CS 6501 Natural Language Processing

Latent Variable Models

Yangfeng Ji

October 31, 2019

Department of Computer Science
University of Virginia

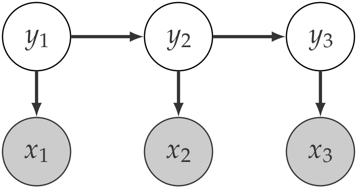
UNIVERSITY of VIRGINIA | ENGINEERING

Overview

1. Latent Variable Models
2. Variational Inference
3. Example: Latent Dirichlet Allocation

Latent Variable Models

Hidden Markov Models



Gaussian Mixture Models

A Gaussian mixture model with K components and each component is a Gaussian distribution

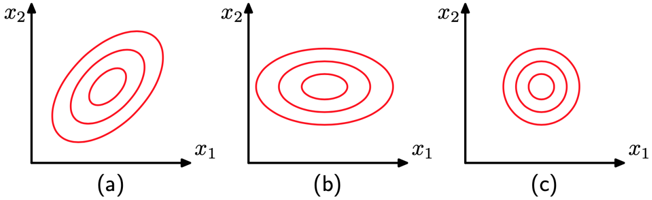
$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

Parameters

- $\boldsymbol{\mu}_k$: mean of the k-th component
- $\boldsymbol{\Sigma}_k$: variance of the k-th component
- π_k: weight of the k-th component with $\sum_k \pi_k = 1$

Gaussian Distribution

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{2}$$

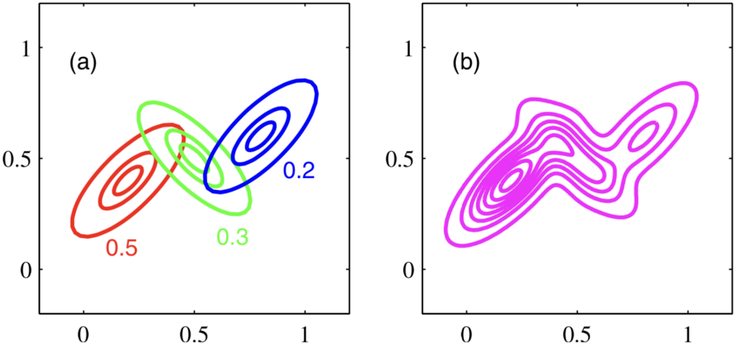


- Mean: $\boldsymbol{\mu}$
- Covariance

  (a) : general form
  (b) : $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$
  (c) : $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$

Gaussian Mixture Models: Example

$$p(\boldsymbol{x}) = \sum_{k=1}^{3} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3}$$



[Bishop, 2006]

GMM as a Latent Variable Model

Given a GMM

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{4}$$

define a K-dimensional binary random vector $\boldsymbol{z}$ to indicate which mixture component a data point comes from

- only one component of $\boldsymbol{z}$ is 1 and all the rest are 0

$$\boldsymbol{z} = [0, \cdots, 0, 1, 0, \cdots, 0] \tag{5}$$

- the probability of each component of $\boldsymbol{z}$, $z^{(k)}$, is defined as

$$p(z^{(k)} = 1) = \pi_k \tag{6}$$

$p(z)$

For each z_k

$$p(z^{(k)} = 1) = \pi_k \tag{7}$$

Overall,

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z^{(k)}} \tag{8}$$

is a categorical distribution with parameters $\{\pi_k\}$

$p(\boldsymbol{x} \mid \boldsymbol{z})$

Using $\boldsymbol{z}$ as an indicator vector, we can redefine $p(\boldsymbol{x} \mid \boldsymbol{z})$ as

- $p(\boldsymbol{x} \mid z^{(k)} = 1)$

$$p(\boldsymbol{x} \mid z^{(k)} = 1) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9}$$

$p(\boldsymbol{x} \mid \boldsymbol{z})$

Using $\boldsymbol{z}$ as an indicator vector, we can redefine $p(\boldsymbol{x} \mid \boldsymbol{z})$ as

- $p(\boldsymbol{x} \mid z^{(k)} = 1)$
$$p(\boldsymbol{x} \mid z^{(k)} = 1) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9}$$
- $p(\boldsymbol{x} \mid \boldsymbol{z})$
$$p(\boldsymbol{x} \mid \boldsymbol{z}) = \sum_{k=1}^{K} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z^{(k)}} \tag{10}$$

