^{k} Dir(\boldsymbol{\beta}_i \mid \boldsymbol{\eta}_i) \prod_{n=1}^{N} q_d(\boldsymbol{\theta}_d, \mathbf{z}_n \mid \phi_n, \boldsymbol{\gamma}_d)$$

where $q_d(\boldsymbol{\theta}_d, \mathbf{z}_n \mid \phi_n, \boldsymbol{\gamma}_d)$ is the variational distribution defined for LDA as above and the update for the new variational parameter **η** is as follow:

$$\boldsymbol{\eta}_{ij} \propto \boldsymbol{\eta} + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \boldsymbol{\phi}^{*}_{dni} w^{j}_{dn}$$

Applications

Document Modeling

- Perplexity is used to indicate the generalization performance of a method.
- Specifically, we estimate a document modeling and use this model to describe the new data set.

$$perplexity(D_{test}) = \exp\left(-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right)$$

- LDA outperforms the other models including pLSI, Smoothed Unigram, and Smoothed Mixt. Unigrams.

Document Classification

- We can use the LDA model results as the features for classifiers. In this way, say 50 topics, we can reduce the feature space by 99.6%.
- The experimental results show that such feature reduction may decrease the accuracy only a little.

Other Applications

Corr–LDA:

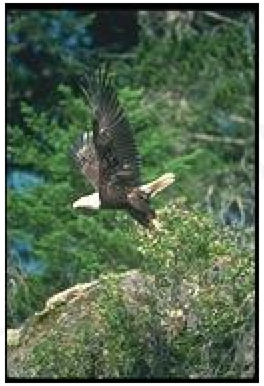
TREE, LIGHT, SUNSET, WATER, SKY

GM–Mixture:

CLOSE–UP, TREE, PEOPLE, MUSHROOMS, LICHEN

GM–LDA:

WATER, SKY, TREE, PEOPLE, GRASS



Corr–LDA:

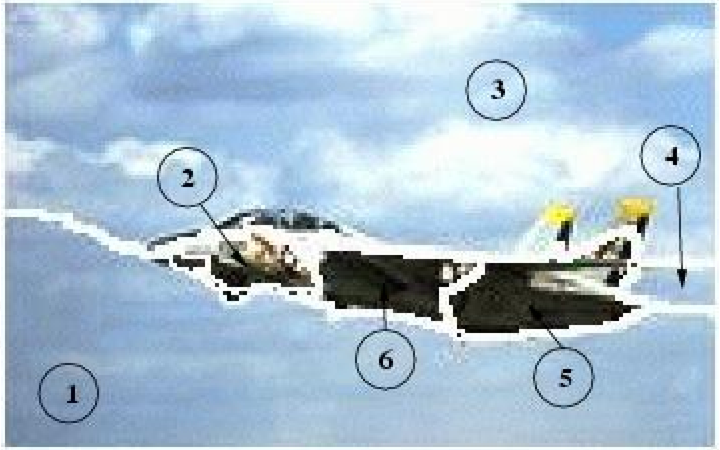
TREE, WATER, GRASS, FLOWERS, BIRDS

GM–Mixture:

TREE, WATER, GRASS, SKY, FIELD

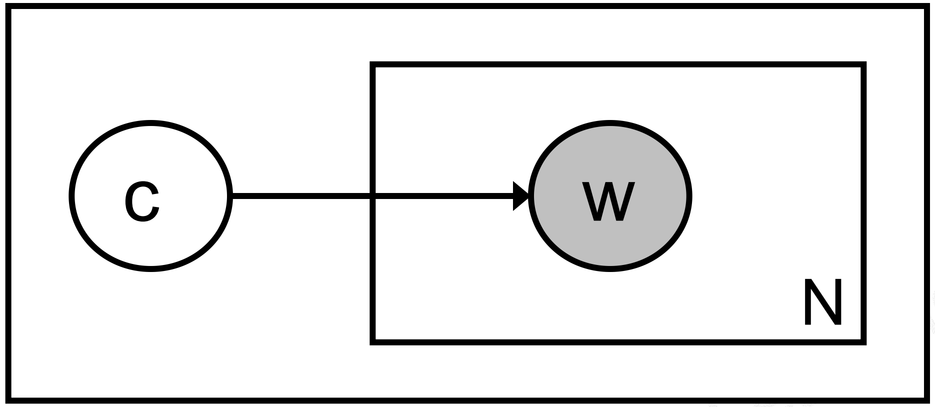
GM–LDA:

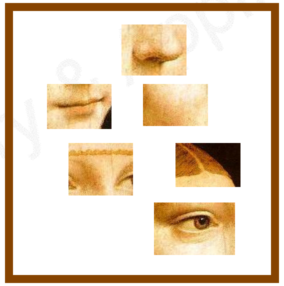
WATER, SKY, TREE, PEOPLE, GRASS



Corr–LDA:	GM–LDA:
1. PEOPLE, TREE	1. HOTEL, WATER
2. SKY, JET	2. PLANE, JET
3. SKY, CLOUDS	3. TUNDRA, PENGUIN
4. SKY, MOUNTAIN	4. PLANE, JET
5. PLANE, JET	5. WATER, SKY
6. PLANE, JET	6. BOATS, WATER

The Naïve Bayes model





$$c^* = \arg\max_c \; p(c \mid w) \propto p(c)\, p(w \mid c)$$

Object class decision

Prior prob. of the object classes

Image likelihood given the class

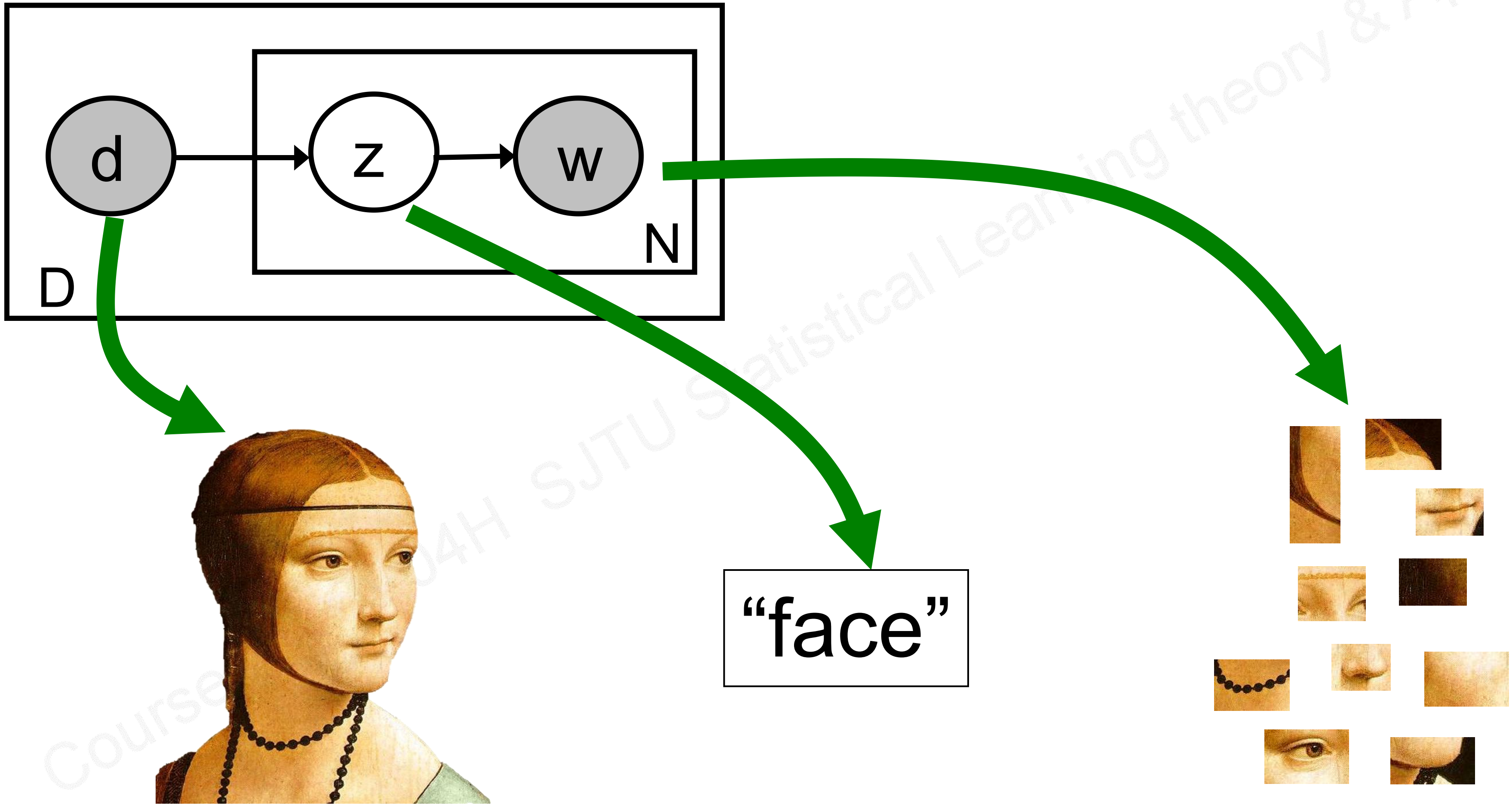
Csurka et al. 2004

in-house database contains 1776 images in seven classes[1]: faces, buildings, trees, cars, phones, bikes and books. Fig. 2 shows some examples from this dataset.

Csurka et al. 2004

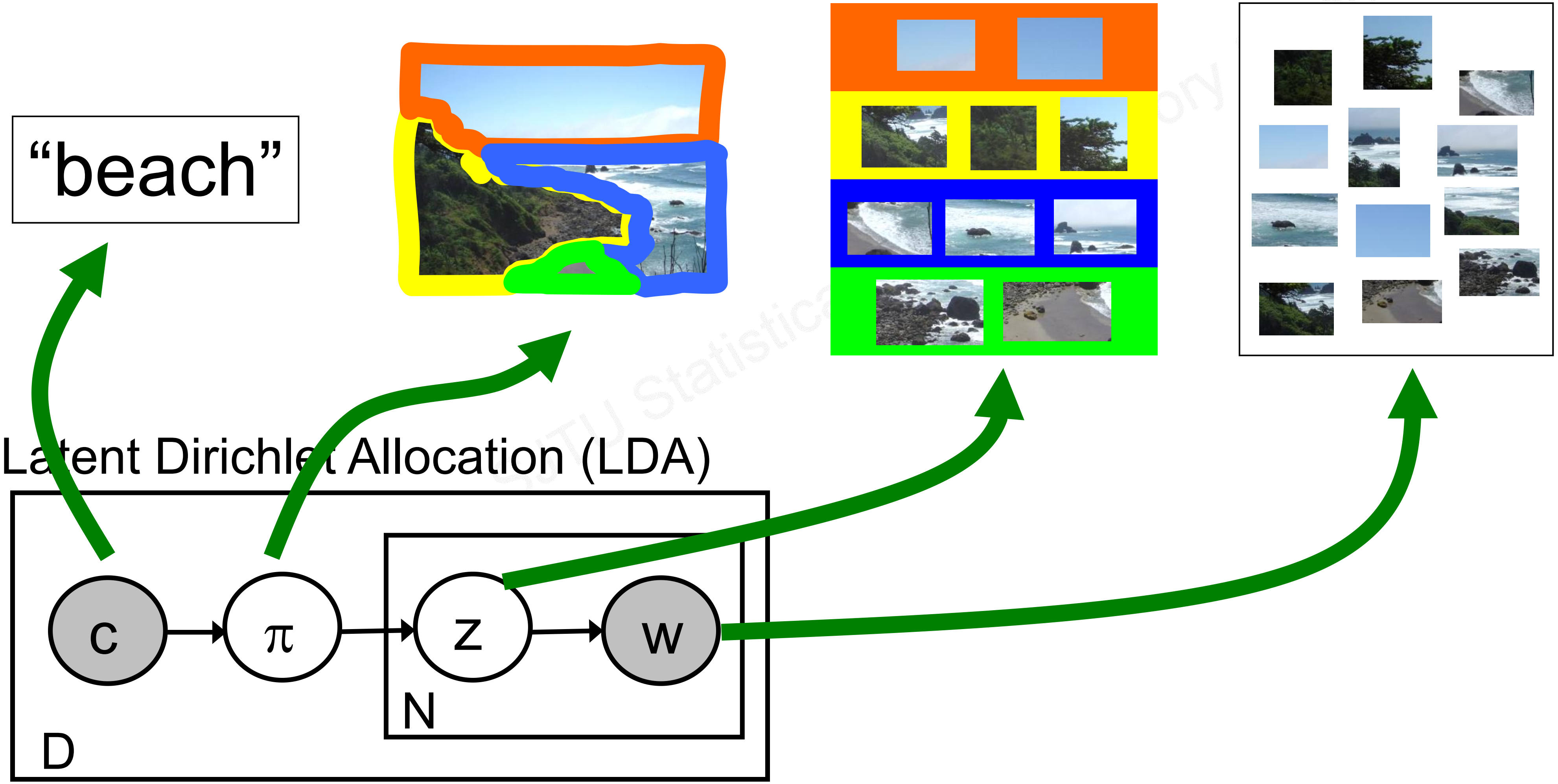
Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



