Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

ISSN NO. 2320-5407

Article DOI:10.21474/IJAR01/11239
DOI URL: http://dx.doi.org/10.21474/IJAR01/11239

RESEARCH ARTICLE

A GENTLE TUTORIAL ON DEVELOPING GENERATIVE PROBABILISTIC MODELS AND DERIVING GIBBS SAMPLING A CASE STUDY ON LDA

Dang Hong Linh
Vinh University, Vietnam.

Manuscript Info

Manuscript History
Received: 25 April 2020
Final Accepted: 30 May 2020
Published: June 2020

Key words:-
Data Analysis, Generative Models, Gibbs Sampling, Latent Dirichlet Allocation

Abstract

We present a tutorial on the basics of Bayesian probabilistic modeling and Gibbs sampling algorithms for data analysis. Particular focus is put on explaining detailed steps to build a probabilistic model and to derive Gibbs sampling algorithm for the model. The tutorial begins with basic concepts that are necessary for understanding the underlying principles and notations often used in generative models. Latent Dirichlet Allocation (LDA) is then explained in details regarding both steps to build the model and to derive its collapsed Gibbs sampling algorithm. Following this LDA case study one can further develop either simple or more complex generative models for a variety of applications.

Introduction:-

Employing generative probabilistic modeling approach in data analysis has become quite popular nowadays. More and more problems in different fields have been successfully solved using this approach. Examples include object extraction and pattern recognition [1, 2], topic models [3], and social community detection [4, 5]. There are two main advantages of this approach compared to others. First, it would be easy to postulate complex latent structures underlying the observed data into a mathematical model. Second, correlations among structures can be realized by conditional probability distributions.

Among three main tasks to develop a complete probabilistic model, as pointed out in [6], knowing how to build a model and to estimate posterior distributions for hidden variables are thought more important for people who are new to this approach. This tutorial aims at providing important background for probabilistic modeling tasks and examining the Latent Dirichlet Allocation (LDA) [3] as a case study to detail the steps to build a model and to derive Gibbs sampling algorithms.

In the context of topic extraction from documents and other related applications, LDA is known to be the best model to date. Since its publication in 2003, LDA has been quickly adopted as a powerful tool for extracting clusters of objects in many application domains. These include the topical analysis in text mining [7, 8, 9], object extraction in computer vision [10, 2], and community detection in social network analysis [4, 11, 12, 13]. Even though several models have been introduced as an extension of LDA, it is interesting to note that research communities tend to employ LDA mostly as a blackbox. There are few studies contributing to the explanation of the model [14, 15] but still, the authors skipped most of the detailed steps especially for the posterior estimation.

Corresponding Author:- Dang Hong Linh
Address**:-** Vinh University, Vietnam.

Basic Background:

By employing a statistical modeling approach to analyzing data, a given dataset consisting of data points (also called observations) $D = \{x_1, x_{2, \ldots,} x_N\}$ is assumed to be generated from some probability distribution having parameter θ . Such an assumption is represented by a likelihood function P(D|θ). Even though θ is unknown, one can give some prior knowledge to the model by considering that the values of θ are generated by some distribution P (θ;α), where is known-value parameter called hyperparameter. This is the under lying key idea of Bayesian statistics approach compared to classical statistics where the parameter is assumed to have a fixed value. The joint distribution of the observed data and the parameters defines a probabilistic model.

$$P(D,\theta;\alpha) = P(D|\theta)P(\theta|\alpha) \qquad (1)$$

Thus, under Bayesian statistics point of view, both the dataset D and the parameter θ are considered random variables. One can, therefore, apply Bayes' theorem to compute the posterior distribution of the parameter θ as follows.

$$P(\theta|D;\alpha) = \frac{P(D|\theta)P(\theta|\alpha)}{P(D|\alpha)} \qquad (2)$$

It is intuitive that one can again model α as to be generated by some distribution having possibly unknown parameters. This leads to a possibility to create a hierarchical Bayesian model. Such a probabilistic model represents the underlying generative process of how the dataset D has been produced, given the defined distributions of the variables in the model. All parameters in a probabilistic model except hyperparameters and variables representing observed data are called hidden variables.