Probability

Joint probability of the observed variable $\boldsymbol{x}$ and latent variable $\boldsymbol{z}$

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{z}) &= p(\boldsymbol{z}) p(\boldsymbol{x} \mid \boldsymbol{z}) && (11) \\
&= \prod_{k=1}^{K} \pi_k^{z^{(k)}} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z^{(k)}} && (12)
\end{aligned}
$$

Probability

Joint probability of the observed variable $\boldsymbol{x}$ and latent variable $\boldsymbol{z}$

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{z}) &= p(\boldsymbol{z}) p(\boldsymbol{x} \mid \boldsymbol{z}) && (11) \\ &= \prod_{k=1}^{K} \pi_k^{z^{(k)}} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z^{(k)}} && (12) \end{aligned}$$

Marginal probability of $\boldsymbol{x}$

$$\begin{aligned} p(\boldsymbol{x}) &= \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) && (13) \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) && (14) \end{aligned}$$

Graphical Representation

With N data points from the GMM, each $\boldsymbol{x}_n$ has a latent variable $\boldsymbol{z}_n$ associated with it

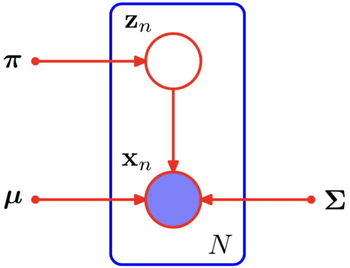
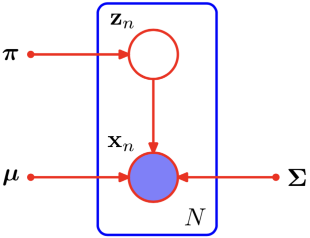


Figure: Graphical representation of GMM [Bishop, 2006]

Generative Story



The generative story of a GMM can be formulated as

1. Randomly pick a mixture component k, with $p(z^{(k)} = 1) = \pi_k$
2. Randomly generate a data point from the k component, $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

The procedure can be repeated multiple times

Parameter Estimation

Given N data points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, parameter estimation on a GMM is an iteration between the following two steps

1. Estimate $p(\boldsymbol{z}_n)$ for every $\boldsymbol{x}_n$
2. Estimate $\{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$ based on $\{p(\boldsymbol{z}_n)\}$ and $\{\boldsymbol{x}_n\}$

Go back to step 1, until convergence

Parameter Estimation

Given N data points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, parameter estimation on a GMM is an iteration between the following two steps

1. Estimate $p(\boldsymbol{z}_n)$ for every $\boldsymbol{x}_n$
2. Estimate $\{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$ based on $\{p(\boldsymbol{z}_n)\}$ and $\{\boldsymbol{x}_n\}$

Go back to step 1, until convergence

Comment

Similar to the K-means algorithm, with $\boldsymbol{z}_n$ as a random variable instead of a determinatic cluster assignment.

Example

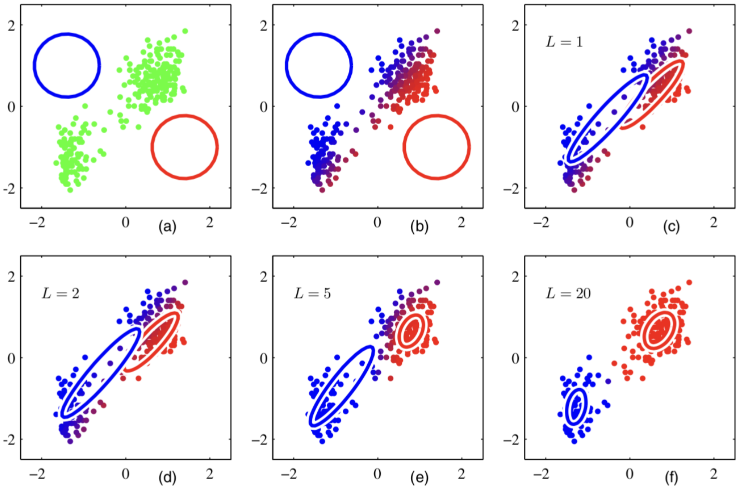
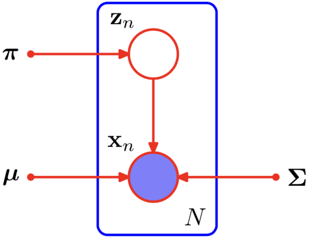


Figure: Illustration of the EM algorithm for GMM parameter estimation.