Sivic et al. ICCV 2005

Hierarchical Bayesian text models



Fei-Fei et al. ICCV 2005

Summary

- LDA is Based on the exchangeability assumption
  - Semantic Representation
  - Viewed as a dimensionality reduction technique
  - Exact inference is intractable, we can approximate it instead
  - Applications in other collection – images and caption for example.

End of The Talk !

Conjugation

The Dirichlet distribution is conjugate to the multinomial distribution in the following sense: if

$$\beta \mid X = (\beta_1, \ldots, \beta_K) \mid X \sim \mathrm{Mult}(X),$$

where β_i is the number of occurrences of i in a sample of n points from the discrete distribution on $\{1, \ldots, K\}$ defined by X, then

$$X \mid \beta \sim \mathrm{Dir}(\alpha + \beta).$$

This relationship is used in Bayesian statistics to estimate the hidden parameters, X, of a categorical distribution (discrete probability distribution) given a collection of n samples.

Intuitively, if the prior is represented as $\mathrm{Dir}(\alpha)$, then $\mathrm{Dir}(\alpha+\beta)$ is the posterior following a sequence of observations with histogram β.

Entropy

If X is a Dir(α) random variable, then we can use the exponential family differential identities to get an analytic expression for the expectation of $\log X_i$:

$$\mathrm{E}\left[\log X_i\right] = \psi(\alpha_i) - \psi(\alpha_0) \qquad \alpha_0 = \sum_{i=1}^{K} \alpha_i$$

where ψ is the digamma function: The logarithmic derivative of the gamma function:

$$\Psi(x) = \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

This yields the following formula for the information entropy of X:

$$H(X) = \log \mathrm{B}(\alpha) + (\alpha_0 - K)\psi(\alpha_0) - \sum_{j=1}^{K}(\alpha_j - 1)\psi(\alpha_j)$$

$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}, \qquad \alpha = (\alpha_1, \ldots, \alpha_K).$$