By integrating both sides of Eq. 2 with respect to θ, the marginal distribution P(D|α) of the dataset D can be represented in terms of the likelihood function P(D|θ) and the prior distribution P(θ|α).

$$P(D|\alpha) = \int_\theta P(D|\theta)P(\theta|\alpha)d\theta \qquad (3)$$

In addition to the computation of the posterior distribution of the parameters in the model for explaining the observed data in the dataset D, one can also derive a prediction for a new coming observation. Specifically, the joint probability of a new observation x_{new} and the parameter θ given the observed data in the dataset D is computed as follows.

$$P(x_{new}, \theta|D;\alpha) = P(x_{new}|\theta)P(\theta|D;\alpha) \qquad (4)$$

By integrating over the parameter θ, the probability of a new data point given the previous ones is computed.

$$P(x_{new}|D;\alpha) = \int_\theta P(x_{new}|\theta)P(\theta|D;\alpha)d\theta \qquad (5)$$

2.1. Conjugate Prior

There are two leaning problems regarding a probabilistic model presented in Eq. 1 for an observed dataset D. These include the estimation of the parameter θ to best explain the underlying patterns existing in the dataset (Eq. 2), and the prediction for a new observation (Eq. 5). Bayesian approach computes the posterior distribution of parameters and uses some statistics (e.g., the expectation and variance) of the derived distribution as the estimation quality or confidence of the parameters. Therefore, it is required to marginalize (i.e., to compute the summation or the integral) over the whole of parameter space, which often becomes quite difficult in terms of computation. The common strategy to get the computation tractable and also to build a framework for prediction is to employ conjugate prior distributions. A prob-ability distribution P(θ|α) is called conjugate prior of a likelihood function P(D|θ) if the posterior distribution P(θ|D;α) has the same functional form as the prior. A detailed discussion of the existence of a prior distribution for a likelihood function built from a probability density in exponential family probability distributions is presented in [16, Section 2.4].

In a probabilistic model, the likelihood function represents our view about the observed dataset (i.e., from which distribution the dataset is generated), which is fixed under the application. Therefore, one tries to seek a prior distribution that is conjugate to the defined likelihood. For a further explanation, we represent the posterior distribution of the probabilistic model in Eq. 1 as follows.

$$P(\theta|D;\alpha) = \frac{P(D|\theta)P(\theta|\alpha)}{\int_\theta P(D|\theta)P(\theta|\alpha)d\theta} \qquad (6)$$

The underlying principle of using a conjugate prior to the likelihood is that it makes the calculation of the integral in the denominator (i.e., the marginal distribution of the dataset) become simple. In particular, each product $P(D|\theta)P(\theta|\alpha)$ returns an expression of the same form as of the prior distribution with the information from the dataset D added to the hyperparameter α. Therefore, the denominator is thus the integral of the unnormalized density function of the updated prior distribution over the parameter space. Consequently, this integral results in an inversion of the normalizing constant of the updated prior distribution with respect to the dataset D. As an example, we consider in the following the conjugacy between the Dirichlet distribution and the Multinomial distribution, which is used later in this paper for explaining examples and applications.

Multinomial variable:

A random variable X that can take one of K categorical values, so that DOM(X) = {1, …, K}, is called a multinomial variable. If we denote the probability that "X has the value k" by a parameter θ_k ($\theta_k \geq 0$ and $\sum_{k=1}^K \theta_k = 1$), then the probability distribution of X is given as follows [17].

$$P(X|\theta_1, \theta_2, \ldots, \theta_k) = \prod_{k=1}^K \theta_k^{\delta(X,k)} \text{ where } \delta(X,k) = \begin{cases} 1 & \text{if } X = k \\ 0 & \text{if } X \neq k \end{cases} \qquad (7)$$

Consider a dataset $D = (x_1, x_2, \ldots, x_N\}$ that is generated by taking N independent trials on the multinomial variable X defined by $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$ then the likelihood function of the dataset is

$$P(D|\theta) = \prod_{i=1}^N P(x_i|\theta) = \prod_{i=1}^N \prod_{k=1}^K \theta_k^{\delta(x_i,k)} = \prod_{k=1}^K \theta_k^{\sum_{i-1}^N \delta(x_i,k)} = \prod_{k=1}^K \theta_k^{c_k} \qquad (8)$$

where c_k is the number of data points in the dataset that has the value k. The likelihood function of a dataset generated as described is the unnormalized Multinomial probability distribution [18, 16].

Dirichlet distribution:

To complete a probabilistic model for the Multinomial dataset D as described, we need to specify a prior distribution for the multinomial parameter θ. The Dirichlet probability distribution is selected because it is conjugate prior to the Multinomial distribution. The Dirichlet distribution is defined as

$$\text{Dirichlet}(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1} \qquad (9)$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ is a K-dimensional hyperparameter and each α_k is a positive real number indicating the prior belief that one puts on the corresponding component θ_k of the multinomial parameter θ.

