Latent Dirichlet Allocation

Hung-An Chang and Chen-Hsiang Yu

December 4, 2007

1 Introduction

In the lectures, we have discussed about modeling techniques such as mixture models and Bayesian networks. In this project, we plan to study an advanced extension of those modeling techniques and work on a generative model for collections of discrete data, the Latent Dirichlet Allocation (LDA) [1]. LDA is a three-level hierarchical Bayesian model that can deal with problems such as document modeling, document classification, and collaborative filtering[1]. However, because of the complex coupling of latent variables in the model, direct computing the marginal probability of a set of data under LDA model is intractable. To handle such intractability, we studied two widely used techniques: Markov Chain Monte Carlo method (MCMC) [3] with Gibbs Sampling [4] and Variational Bayesian (VB) Inference method [1]-[2]. Basically, the MCMC method tries to sample the latent variables according to the posterior probability given the data and the model parameters, and then bases on the sampled values to do inference. The VB inference method, on the other hand, tries to maximize a variational lower bound for the marginal probability by introducing a set of variational parameters. In this project, we will conduct a detailed study of LDA, MCMC, and VB inference. The followings are the tasks we plan to do in the project. The higher level scope of the project can be illustrated as Figure 1 and the topics we will cover are shown in bold.

- Describe the main idea of LDA, MCMC and Variational Bayesian Inference
- Derive the process of estimating the parameters of LDA
- Implement LDA model with Variational Bayesian Inference method to do document classification

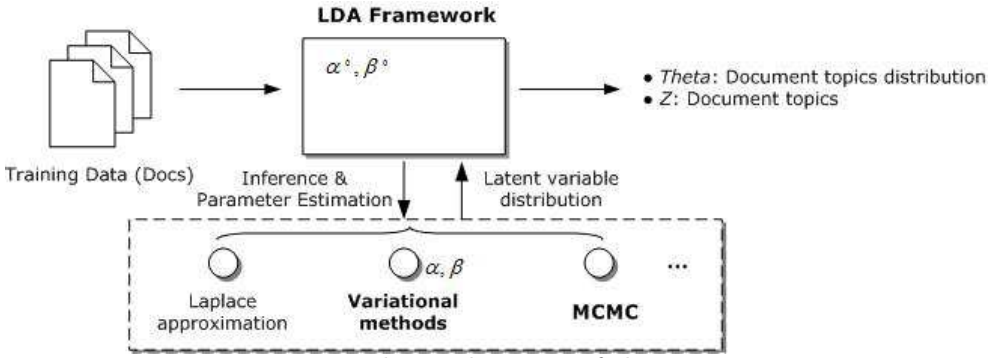

Figure 1: The study scope of the project.

The whole project report will be divided into following sections. In section 2, we will introduce the basic idea of Latent Dirichlet Allocation (LDA) model associated with the notations that will be used in the following

sections. In section 3, MCMC method with Gibbs sampling for LDA will be discussed. Section 4 will discuss Variational Bayesian Inference for LDA, and the comparison of MCMC and VB will be shown in section 5. To further verify the understanding, we have an implementation of LDA with Variational Bayesian Inference in section 6 to demonstrate how to do document classification by using LDA model. Finally, we have a discussion in section 7.

2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for a corpus of discrete data, such as the words in a set of documents. LDA models the words in the documents under the "bag-of-words" assumption, which basically ignores the orders of the words in the documents. Following this "exchangeability", the distribution of the words would be independent and identically distributed given conditioned on some parameters [8]. This conditionally independence allows us to build a hierarchial Bayesian model for a corpus of documents and words. More specifically, the process of how LDA generates the words in a corpus can be illustrated by the graphical model representation in Figure 2. For each document d in the corpus, the LDA model first picks a multinomial distribution $\theta_d = [\theta_{d1} \dots \theta_{dK}]^T$ from the Dirichlet distribution $\alpha = [\alpha_1 \dots \alpha_K]^T$, and then the model assigns a topic $z_{id} = k$ to the ith word in the document according to the multinomial distribution θ_d. Given the topic $z_{id} = k$, the model then pick a word w_{id} from the vocabulary of V words according to the multinomoial distribution $[\phi_{k1} \dots \phi_{kV}]^{\mathrm{T}}$ which is generated from the Dirichlet distribution $[\beta_{k1} \dots \beta_{kV}]^{\mathrm{T}}$ for each topic k.

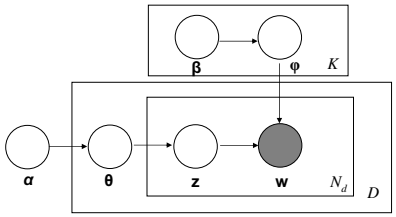

Figure 2: Graphical model representation of LDA.

For convenience, we use the following notations and subscripts for the remaining part of the paper.