Comments



A few things about latent variable formulation of GMMs

- z_n is defined on each data point
- the model can be interpreted as a generative story
- marginalizing over z in $p(x, z)$ is tractable

Variational Inference

Ideal Cases

Recall the previous example

1. Estimate the probability of $\boldsymbol{z}$ using

$$p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{x})} \tag{15}$$

2. Estimate $\boldsymbol{\theta}$ using

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}) \tag{16}$$

Ideal Cases

Recall the previous example

1. Estimate the probability of z using

$$p(z \mid x; \theta) = \frac{p(z, x)}{p(x)} \tag{15}$$

2. Estimate θ using

$$\underset{\theta}{\operatorname{argmax}} \log p(z \mid x; \theta) \tag{16}$$

Key Requirement

$$p(x) = \sum_{z} p(x, z) \tag{17}$$

is tractable

Evidence

However, the challenge comes from

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) \tag{18}$$

- The space of $\boldsymbol{z}$ could be exponentially large, when $\boldsymbol{z}$ is discrete
- Integral may be intractable, when $\boldsymbol{z}$ is continuous

Variational Inference

Instead of computing $p(\boldsymbol{z} \mid \boldsymbol{x})$, we define a family of distribution $\mathbb{Q}$, and compute the following optimization problem

$$\tilde{q}(\boldsymbol{z}) = \underset{q(\boldsymbol{z}) \in \mathbb{Q}}{\operatorname{argmin}} \mathrm{KL}(q(\boldsymbol{z}) \| p(\boldsymbol{z} \mid \boldsymbol{x})) \tag{19}$$

where KL divergence is defined as

$$\mathrm{KL}(q \| p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z} \mid \boldsymbol{x})] \tag{20}$$

More about KL Divergence

The Kullback–Leibler divergence measure the difference between two distributions

$$
\begin{aligned}
\mathrm{KL}(q(x)\|p(x)) &= \sum_{q} q(x) \log \frac{q(x)}{p(x)} &\quad (21)\\
&= E_q[\log q(x)] - E_q[\log p(x)] &\quad (22)
\end{aligned}
$$

- $\mathrm{KL}(q\|p) = 0$, if $q = p$
- $\mathrm{KL}(q\|p) \geq 0$

ELBO

$$\mathrm{KL}(q \| p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z} \mid \boldsymbol{x})] \tag{23}$$

ELBO

$$\mathrm{KL}(q \| p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z} \mid \boldsymbol{x})] \tag{23}$$

One more step we need

$$\mathrm{KL}(q \| p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z}, \boldsymbol{x})] + \log p(\boldsymbol{x}) \tag{24}$$

ELBO

$$\mathrm{KL}(q \| p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z} \mid \boldsymbol{x})] \tag{23}$$

One more step we need

$$\mathrm{KL}(q \| p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z}, \boldsymbol{x})] + \log p(\boldsymbol{x}) \tag{24}$$

Evidence lower bound

$$\mathrm{ELBO} = E_q[\log p(\boldsymbol{z}, \boldsymbol{x})] - E_q[\log q(\boldsymbol{z})] \tag{25}$$

Justification

Consider maximizing the log-likelihood of the observed variable

$$\log p(\boldsymbol{x}; \boldsymbol{\theta}) \quad = \quad \log \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$$

Justification

Consider maximizing the log-likelihood of the observed variable

$$\begin{aligned}
\log p(\boldsymbol{x};\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) \\
&= \log \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})}
\end{aligned}$$

Justification

Consider maximizing the log-likelihood of the observed variable

$$\begin{aligned}
\log p(\boldsymbol{x};\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) \\
&= \log \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})}
\end{aligned}$$

Justification

Consider maximizing the log-likelihood of the observed variable

$$
\begin{aligned}
\log p(\boldsymbol{x};\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) \\
&= \log \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) - \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log q(\boldsymbol{z};\boldsymbol{\psi})
\end{aligned}
$$