Dirichlet distribution has a number of interesting properties. For example, if all components of the hyperparameter α are assigned a small value (i.e., $\sum_{k=1}^K \alpha_k \rightarrow 0$) then the distribution can be simplified as in Eq. 10, which leads to a phenomenon that the θ_s with many zero-components are heavily favored.

$$\text{Dirichlet}(\theta|\alpha)\ \alpha \prod_{k=1}^K \theta_k^{\alpha_k - 1}\ \alpha \prod_{k=1}^K \frac{1}{\theta_k} \qquad (10)$$

The expectation of the Dirichlet distribution, i.e., the expectation of a component θ_k in θ, is computed as follows.

$$E[\theta_k|\alpha] = \frac{\alpha_k}{\alpha_0} \quad \text{where} \quad \alpha_0 = \sum_{k=1}^K \alpha_k \qquad (11)$$

Posterior distribution:

Having the likelihood function (Eq. 8) and the Dirichlet prior distribution (Eq. 10) described, the posterior distribution of the parameter θ (Eq. 6) is now computed by

$$P(\theta|D;\alpha) = \frac{\prod_{k=1}^K \theta_k^{c_k} \prod_{k=1}^K \theta_k^{\alpha_k - 1}}{\int_\theta \prod_{k=1}^K \theta_k^{c_k} \prod_{k=1}^K \theta_k^{\alpha_k - 1} d\theta} = \frac{\prod_{k=1}^K \theta_k^{c_k + \alpha_k - 1}}{\int_\theta \prod_{k=1}^K \theta_k^{c_k + \alpha_k - 1} d\theta} \qquad (12)$$

By multiplying the denominator of the above equation with 1 represented by

$$\frac{\prod_{\mathrm{k}=1}^{\mathrm{K}} \Gamma(\mathrm{c}_{\mathrm{k}} + \alpha_{\mathrm{k}})\, \Gamma(\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{c}_{\mathrm{k}} + \alpha_{\mathrm{k}})}{\Gamma(\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{c}_{\mathrm{k}} + \alpha_{\mathrm{k}})\, \prod_{\mathrm{k}=1}^{\mathrm{K}} \Gamma(\mathrm{c}_{\mathrm{k}} + \alpha_{\mathrm{k}})}$$

the denominator becomes

$$\int_{\theta} \prod_{k=1}^{K} \theta_k^{c_k+\alpha_k-1} d\theta = \frac{\prod_{k=1}^{K} \Gamma(c_k + \alpha_k)\, \Gamma(\sum_{k=1}^{K} c_k + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_k + \alpha_k)\, \prod_{k=1}^{K} \Gamma(c_k + \alpha_k)} \int_{\theta} \prod_{k=1}^{K} \theta_k^{c_k+\alpha_k-1} d\theta$$

$$= \frac{\prod_{k=1}^{K} \Gamma(c_k + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_k + \alpha_k)} \int_{\theta} \frac{\Gamma(\sum_{k=1}^{K} c_k + \alpha_k)}{\prod_{k=1}^{K} \Gamma(c_k + \alpha_k)} \prod_{k=1}^{K} \theta_k^{c_k+\alpha_k-1} d\theta = \frac{\prod_{k=1}^{K} \Gamma(c_k + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_k + \alpha_k)}$$

(13)

Finally, the posterior distribution of θ is

P(θ|D;α) = $\frac{\Gamma(\sum_{k=1}^{K} c_k+\alpha_k)}{\prod_{k=1}^{K} \Gamma(c_k+\alpha_k)} \prod_{k=1}^{K} \theta_k^{c_k+\alpha_k-1} = Dirichlet(\theta|c+\alpha)$ (14)

where c = (c_1, c_2,…, c_K). Thus, the posterior distribution of the parameter θ is the Dirichlet distribution where the information from the dataset (i.e., the count of the number of data points for each category) is added to the hyperparameter α. One can now, for example, estimate each component of θ using the expectation of the Dirichlet distribution.

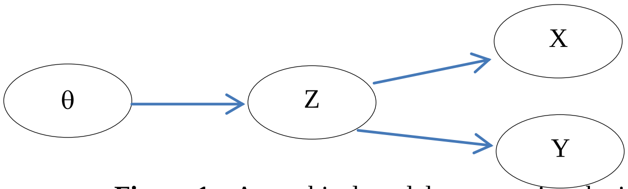
E$[\theta_k|c+\alpha]$ = $\frac{c_k+\alpha_k}{\alpha_0}$ where $\alpha_0 = \sum_{k=1}^{K} c_k + \alpha_k$ (15)

Graphical Model

One of the challenges in presenting a probabilistic model is that it is hard to explain the joint distribution of all random variables in the model. This is because of a huge number of combinations of the values of variables in the model. Even in the simplest case where the model has N binary valued random variables, the joint distribution requires a specification of 2^N numbers - the probabilities of 2^N different assignments of the values of variables X_1 ,…, X_N . Graphical model is a language that uses graph notations for intuitively representing a probabilistic model in a compact way and for interpreting the underlying generative process of how the observations in a dataset D are generated from the model. The main idea of graphical model is to exploit the independent of variables to factor the representation of the model into modular components [19].