- α: The Dirichlet parameters for topics. α_k refers to the prior for topic k.
- β: The topic dependent Dirichlet parameters for word index. β_{kv} refers to the prior for the vth word in the vocabulary under the topic k.
- $\mathbf{w}$: The words in the corpus. w_{id} denotes the ith word in the dth document. For convenience, we use N_d to denote the number of words in the document d.
- θ: The multinomial distributions of topics for the documents in the corpus. θ_{dk} denotes the probability of assign words in d to topic k. $\theta_d = [\theta_{d1} \cdots \theta_{dK}]^{\mathrm{T}}$.
- $\mathbf{z}$: Topic indices. $z_{id} = k$ means that the ith word in document d is assigned to topic k.
- ϕ: The multinomial distributions of word values under certain topic. ϕ_{kv} denotes the probability of generate the vth word in the vocabulary under topic k. $\phi_k = [\phi_{k1} \cdots \phi_{kV}]^{\mathrm{T}}$.
- $\mathbf{n}$: Word counts in the corpus. n_{dkv} denotes the number of words that are inside document d and are assigned to topic k and are with value v.

Also we use $\cdot$ in the subscripts to denote summation over certain valuable. For example, $\alpha_\cdot = \sum_{k=1}^{K} \alpha_k$ and $n_{\cdot kv} = \sum_{d=1}^{D} (n_{dkv})$.

According to the graphic representation in Figure 2, we can express the joint distribution by

$$\begin{aligned} p(\mathbf{w},\mathbf{z},\theta,\phi|\alpha,\beta) &= p(\theta|\alpha)p(\mathbf{z}|\theta)p(\phi|\beta)p(\mathbf{w}|\mathbf{z},\phi) \\ &= (\prod_{d=1}^{D}\frac{\Gamma(\alpha_.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{dk}^{\alpha_k-1})(\prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{dk}^{n_{dk\cdot}})(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k.})}{\prod_{k=1}^{V}\Gamma(\beta_{kv})}\prod_{v=1}^{V}\phi_{kv}^{\beta_{kv}-1})(\prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{kv}^{n_{\cdot kv}}) \\ &= (\prod_{d=1}^{D}\frac{\Gamma(\alpha_.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{dk}^{\alpha_k+n_{dk\cdot}-1})(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k.})}{\prod_{v=1}^{V}\Gamma(\beta_{kv})}\prod_{v=1}^{V}\phi_{kv}^{\beta_{kv}+n_{\cdot kv}-1}), \end{aligned} \tag{1}$$

where $\Gamma(y)=\int_0^\infty t^{y-1}\exp(-t)dt$ is the Γ function and $\Gamma(y)=(y-1)!$ for integer value of y. Note that although we do not express $\mathbf{z}$ explicitly in (1), the effect of $\mathbf{z}$ is reflected by the word counts $n_{dk\cdot}$ and $n_{\cdot kv}$ under the "bag-of-words" assumption.

As other kinds of generative models, it is important to know the probability of a set of data $\mathbf{w}$ generated by LDA, because knowing that probability allows us to do maximum-likelihood estimation of the model parameters and to infer the distribution of latent variables given $\mathbf{w}$. To compute such probability, we have to marginalize all the latent variables θ, ϕ, and $\mathbf{z}$ in (1), and the resulting margin probability can be expressed by

$$p(\mathbf{w}|\alpha,\beta)=\int_\phi\int_\theta\sum_{\mathbf{z}}(\prod_{d=1}^{D}\frac{\Gamma(\alpha_.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{dk}^{\alpha_k+n_{dk\cdot}-1})(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k.})}{\prod_{v=1}^{V}\Gamma(\beta_{kv})}\prod_{v=1}^{V}\phi_{kv}^{\beta_{kv}+n_{\cdot kv}-1})d\theta d\phi. \tag{2}$$

Although the integrations in the above equation related to θ and ϕ are of the form of Dirichlet distribution, which are manageable to compute, we still need to sum over all possible combinations of the topic assignments, all possible combination of $\mathbf{z}$ ($n_{dk\cdot}$ and $n_{\cdot kv}$). Such coupling of $\mathbf{z}$ with θ and ϕ makes direct computation of the marginal probability $p(\mathbf{w}|\alpha,\beta)$ intractable. Because of such intractability, we can not use conventional EM algorithm to estimate the LDA model parameters. Also, we can not directly perform inference tasks since the posterior distributions are not directly computable due to the intractability of the marginal distribution. To handle the intractability, we have studied two approaches, Markov Chain Monte Carlo (MCMC) method and Variational Bayesian (VB) inference method. The details of these two methods will be illustrated in the following sections.

3 Markov Chain Monte Carlo Method using Gibbs Sampling for LDA

Markov Chain Monte Carlo (MCMC) method [4] is a general method to obtain samples from complex distribution. The basic idea of this method can be summarized in the following steps: 1. Map the variable assignments of the distribution to the states of a Markov chain; 2. Specify the transition probabilities of the chain; 3. Start from an initial state and make a sufficient amount of transitions such that the chain can reach its stationary distribution; 4. After the chain reach the stationary distribution, record the state of the chain as the desired sample; 5. If more sample needed, keep recording the states after several transitions.