Justification

Consider maximizing the log-likelihood of the observed variable

$$
\begin{aligned}
\log p(\boldsymbol{x};\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) \\
&= \log \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) - \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log q(\boldsymbol{z};\boldsymbol{\psi}) \\
&= E_q[\log p(\boldsymbol{z},\boldsymbol{x};\boldsymbol{\theta})] - E_q[\log q(\boldsymbol{z};\boldsymbol{\psi})]
\end{aligned}
$$

Justification

Consider maximizing the log-likelihood of the observed variable

$$
\begin{aligned}
\log p(\boldsymbol{x};\boldsymbol{\theta}) &= \log \sum_{z} p(\boldsymbol{x}, z; \boldsymbol{\theta}) \\
&= \log \sum_{z} q(z;\boldsymbol{\psi}) \frac{p(\boldsymbol{x}, z; \boldsymbol{\theta})}{q(z;\boldsymbol{\psi})} \\
&\geq \sum_{z} q(z;\boldsymbol{\psi}) \log \frac{p(\boldsymbol{x}, z; \boldsymbol{\theta})}{q(z;\boldsymbol{\psi})} \\
&= \sum_{z} q(z;\boldsymbol{\psi}) \log p(\boldsymbol{x}, z; \boldsymbol{\theta}) - \sum_{z} q(z;\boldsymbol{\psi}) \log q(z;\boldsymbol{\psi}) \\
&= E_q[\log p(z, \boldsymbol{x}; \boldsymbol{\theta})] - E_q[\log q(z;\boldsymbol{\psi})] \\
&= \underbrace{E_q[\log p(z, \boldsymbol{x}; \boldsymbol{\theta})] + H(q)}_{\text{ELBo}}
\end{aligned}
$$

Mean-field Approximation

A special case of variational inference is called **mean field approximation**, in which different latent variables $\boldsymbol{z}_i$ are independent with each other

$$q(\boldsymbol{z}; \boldsymbol{\psi}) = \prod_i q(\boldsymbol{z}_i; \boldsymbol{\psi}_i) \tag{26}$$

Mean-field Approximation

A special case of variational inference is called **mean field approximation**, in which different latent variables $\boldsymbol{z}_i$ are independent with each other

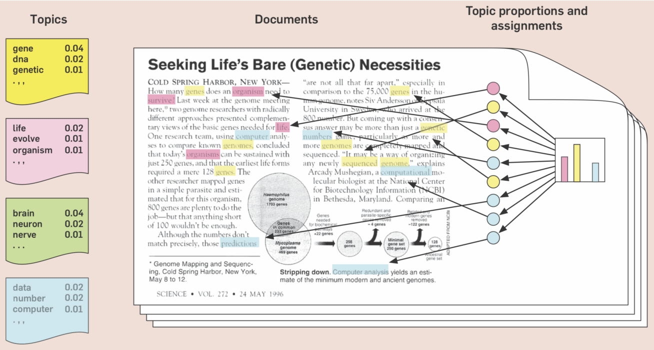
$$q(\boldsymbol{z};\boldsymbol{\psi}) = \prod_i q(\boldsymbol{z}_i;\boldsymbol{\psi}_i) \tag{26}$$

Requirements:

- $p(\boldsymbol{z};\boldsymbol{\psi})$ is factorial
- ψ_i can be compuated with a close-form solution

Example: Latent Dirichlet Allocation

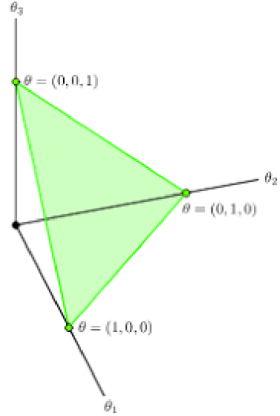
Illustration



The basic idea is that a document is represented as a random *mixture* over latent topics, where each topic is characterized by a distribution over words. [Blei, 2012]

Dirichlet Distribution

$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{27}$$



Generative Story

1. Choose $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$
2. For each word $\boldsymbol{w}_n$
   2.1 Choose a topic $\boldsymbol{z}_n \sim \text{Categorical}(\boldsymbol{\theta})$
   2.2 Choose a word $\boldsymbol{w}_n \sim p(\boldsymbol{w}_n \mid \boldsymbol{z}_n, \boldsymbol{\beta})$, a multinomial probability conditioned on the topic $\boldsymbol{z}_n$.