There are two main classes of graphical models, which are called Bayesian networks and Markov networks. A Bayesian network is represented by a directed graph and hence it is also called directed graphical model. A Markov network is represented by an undirected graph and is called Markov random fields (MRFs) or undirected graphical model. In the following paragraph, we briefly give some basics of a Bayesian network that will be employed to develop generative probabilistic models. For detailed discussions of graphical models, we refer the reader to [16, 20, 19, 21].

A graphical model for a Bayesian network representing the joint distribution P (X_1, X_2,…, X_N) of random variables X_1, …, X_N is a directed acyclic graph G. Nodes of the graph are random variables in the model. Each directed edge is created to connect two variables that have a conditional (probability) distribution relationship in the factorization of the joint distribution. Specifically, if there is a conditional distribution P(X_k|Pa_{Xk}) in the factorization of the joint distribution P(X_1, X_2,…, X_N) then for each variable $X_i \in Pa_{Xk}$ there is a directed edge connecting X_i to X_k. Variables in Pa_{Xk} are called parent variables of X_k. Intuitively, each node X_k in a graphical model represents the conditional distribution of X_k given its parent variables. An important property of a graphical model is that it encodes the local Markov assumption for random variables in the graph. That means each variable X_k in the graph is conditionally independent of its non-descendants given its parent variables [19]. Figure 1 shows a graphical model presenting a probabilistic model consisting of four random variables X, Y, Z and θ where X and Y are conditionally independent given Z, and Z depends on θ.



P(X, Y, Z, θ) =

P(X|Z)P(Y|Z)P(Z|θ)P(θ)

Figure 1:- A graphical model representing the joint distribution of X; Y; Z; and θfactorized based on the (assumption) dependency between variables: P (X, Y, Z;) = (X|Z)P(Y|Z)P(Z|θ)P(θ).

A graphical model can be represented in a more compact way by using plate notations in which several random variables of the same kind are shown in the graph by only one representative node with an index and that node is covered by a box labeled with a number indicating the cardinality of such variables [16, Chapter 8]. Another notation used in graphical model is that nodes represented for observed random variables (i.e., variables encode the observed features of data points in a dataset) are shaded.

As an example, we consider the joint distribution shown in Eq. 16 that represents a generative probabilistic model for a dataset D = {x_1, x_2, ---, x_N}. Here we assume that data points x_i are generated by P(x_i|θ) and the prior distribution for parameter has some hyperparameter . The corresponding graphical models are shown in Figure 2.

$P(x_1, x_2, \ldots, x_N) = P(\theta;\alpha)\prod_{i=1}^{N} P(x_i|\theta)$ (16)

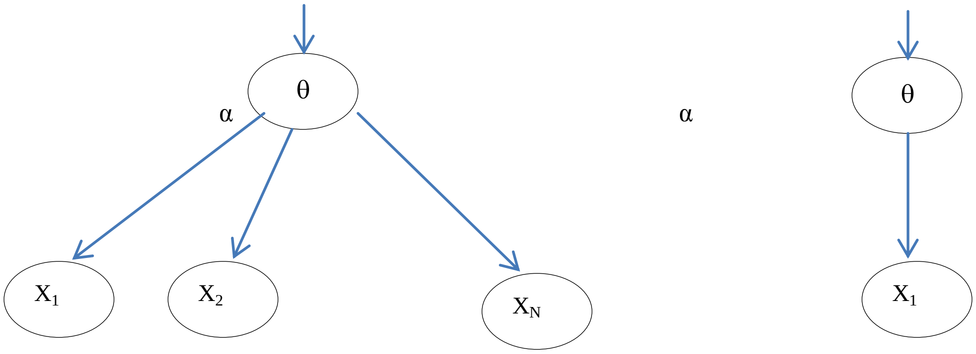


Figure 2:- Two graphical models representing the same probabilistic generative process for a dataset D. The graphical model on the right is presented using plate notations.

Gibbs Sampling for Posterior Estimation:

Computing the posterior distribution of hidden variables, given a dataset and the hyperparameters of the prior distribution of hidden parameters, in a probabilistic model is the main goal for explaining the observed data in the context described by the model. Such a computation is often intractable because of the marginalization, as described above. Note that the integral or summation appears not only in the denominator of Eq. 6 but also in the likelihood function P(D|θ) if one is interested in only some hidden variables and, therefore, needs to integrate out the others.

There are three popular strategies to approximate the posterior distribution of a probabilistic model. These include the sampling based on Markov Chain Monte Carlo [22], Expectation Maximization (EM), and variational parameter methods (optimizationbased). Gibbs sampling [23], a special form of the Metropolis-Hastings algorithm [24], is discussed in this section as we will employ Gibbs sampling in this tutorial. For further details of the EM and variational parameter methods, we refer the reader to [25, 26].