As shown in the class, we have to construct a Markov chain that is irreducible, aperiodic, and reversible in order to make the chain have a unique stationary distribution. Such properties are guaranteed if we apply the Gibbs sampling for the state transitions [4].

The key idea of the Gibbs sampling is to sequentially update each variable of the distribution according the conditional probability given all the other variables. Take a distribution of three variable (X,Y,Z) for example, the Gibbs sampler makes the Markov chain transit from (x_i,y_i,z_i) to $(x_{i+1},y_{i+1},z_{i+1})$ by the following steps:

- Draw x_{i+1} from $p(X|Y=y_i,Z=z_i)$.
- Draw y_{i+1} from $p(Y|X=x_{i+1},Z=z_i)$.
- Draw z_{i+1} from $p(Z|X=x_{i+1},Y=y_{i+1})$.

The power of Gibbs sampling is that the conditional probability given all the other variables can generally be computed efficiently. We can see this fact as we apply this technique to the inference problem of LDA.

To apply the MCMC with Gibbs sampling for LDA [3], we first need to construct the state space of the Markov chain. As we see in Figure 2, if we are given the topic assignments $\mathbf{z}$, the inference problems for θ and ϕ become independent. Therefore, we can select $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$ as the target distribution, apply MCMC, and use the $\mathbf{z}$ drawn from the stationary distribution of the chain to estimate the values of θ and ϕ. To do this, let us first consider the joint distribution of $\mathbf{z}$ and $\mathbf{w}$

$$\begin{aligned} p(\mathbf{w}, \mathbf{z}|\alpha, \beta) &= \int_\theta \int_\phi p(\mathbf{w}, \mathbf{z}, \theta, \phi|\alpha, \beta) d\theta d\phi \\ &= [\prod_{d=1}^{M} \frac{\Gamma(\alpha_\cdot)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \frac{\prod_{k=1}^{K}\Gamma(\alpha_k + n_{dk\cdot})}{\Gamma(\alpha_\cdot + n_{d\cdot\cdot})}][\prod_{k=1}^{K} \frac{\Gamma(\beta_{k\cdot})}{\prod_{v=1}^{V}\Gamma\beta_{kv}} \frac{\prod_{v=1}^{V}\Gamma(\beta_{kv} + n_{\cdot kv})}{\Gamma(\beta_{k\cdot} + n_{\cdot k\cdot})}]. \end{aligned} \tag{3}$$

Note that the second terms in the brackets result from that the products $\prod_{k=1}^{K} \theta_{dk}^{\alpha_k + n_{dk\cdot} - 1}$ and $\prod_{v=1}^{V} \phi_{kv}^{\beta_{kv} + n_{\cdot kv} - 1}$ in (1) are in the form of Dirichlet distribution without normalization constant, and therefore the results of the integrations are just the inverse of the normalization constants. Given the joint distribution of $\mathbf{w}$ and $\mathbf{z}$ under LDA, we can compute the conditional probability for the Gibbs sampler by

$$p(z_{id} = k|\mathbf{z}^{\neg id}, \mathbf{x}, \alpha, \beta) = \frac{p(\mathbf{z}^{\neg id}, z_{id} = k|\mathbf{x}, \alpha, \beta)}{\sum_{k'=1}^{K} p(\mathbf{z}^{\neg id}, z_{id} = k'|\mathbf{x}, \alpha, \beta)} = \frac{(\alpha_k + n_{dk\cdot}^{\neg id})(\beta_{kv} + n_{\cdot kv}^{\neg id})(\beta_{k\cdot} + n_{\cdot k\cdot}^{\neg id})}{\sum_{k'=1}^{K}(\alpha'_k + n_{dk'\cdot}^{\neg id})(\beta_{k'v} + n_{\cdot k'v}^{\neg id})(\beta_{k'\cdot} + n_{\cdot k'\cdot}^{\neg id})}, \tag{4}$$

where the superscript $\neg id$ means to neglect the effect of the ith word in the dth document, and $v = w_{id}$. The intuitions of deriving the above equation are that assigning w_{id} to different topic only changes the word counts by one and that $\frac{\Gamma(y+1)}{\Gamma(y)} = y$. Note that since the word counts with the $\neg id$ superscript differ from original word counts at most by one, the conditional probability can be computed efficiently given the original counts.