where

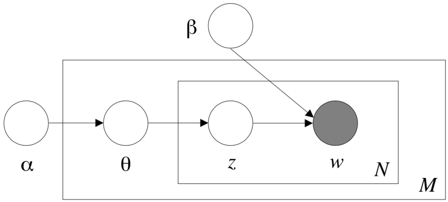
- $\boldsymbol{\theta} \in \mathbb{R}^K$ is a K-dimensional random vector from the Dirichlet distribution with parameter $\boldsymbol{\alpha}$
- $\boldsymbol{\beta} \in \mathbb{R}^{K \times V}$ is a matrix with $\beta_{ij} = p(w_j = 1 \mid z_i = 1)$

Joint Probability

For one document

$$p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}; \boldsymbol{\alpha}) \prod_{n=1}^{N} \left\{ p(\boldsymbol{z}_n; \boldsymbol{\theta}) p(\boldsymbol{w}_n \mid \boldsymbol{z}_n; \boldsymbol{\beta}) \right\} \tag{28}$$

M documents in a corpus



Inference

The key inference problem is to compute the posterior distribution of the hidden variable given a document

$$p(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{d}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{d}; \boldsymbol{\alpha}, \boldsymbol{\beta})} \tag{29}$$

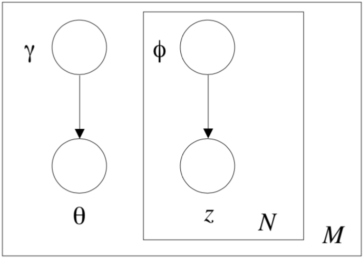
Recall that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the parameters of the original model. $\boldsymbol{\theta}$ and $\boldsymbol{z}$ are latent variables.

Variational Distribution

For one document

$$q(\theta, z; \gamma, \phi) = q(\theta; \gamma) \prod_n q(z_n; \phi) \tag{30}$$

M documents in a corpus

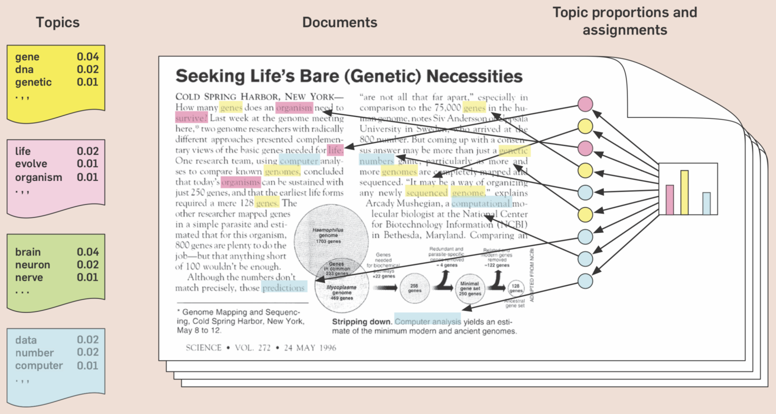


ELBo

$$\begin{aligned}\mathrm{ELBo}_{\mathrm{LDA}} = & E_q[\log p(\boldsymbol{\theta}; \boldsymbol{\alpha}] + E_q[\log p(\boldsymbol{z}; \boldsymbol{\theta}) \\ & + E_q[\log p(\boldsymbol{w} \mid \boldsymbol{z}; \boldsymbol{\beta})] \\ & - E_q[\log q(\boldsymbol{\theta}; \boldsymbol{\gamma})] - E_q[\log q(\boldsymbol{z}; \boldsymbol{\phi})]\end{aligned} \tag{31}$$

As shown in [Blei et al., 2003], every item in Eq. 31 has an analytic form, therefore we can have a closed form solution.

θ and β



[Blei, 2012]

Summary

1. Latent Variable Models
2. Variational Inference
3. Example: Latent Dirichlet Allocation

Reference

Bishop, C. M. (2006).
Pattern Recognition and Machine Learning.
Springer-Verlag.

Blei, D. M. (2012).
Probabilistic topic models.
Communications of the ACM, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
Journal of machine Learning research, 3(Jan):993–1022.