Monte Carlo method:

The underlying idea for deriving the posterior distribution of hidden variables is that if such a probability distribution is computed (or is approximated in most of the cases) then one can use typical statistics such as the expectation and the variance of the distribution to summarize the values of hidden variables. Monte Carlo method is based on the idea that one can learn a complex distribution by repeatedly drawing samples from it and empirically summarizing those samples. For example, the expectation of the posterior distribution defined in Eq. 6 is analytically derived from

$E[\theta|D;\alpha] = \int_\theta \theta P(\theta|D;\alpha)d\theta$ (17)

However, if it is able to produce a large enough number of samples $\theta^{(1)}$, $\theta^{(2)}$, …, $\theta^{(N)}$ from P(θ|D;α) then one can approximate the expectation of with respect to the given dataset D and the hyperparameter by computing the average of such samples.

$E[\theta|D;\alpha] = \int_\theta \theta P(\theta|D;\alpha)d\theta \approx \frac{1}{N}\sum_{i=1}^{N}\theta^{i}$ (18)

The variance of θ is therefore derived from the approximated expectation.

$\mathrm{Var}(\theta|D;\alpha) = E[\theta^2|D;\alpha] - E[\theta|D;\alpha]^2$ (19)

Gibbs Sampling:

It is clear that in order to employ the Monte Carlo strategy to summarize a probability distribution one needs to find a method to correctly draw samples from that distribution. In our scenario of approximating the posterior distribution $P(\theta|D;\alpha)$ of hidden variables, we need to draw $\theta^{(1)}$, $\theta^{(2)}$, …, $\theta^{(N)}$ from $P(\theta|D;\alpha)$. Gibbs sampling is one of the algorithms designed to do so. The basic idea of Gibbs sampling is that it produces a Markov chain of states of hidden variables. The value of a variable at each state is drawn conditionally on the values of other variables. Assume that we need to draw samples from a distribution $P(\theta|D;\alpha)$ where θ consists of K hidden variables $\theta=\{\theta_1, \theta_2,\ldots, \theta_K)$, then the general schema of a Gibbs sampling for that model is as follows.

Algorithm:

A genneral Gibbs sampling algorithm

```
/* State intialization */
$\theta^{(0)} \leftarrow \theta_1^{(0)}, \theta_2^{(0)},.., \theta_K^{(0)}$;
/* Markov chain */
For each t = 1..T do
    For each i = 1..K do
        $\theta_i^{(t+1)} \sim P(\theta_i|\theta_1^{(t+1)}, \theta_2^{(t+1)},\ldots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)},\ldots, \theta_K^{(t)}, D;\alpha)$
```

It is important to note that samples drawn frome a Gibbs sampling algorithm only get to a steady state or converge to the real distribution after a number of iteration called Burn-in stage [23]. Therefore, one needs to discard the results obtained from the first Burn-in steps before collecting samples for summarizing the distribution.

LDA probabilistic model:

LDA is a probabilistic model originally proposed for extracting semantic topics from a corpus of documents. The key idea of the model is that it considers a document as a mixture of topics, a topic being a mixture of terms, and topics are shared among documents [3, 27]. Particularly, given a corpus of documents $D = \{d_1, d_2, ,,,, d_{|D|}\}$ built from a vocabulary set consisting of |V| terms, $V = \{w_1, w_2, \ldots, w_{|V|}\}$, LDA considers words occurring in any document d in the corpus to be independently sampled from a common number of topics $Z = \{z_1, z_2, \ldots, z_{|Z|}\}$. One can, therefore, assume that the topics Z are shared among documents. Another assumption employed in LDA is that documents as well as words within each document are considered to be exchangeable, respectively. To learn the mixture of topics in a document and the mixture of terms in a topic, a probabilistic framework was introduced, which works as follows.

The mixture of terms in a topic $z \in Z$ is modeled as a multinomial dis-tribution specified by a multinomial parameter $\phi_z = \{\phi_{z,w1}, \phi_{z,w2}, \ldots, \phi_{z,w|V|}\}$. Each $\phi_{z,w}$ is the probability that term w belongs to topic z, denoted $P(w|\phi_z)$, such that $\sum_{w \in V} P(w|\phi_{wz}) = 1$. The mixture of topics in a document d, usually referred to as the topic proportion of the document, is also modeled as a multinomial parameter $\theta_d = \{\theta_{d,z1}, \theta_{d,z2}, \ldots, \theta_{d,z|Z|}\}$. Each $\theta_{d,z}$ indicates the likelihood of topic z in document d, denoted $P(z|\theta_d)$, such that $\sum_{z \in Z} P(z|\theta_d) = 1$.