After the Markov chain reach the stationary distribution, we can start drawing samples from the chain. As shown in [3], given a sampled $\mathbf{z}$, we can estimate the values of the other latent variables by

$$\hat{\theta}_{dk} = \frac{\alpha_k + n_{dk\cdot}}{\alpha_\cdot + n_{d\cdot\cdot}} \tag{5}$$

$$\hat{\phi}_{kv} = \frac{\beta_{kv} + n_{\cdot kv}}{\beta_{k\cdot} + n_{\cdot k\cdot}}, \tag{6}$$

where the counts are obtained from the assignment $\mathbf{z}$. The above two equations are derived by computing the expectance of the Dirichlet distribution in the posterior form.

Given a set of training documents, we can construct a Markov chian for the training documents, and run the above MCMC procedures to reach the stationary distribution of the chain. After the stationary distribution is reached, we can store the corresponding statistics. When a new document comes, we can infer the corresponding latent variables for the new document by expanding the state space of the chain with respect to the new document. Because the size of the training data is generally much larger than the new document, the new stationary distribution can be reached reasonably soon if we start from the original stationary distribution.

4 Variational Bayesian Inference for LDA

4.1 Variational Bayesian Inference

Variational Bayesian (VB) inference is another useful tool to deal with complex distributions. Consider the problem of maximum likelihood estimation for generative models; that is,

$$\hat{\mathbf{M}} = \arg\max_{\mathbf{M}} \log(p(\mathbf{X}|\mathbf{M})). \tag{7}$$

As in many machine learning problems, computing the above probability can involve marginalization over some latent variables. Without loss of generality, we can express the above problem by

$$\hat{\mathbf{M}} = \arg\max_{\mathbf{M}} log(p(\mathbf{X}|\mathbf{M})) = \arg\max_{\mathbf{M}} \log(\sum_{\mathbf{Y}} p(\mathbf{X}, \mathbf{Y}|\mathbf{M})), \tag{8}$$

where $\mathbf{Y}$ denotes the latent variables. However, the form of the marginal probability above can be very complex and make direct maximization infeasible as in the LDA case.

Therefore, instead of direct maximizing the log probability above, the VB method tries to maximize a variational lower bound of the above probability by the following. First, the method specifies a set of variational parameters $\mathbf{Q}$, and a corresponding form of variational distribution $q(\mathbf{Y}|\mathbf{Q})$. Using Janson's inequity, we can derive a lower bound for $p(\mathbf{X}|\mathbf{M})$ that does not involve marginalization over $\mathbf{Y}$:

$$\begin{aligned} \log(\sum_{\mathbf{Y}} p(\mathbf{X}, \mathbf{Y}|\mathbf{M})) &= \log(\textstyle\sum_{\mathbf{Y}} \frac{p(\mathbf{X},\mathbf{Y}|\mathbf{M})}{q(\mathbf{Y}|\mathbf{Q})} q(\mathbf{Y}|\mathbf{Q})) \\ &\geqslant \textstyle\sum_{\mathbf{Y}} q(\mathbf{Y}|\mathbf{Q}) \log(\frac{p(\mathbf{X},\mathbf{Y}|\mathbf{M})}{q(\mathbf{Y}|\mathbf{Q})}) \\ &= \mathrm{E}_q[\log(p(\mathbf{X}, \mathbf{Y}|\mathbf{M}))] - \mathrm{E}_q[\log(q(\mathbf{Y}|\mathbf{Q}))], \end{aligned} \tag{9}$$

where E_q means to take expectation over the conditional distribution $q(\mathbf{Y}|\mathbf{Q})$. Note that the inequity in the above formula becomes equity if the variational distribution $q(\mathbf{Y}|\mathbf{Q})$ is the same as the posterior probability $p(\mathbf{Y}|\mathbf{X}, \mathbf{M})$. The difference between the two sides of the inequity is basically the KL divergence $D(q(\mathbf{Y}|\mathbf{Q})||p(\mathbf{Y}|\mathbf{X}, \mathbf{M}))$. Another important fact is that by using the lower bound provided in (9), the problematic marginalization over latent variables in (8) is transformed to the task of taking the expectance over the variational distribution $q(\mathbf{Y}|\mathbf{Q})$.

Therefore, the spirit of the VB method is to find the closest approximation $q(\mathbf{Y}|\hat{\mathbf{Q}})$ for the posterior distribution $p(\mathbf{Y}|\mathbf{X}, \mathbf{M})$ that can minimize the KL divergence. Note that the form of $q(\mathbf{Y}|\mathbf{Q})$ should follow some constraints such that the expectance in (9) is tractable. After we obtained $\hat{\mathbf{Q}}$, we can then plug in $\hat{\mathbf{Q}}$ and maximize the lower bound in (9) with respect to $\mathbf{M}$ using numerical optimization techniques such as Newton-Raphson method.