Obviously, if one knows th distribution of term in topic z and the topic proportion of document d beforehand, then the probability that a word w in d belongs to topic z would be

$$P(w, z|\phi_z, \theta_d) = P(z|\theta_d)P(w|\phi_z) = \theta_{d,z}\,\phi_{z,w} \qquad (20)$$

However, generally, we are given a corpus of documents and asked to find some topics in these documents without having knowledge about the distribution of terms in topics and the proportion of topics in documents. In other words, not only the topic that a word should be assigned to but also the distribution of terms in any topic (ϕ_z) and the topic proportion of any document (θ_d) are hidden. One therefore has to learn such hidden variables from the occurrences of terms in the corpus.

Suppose each of the two variables ϕ_z and θ_d is generated by a probability distribution, denoted $P(\phi_z|\beta)$ and $P(\theta_d|\alpha)$, respectively, where α and β are the hyperparameters of the corresponding distribution; then the joint probability of word w and topic z in document d is

$$P(w, z, \phi_z, \theta_d|\alpha,\beta) = P(\phi_z|\beta)P(\theta_d|\alpha)P(z|\theta_d)P(w|\phi_z) \qquad (21)$$

and the joint distribution of all words and topics in document d becomes

$$P(d, z, \phi, \theta_d|\alpha,\beta) = \prod_{z\in Z} P(\phi_z|\beta)\times P(\theta_d|\alpha) \prod_{w\in d} P(z_z|\theta_d)P(w|\emptyset_{z_w}) \qquad (22)$$

where $\phi = \{\phi_z\}$, $z = \{z_w\}$, $w \in d$. Each $z_w \in z$ is a topic index (i.e., 1..|Z|) indicating the topic assignment of word w in document d. Finally, the joint distribution of words and topics in the entire corpus, which is referred to as the joint distribution of the LDA model, is

$$P(D, z, \phi, \theta|\alpha,\beta) = \prod_{z\in Z} P(\phi_z|\beta)\times \prod_{d\in d} P(\theta_d|\alpha) \prod_{w\in d} P(z_w|\theta_d)P(w|\emptyset_{z_w}) \qquad (23)$$

where $\theta = \{\theta_d\}$, $d \in D$.

Substituting $P(z_w|\theta_d)$ and $P(w|\phi_{zw})$ in Eq. 23 by the respective multinomial components, i.e., $\theta_{d,zw}$ of the topic proportion θ_d, and $\phi_{zw,w}$ of the distribution ϕ_{zw} of terms in topic z_w, we have

$$P(D, z, \phi, \theta|\alpha,\beta) = \prod_{z\in Z} P(\phi_z|\beta)\times \prod_{d\in d} P(\theta_d|\alpha) \prod_{w\in d} \theta_{d,z_w} \emptyset_{z_w,w}) \qquad (24)$$

To complete the model, one needs to specify the probability distributions that generate samples of the distribution ϕ_z of terms in a topic, and the topic proportion θ_d of a document. As presented above, both ϕ_z and θ_d are modeled as multinomial parameters. Therefore, the Dirichlet distribution is used as prior of ϕ_z and θ_d. This is due to the conjugacy between the Dirichlet and Multinomial distributions [16].

Thus, one can now present the joint distribution of the LDA model in a more specific way as

$$P(D, z, \phi, \theta|\alpha,\beta) = \prod_{z\in Z} Dir(\phi_z|\beta)\times \prod_{d\in d} Dir(\theta_d|\alpha) \prod_{w\in d} \theta_{d,z_w} \emptyset_{z_w,w} \qquad (25)$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{|Z|})$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_{|V|})$ are the hyperparameters of the Dirichlet distributions, which present prior knowledge for the topic proportion of a document and the distribution of terms in a topic respectively.

Figure 3 shows the graphical models explaining three main joint distributions in the LDA model. (a) and (b) are the graphical models of Eq. 21 and Eq. 22, respectively; (c) is the complete graphical model of LDA represented by Eq. 23.