The above paragraph has summarized how to do parameter estimation for generative models with complex marginal distribution. For the inference problem, we can solve it by the following. Given new data $\mathbf{X}'$ and the model parameters $\hat{\mathbf{M}}$, we can infer the corresponding latent variables $\mathbf{Y}'$ by $q(\mathbf{Y}'|\hat{\mathbf{Q}}')$, where

$$\hat{\mathbf{Q}}' = \arg\max_{\mathbf{Q}'} \mathrm{E}_q[\log(p(\mathbf{X}', \mathbf{Y}'|\hat{\mathbf{M}}))] - \mathrm{E}_q[\log(q(\mathbf{Y}'|\mathbf{Q}'))]. \tag{10}$$

4.2 VB inference for LDA

In this subsection, we summarize how to apply the VB inference method to LDA [1]. The first step of the VB method is to specify a form of the variational distribution. As shown in section 2, the LDA model has three sets of latent variables θ, $\mathbf{z}$, and ϕ; as a result, the most computationally friendly family of the variational distribution would be the one that can be fully factorized into the following form:

$$q(\theta, \mathbf{z}, \phi|\eta, \gamma, \varphi) = q(\mathbf{z}|\gamma)q(\theta|\eta)q(\phi|\varphi) = \prod_{id} q(z_{id}|\gamma_{id}) \prod_{d} q(\theta_d|\eta_d) \prod_{k} q(\phi_k|\varphi_k), \tag{11}$$

where $q(z_{id}|\gamma_{id})$ is a multinomial distribution with $\gamma_{id} = [\gamma_{id1} \cdots \gamma_{idK}]^{\mathrm{T}}$, $q(\theta_d|\eta_d)$ is Dirichlet with η_d, and $q(\phi_k|\varphi_k)$ is Dirichlet with φ_k. The graphical model representation for this family of distribution can be illustrated by Figure 3. As in (9), we can provide a lower bound for $\log(p(\mathbf{w}|\alpha, \beta))$ as

$$\begin{aligned} \ell(\eta, \gamma, \varphi; \alpha, \beta) &= \mathrm{E}_q[\log(p(\mathbf{w}, \theta, \mathbf{z}, \phi|\alpha, \beta))] - \mathrm{E}_q[\log(q(\theta, \mathbf{z}, \phi|\eta, \gamma, \varphi))] \\ &= \mathrm{E}_q[\log(p(\theta|\alpha))] + \mathrm{E}_q[\log(p(\mathbf{z}|\theta))] + \mathrm{E}_q[\log(p(\mathbf{w}|\phi, \mathbf{z}))] + \mathrm{E}_q[\log(p(\phi|\beta))] \\ &\quad -\mathrm{E}_q[\log(q(\theta|\eta))] - \mathrm{E}_q[\log(q(\mathbf{z}|\gamma))] - \mathrm{E}_q[\log(q(\phi|\varphi))], \end{aligned} \tag{12}$$

where the second equity resulted from the independencies specified in Figure 2 and 3. Note that the ';' in above equation is to distinguish variational parameters to model parameters.

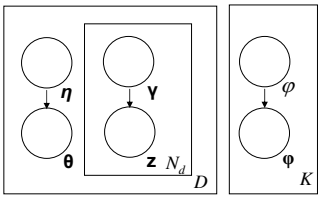

Figure 3: Graphical model representation of the variational distribution.

Given (12), we can derive the values of the variational parameters by the following steps. 1. Use the independencies introduced in Figure 2 to decompose the expectation terms in (12). 2. For each variational parameter ν (can be η_{dk}, φ_{kv}, or γ_{idk}), pull out the terms in the decomposed expressions of (12) that are related to ν and form a function $\ell_{[\nu]}$. 3. Add Lagurange multiplier to $\ell_{[\nu]}$ if needed. 4. Take partial derivative of $\ell_{[\nu]}$ with respect to ν. 5. Solve ν by setting the partial derivative to 0.

Following the above steps and using some insightful mathematical techniques, we can derive the following formula for the variational parameters:

$$\eta_{bk} = \alpha_k + \gamma_{\cdot dk}. \tag{13}$$

$$\varphi_{kv} = (\beta_{kv} + \sum_{d=1}^{D}\sum_{i=1}^{N_d} \delta(w_{id} = v)\gamma_{idk}). \tag{14}$$

$$\gamma_{idk} \propto \exp(\Psi(\eta_{dk}) + \Psi(\varphi_{kw_{id}}) - \Psi(\varphi_{k\cdot})). \tag{15}$$

The $\Psi(\cdot)$ in the above formula is called digmma function, where $\Psi(y) = \frac{d}{dy}\log(\Gamma(y))$, and the value of the digmma function can be computed by Taylor series expansion. Because of the page limits, we put the complete derivations of (13)-(15) in appendix A. Note that the optimal values of η_{dk} and φ_{kv} depend on γ_{idk}, and vice verse. Therefore, we can update the two sets of parameters alternatively until the approximated likelihood achieves a local maximum.

After we have computed the values of the variational parameters, we can then start optimizing over the original parameters α and β. Although the solutions for α and β do not have closed-forms as the variational parameters, their gradients and Hessian matrices are easily computable. Therefore, we can use gradient-based methods such as Newton-Raphson optimization method to find the optimal values.

The overall procedures of VB inference for LDA can be summarized by the following.

Algorithm 1: VB inference and parameter estimation for LDA

Input: A corpus of document **w** and initial model parameters α and β.

Output: Updated parameters α^* and β^* and variational parameters η^*, φ^* , and γ^*.

1 Initialize η, φ, and γ.
2 Update γ_{idk} by (15) for all i, d, and k.
3 Update η_{dk} by (13) for all d and k.
4 Update φ_{kv} by (14) for all k and v.
5 Repeat step 2-4 until converge and return current setting of η, φ, and γ.
6 Using Newton-Raphson method to update α and β.
7 Check if α and β converge; return to step 2 if not.
8 return the values of the parameters.

Given a new sets of documents $\mathbf{w}'$, we can infer the topic information we wants by running the above algorithm from step 1 to step 4.

5 Comparison of MCMC and VB Inference

In the previous two sections we have illustrated the MCMC method and the VB inference method to handle the problem of complex marginal distribution introduced by the coupling of the latent variables in LDA. In this section, we will further investigate the intuitions behind the two methods and compare the pros and cons of these two methods.

The way that the MCMC method avoids the intractability is by skipping away from the complex normalization factor of the posterior distribution $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$. By constructing the Markov chain using Gibbs sampling, the probability of drawing a sample $\mathbf{z}$ is proportional to the posterior probability $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$ and thus the average of the samples drawn by MCMC forms an unbiased estimator for the topic's actual assignment. Since the topic assignment drawn by MCMC is unbiased, the following inferences for θ and ϕ are also unbiased. However, the MCMC method also suffers some problems. For example, it is hard to find how many transitions are sufficient for the chain to mix to its stationary distribution (mixing time is uncertain). Also, the MCMC method suffers sampling noise by its intrinsic randomness. Also, because the MCMC method needs to store the stationary state for further inference of the new document, the size of storage grows linearly with the size of training corpus.

The VB inference, on the other hand, tries to maximize a tractable variational lower bound for the intractable marginal distribution. Although this idea is effective and can be implemented efficiently, the VB method suffers some intrinsic bias and may converges to local maximum. The amount of bias can be reduced by introducing a less constrained (more complex) form of variational distribution as in [2]. However, by doing so, the complexity of finding the variational parameters may increase a lot and further approximation may be needed [2]. As for the amount of storage the method needs after the training, it does not grow linearly as the size of training data as the MCMC method because the information of the training data is kept in the updated parameters of the model.

Methods / Items	MCMC	VB
Main Idea	Randomized algorithm that uses Gibbs sampling to draw samples with probability proportional to posterior of the latent variable.	Deterministic algorithm that maximizes a tractable lower bound of original marginal probability.
Advantages	• Unbiased estimation • No other assumptions	• Deterministic • Size of storage invariant over the size of the corpus
Disadvantages	• Uncertainty in mixing time. • Sampling noise • Size of storage	• Biased • Complexity and performance depends on the assumptions of the variational distribution

Figure 4: Comparisons of MCMC and VB.

6 Implementation

To verify our understanding of the LDA model, we implemented LDA with VB inference method presented in section 4. To test our implementation, we evaluated it the on the document classification task with a synthetic corpus. We compared the classification result of our implementation with that of GibbsLDA++ library [6] and found that the two methods have similar performances; showing the correctness of our implementation. The

details of the experiment are described in the following subsections and the process of our implementation can be illustrated as Figure 5.

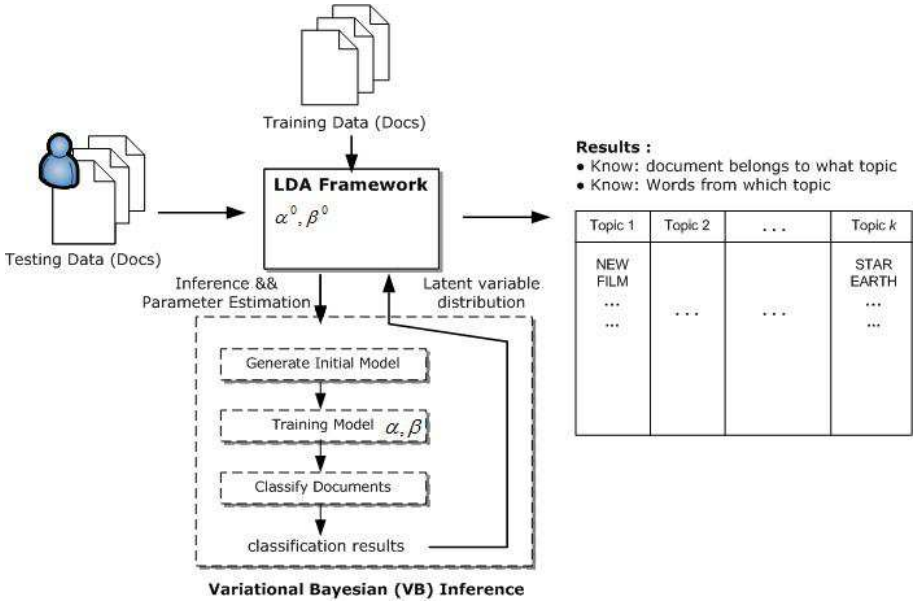

Figure 5: The process of project implementation.