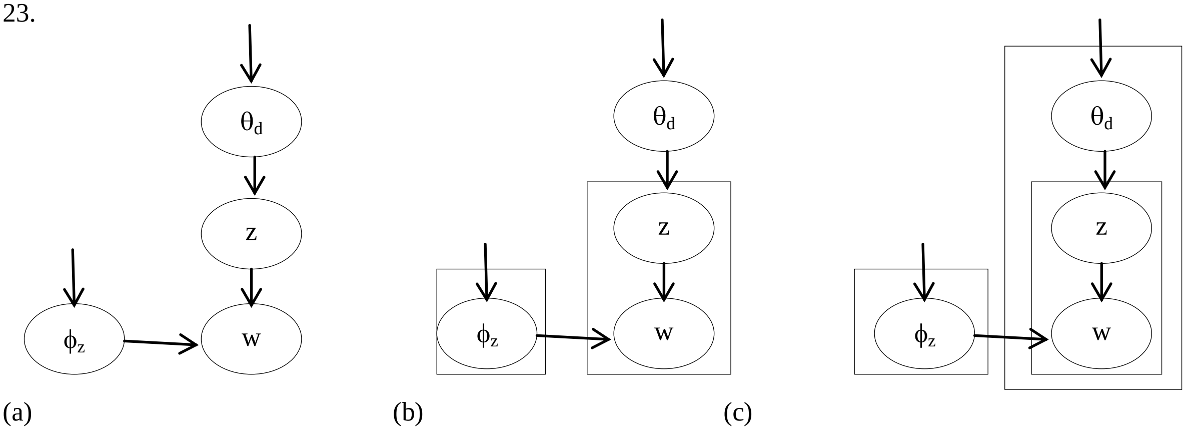


Figure 3:- Graphical models representing selected joint distributions in the LDA model. (a) is the joint distribution of word w in topic z of document d; (b) is the joint distribution of all words and topics in document d; (c) is the complete graphical model of LDA.

Generative process:

Having the graphical model shown in Figure 3(c), the generative process of the LDA model is as follows.

Sample the distribution of term in topic

$\phi = \{\phi_z \sim Dir_{|V|}(\beta)\}$, $z \in Z$

For each document d:

sample topic proportion $\theta_d \sim Dir_{|Z|}(\alpha)$

For each word w in document d:

1. sample a topic index $z \sim Mult(\theta_d)$
2. sample term w in th selected topic z, i.e., $w \sim Mult(\phi_z)$

In the following section, the detailed steps to derive the Gibbs sampling rules for estimating the distributions of hidden variables in LDA are presented.

Gibbs Sampling for LDA:

There are hidden variables represented by z (topic assignments), ϕ (distributions of terms in topics), and θ (topic proportions of documents) in the LDA model. The posterior distribution of such variables is analytically obtained using Bayes' theorem as in Eq. 26. This distribution is, however, intractable to compute due to the marginalization over the hidden variables [3].

$$\mathrm{P}(\mathrm{z}, \phi, \theta|\mathrm{D}; \alpha, \beta) = \frac{P(D,z,\emptyset,\theta|\alpha,\beta)}{P(D|\alpha,\beta)} = \frac{P(D,z,\emptyset,\theta|\alpha,\beta)}{\int_{\emptyset} \int_{\theta} \sum_{z\in Z} P(D,z,\emptyset,\theta|\alpha,\beta) d\theta d\emptyset} \qquad (26)$$

By applying sampling, the posterior distribution is approximated through the samples of the joint distribution as shown in Eq. 27.

$$\mathrm{P}(\mathrm{z}, \phi, \theta|\mathrm{D}; \alpha, \beta) = \frac{P(D,z,\emptyset,\theta|\alpha,\beta)}{P(D|\alpha,\beta)} \alpha P(D, z, \phi, \theta|\alpha, \beta \qquad (27)$$

Generally, implementing a Gibbs sampling algorithm for all variables in the LDA model is straightforward. However, it is inefficient due to the sam-pling for the multinomial parameters ϕ and θ, which can be computed from the topic assignment variables z. In other words, it is better to make use of the conjugacy between the Dirichlet and the Multinomial distributions to integrate out the multinomial parameters θ and ϕ in Eq. 27 and build a collapsed Gibbs sampling for z from which θ and ϕ are then derived. In the following, the detailed steps to integrate out θ and ϕ are given.

First, from Eq. 27, the joint distribution of the topic assignments of all words in the corpus is obtained by

$$\mathrm{P}(\mathrm{z}|\mathrm{D}; \alpha,\beta) = \int_{\emptyset} \int_{\theta} P(z, \emptyset, \theta|D; \alpha, \beta) d\theta d\emptyset \ \alpha \int_{\emptyset} \int_{\theta} P(D, z, \emptyset, \theta|\alpha, \beta) d\theta d\emptyset \qquad (28)$$

It is noted that the second term in Eq. 23 can be represented as

$$\prod_{d\in d} P(\theta_d|\alpha) \prod_{w\in d} P(z_w|\theta_d) P(w|\emptyset_{z_w}) = \prod_{d\in D} \prod_{w\in d} P(w|\emptyset_{z_{w)}} \times \prod_{d\in d} P(\theta_d|\alpha) \prod_{w\in d} P(z_w|\theta_d)$$

(29)

Therefore, the joint distribution of the LDA model (Eq. 23) can be rewritten as follows.

$$P(D, z, \phi, \theta|\alpha; \beta) = \prod_{z\in Z} P(\emptyset_z|\beta) \times \prod_{\mathrm{d}\in D} \prod_{w\in d} P(w|\emptyset_{z_{w)}} \times \prod_{d\in d} P(\theta_d|\alpha) \prod_{w\in d} P(z_w|\theta_d) \quad (30)$$

Summary:

We have presented a tutorial for people who are new to the field of applying generative probabilistic modeling approach to detecting hidden structures in the data. Basic ideas and concepts of Bayesian statistics were first recalled. We then particularly motivated the Gibbs sampling method for estimating the posterior distribution of a probabilistic model. LDA, the most well-known probabilistic model, was studied and explained in detail. In terms of applications, LDA was initially designed for the extraction of topics from a corpus of documents. However, it can be employed to cluster observations in a dataset from various applications, often applying these three assumptions:

1. observations are organized in groups (e.g., a group is a document);
2. it is desirable to share clusters among groups (e.g., topics are shared among documents);
3. both groups as well as observations in each group are exchangeable [27]. Based on such principles, one can

Develop more complex probabilistic models for applications where each data point is described by multiple features. That is, for each observation, more than one feature needs to be jointly considered to compute the likelihood of the observation in a cluster.

References:-

1. I. V. Cadez, S. Gaffney, P. Smyth, A general probabilistic framework for clustering individuals and objects, in: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '00, ACM, New York, NY, USA, 2000, pp. 140{149.
2. X. Wang, E. Grimson, Spatial latent dirichlet allocation, in: Proceedings of Neural Information Processing Systems Conference (NIPS) 2007.
3. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993{1022.

4. T. V. Canh, M. Gertz, rlinktopic: A probabilistic model for discovering
5. Regional linktopic communities, in: Advances in Social Networks Analy-sis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, pp. 21{26.
6. N. Natarajan, P. Sen, V. Chaoji, Community detection in content-sharing social networks, in: Proceedings of the 2013 IEEE/ACM In-ternational Conference on Advances in Social Networks Analysis and Mining, ASONAM '13, ACM, New York, NY, USA, 2013, pp. 82{89.
7. A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin, Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science), Chapman and Hall/CRC, London, third edition, 2013.
8. Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, L. Giles, Detecting topic evolu-tion in scientific literature: How can citations help?, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, ACM, New York, NY, USA, 2009, pp. 957{966.
9. C. Zhang, J. Sun, Large scale microblog mining using distributed mb-lda, in: Proceedings of the 21st International Conference Companion on World Wide Web, WWW '12 Companion, ACM, New York, NY, USA, 2012, pp. 1035-1042.
10. S. Moghaddam, M.Ester, On the design of Ida models for aspectbased opinion mining, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, ACM, New York, NY, USA, 2012, pp. 803-812.
11. T. S. F. Haines, T. Xiang, Video topic modelling with behavioural segmentation, in: Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis, MPVA '10, ACM, New York, NY, USA, 2010, pp. 53-58.
12. X. Wang, N. Mohanty, A. McCallum, Group and topic discovery from relations and text, in: Proceedings of the 3rd international workshop on Link discovery, LinkKDD '05, ACM, New York, NY, USA, 2005, pp. 28-35.
13. G. Zheng, J. Guo, L. Yang, S. Xu, S. Bao, Z. Su, D. Han, Y. Yu, Mining topics on participations for community discovery, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11, ACM, New York, NY, USA, 2011, pp. 445-454.
14. D. Zhou, E. Manavoglu, J. Li, C. L. Giles, H. Zha, Probabilistic models for discovering e-communities, in: Proceedings of the 15th international conference on World Wide Web, WWW '06, ACM, New York, NY, USA, 2006, pp. 173-182.
15. T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences 101 (2004) 5228-5235.
16. G. Heinrich, Parameter estimation for text analysis, Technical Report, Univerity of Lepzig, Germany, 2008.
17. C. M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
18. E. B. Sudderth, Graphical Models for Visual Object Recognition and Tracking, Ph.D. thesis, Massachusetts Institute of Technology, Cam-bridge, Massachusetts, 2006.
19. J. M. Bernardo, A. F. M. Smith, Bayesian Theory, Wiley, 1 edition, 1994.
20. D. Koller, N. Friedman, L. Getoor, B. Taskar, Graphical Models in a Nutshell, in: L. Getoor, B. Taskar (Eds.), Introduction to Statistical Relational Learning, MIT Press, 2007.
21. M. I. Jordan, Graphical models, Statistical Science 19 (2004) 140-155.
22. M. J. Wainwright, M. I. Jordan, Graphical models, exponential families, and variational inference, Found. Trends Mach. Learn. 1 (2008) 1-305.
23. N. Metropolis, S. M. Ulam, The Monte Carlo Method, Journal of the American Statistical Association 44 (1949) 335-341.
24. S. Geman, D. Geman, Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, Pattern Analysis and Machine In-telligence, IEEE Transactions on PAMI-6 (1984) 721-741.
25. W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97-109.
26. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal statistical society, Series B 39 (1977) 1-38.
27. M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, L. K. Saul, An intro-duction to variational methods for graphical models, Mach. Learn. 37 (1999) 183-233.
28. Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei, Hierarchical dirichlet processes, Journal of the American Statistical Association 101 (2004).