6.1 Data Generation

The process we used for generating the synthetic corpus is close to LDA except that we used a fixed set of θ and ϕ values instead of using the Dirichlet distribution to dynamically generate those values. The corpus we used contains three possible topics and has a six-word vocabulary. The values we used for θ and ϕ are as follows:

$\theta_1 = [0.6\ 0.2\ 0.2]^{\mathrm{T}}$
$\theta_2 = [0.2\ 0.6\ 0.2]^{\mathrm{T}}$
$\theta_3 = [0.2\ 0.2\ 0.6]^{\mathrm{T}}$
$\phi_1 = [0.25\ 0.25\ 0.125\ 0.125\ 0.125\ 0.125]^{\mathrm{T}}$
$\phi_2 = [0.125\ 0.125\ 0.25\ 0.25\ 0.125\ 0.125]^{\mathrm{T}}$
$\phi_3 = [0.125\ 0.125\ 0.125\ 0.125\ 0.25\ 0.25]^{\mathrm{T}}$

For each document d, we first pick θ_d uniformly from the 3 choices shown above. For each word w_{id} in d, we first pick a topic assignment z_{id} with multinomial distribution θ_d and then pick a word value v with multinomial distribution $\theta_{z_{id}}$. Using this process, we generated 100 documents as the training set and 20 documents as the test set, with each document of 100 words in length. For each document d, we label this document as having topic k if θ_{dk} is of the greatest value in θ_d.

6.2 Experiments

For the VB implementation, we first initialize the model with $\alpha_k^0 = 0.5$ for all k and $\beta_{kv}^0 = \frac{N(k,v)}{\sum_v N(k,v')}$, where the function $N(k, v)$ denotes the number of words v in the document labeled with k. After we initialize the model, we

run the implemented training algorithm and keep the resulting model parameters α^* and β^*. After the training is complete, we then evaluate the model on a 20-document test set. The resulting classification accuracy is 80

For the MCMC method, because it is a randomized algorithm, we run the training scheme of GibbsLDA++ for 5 times. For each of the training repetitions, we run the inference scheme of GibbsLDA++[6]for 10 times, and resulted in totally 50 inference results. If we pick a random inference result and evaluate the document classification performance, the accuracy is around $60 - 70\%$. However, if we average the 50 inference results and use the averaged result to conduct classification, the accuracy increases to 85%. This phenomenon matches our expectation that average the inference results of different samples can effectively reduce sampling noise. Since the performance of our VB implementation is close to that of GibbsLDA++, we can have confidence in the correctness of our implementation. The experiment results are summarized in the following table (figure 6)

Items \ Methods	LDA + Variational Bayesian (VB) Inference	LDA + MCMC (GibbsLDA++) Single Run	LDA + MCMC (GibbsLDA++) Average of 50 Runs
Total Training Docs	100	100	100
Total Testing Docs	20	20	20
Successful Classification	16	13 ~ 14	17
Success Rate	80%	60% ~ 70%	85%

Figure 6: The comparison of LDA+VB and LDA+MCMC in document classification.

7 Discussion

In this project, we have studied a generative model, LDA, for collections of discrete data. Because LDA has complex coupling between its latent variables, the marginal distribution of observing a set of data is not tractable. According to such intractability, we also have studied and compared two approaches, MCMC and VB inference, that are commonly used in the field of machine learning to handle complex distributions. To further understand the effect of LDA with two different inference algorithms, we implement LDA model with Variational Bayesian (VB) Inference method to do document classification, and compare the result with running GibbsLDA++[6] on the same training and testing documents we generated for the experiments. Because the classification results of these two methods are similar, we can have confidence in the correctness of our implementation and that matches to our original expectation.

References

[1] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning*, Research 3, 993-1022, 2003.

[2] Y. W. Teh, D. Newman and M. Welling, "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation," *NIPS*, 2006.

[3] T. L. Griffiths, "Gibbs sampling in the generative model of Latent Dirichlet Allocation," Standford University.

[4] B. Walsh, "Markov Chain Monte Carlo and Gibbs Sampling," *Lecture Notes for EEB 581*, Columbia University, 2004.

[5] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *National Academy of Sciences*, 101. 5228-5235 (2004).

[6] X.-H. Phan, "GibbsLDA++," http://gibbslda.sourceforge.net/

[7] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," *Springer Series in Statistics.*

[8] B. de Finetti, *Theory of probability. Vol. 1-2.* John Wiley & Sons Ltd., Chichester, 1990. Reprint of the 1975 translation.

[9] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed Inference for Latent Dirichlet Allocation," *NIPS*, 2007.

APPENDIX

A Derivations of Variational Parameters

In this appendix we introduce how to derive (13), (14), and (15) in detail. The first step is to decompose the expectation terms in (12). To do so, we need some auxiliary tools. The first one is to compute the expectance $\mathrm{E}_q[\log(\theta_{dk})]$. Given that θ_d is Dirichlet distributed under η_d, we can compute $\mathrm{E}_q[\log(\theta_{dk})]$ by taking the advantage that the Dirchlet distribution belongs to the exponential family. More specifically, consider the following equities

$$\begin{aligned}\frac{\partial}{\partial\eta_{dk}}(\int_{\theta_d} p(\theta_d|\eta_d)d\theta_d) &= \textstyle\int_{\theta_d} \frac{\partial}{\partial\eta_{dk}} p(\theta_d|\eta_d)d\theta_d \\ &= \textstyle\int_{\theta_d} \frac{\partial}{\partial\eta_{dk}} \exp(\sum_{k=1}^K((\eta_{dk}-1)\log(\theta_{dk}) - \log(\Gamma(\eta_{dk}))) + \log(\Gamma(\eta_{d\cdot})))d\theta_d \\ &= \textstyle\int_{\theta_d} (\log(\theta_{dk}) - \frac{\partial}{\partial\eta_{dk}}\log(\Gamma(\eta_{dk})) + \frac{\partial}{\partial\eta_{dk}}\log(\Gamma(\eta_{d\cdot})))p(\theta_d|\eta_d)d\theta_d \\ &= \textstyle\frac{\partial}{\partial\eta_{dk}} 1 = 0.\end{aligned} \tag{16}$$

If we define the $\Psi(y) = \frac{\partial}{\partial y}\log(\Gamma(y))$, we can find that

$$\mathrm{E}_q[\log(\theta_{dk})] = \int_{\theta_d} \log(\theta_{dk})p(\theta_d|\eta_d)d\theta_{\mathbf{d}} = \Psi(\eta_{d\cdot}) - \Psi(\eta_{dk}) \tag{17}$$

by moving the other two terms in (16) to the other side of the equity. Following the same spirit, we can also have

$$\mathrm{E}_q[\log(\phi_{kv})] = \Psi(\varphi_{kv}) - \Psi(\varphi_{k\cdot}). \tag{18}$$

Another key is that

$$\mathrm{E}_q[n_{dk\cdot}\log(\theta_{dk})] = \mathrm{E}_q[n_{dk\cdot}]\mathrm{E}_q[\theta_{dk}], \tag{19}$$

because $\mathbf{z}$ and θ are conditional independent given that the variational distribution is of the form in (11). Similarly, we can also have $\mathrm{E}_q[n_{\cdot kv}\log(\phi_{kv})] = \mathrm{E}_q[n_{\cdot kv}]\mathrm{E}_q[\phi_{kv}]$. Using the two techniques above, we can then decompose each expectation terms in (12).

$$\begin{aligned}\mathrm{E}_q[\log(p(\theta|\alpha))] &= \textstyle\sum_{d=1}^D[\log(\Gamma(\alpha_\cdot)) - \sum_{k=1}^K\log(\alpha_k) + \sum_{k=1}^K((\alpha_k-1)\mathrm{E}_q[\log(\theta_{dk})])] \\ &= \textstyle\sum_{d=1}^D[\log(\Gamma(\alpha_\cdot)) - \sum_{k=1}^K\log(\alpha_k) + \sum_{k=1}^K((\alpha_k-1)(\Psi(\eta_{dk}) - \Psi(\eta_{d\cdot})))].\end{aligned} \tag{20}$$

$$\begin{aligned}\mathrm{E}_q[\log(p(\phi|\beta))] &= \textstyle\sum_{k=1}^K[\log(\Gamma(\beta_{k\cdot})) - \sum_{v=1}^V\log(\beta_{kv}) + \sum_{v=1}^V((\beta_{kv}-1)\mathrm{E}_q[\log(\phi_{kv})])] \\ &= \textstyle\sum_{k=1}^K[\log(\Gamma(\beta_{k\cdot})) - \sum_{v=1}^V\log(\beta_{kv}) + \sum_{v=1}^V((\beta_{kv}-1)(\Psi(\varphi_{kv}) - \Psi(\varphi_{k\cdot})))].\end{aligned} \tag{21}$$

$$\mathrm{E}_q[\log(p(\mathbf{z}|\theta))] = \sum_{d=1}^D\sum_{k=1}^K(\mathrm{E}_q[n_{dk\cdot}]\mathrm{E}_q[\log(\theta_{dk})]) = \sum_{d=1}^D\sum_{k=1}^K\gamma_{\cdot dk}(\Psi(\eta_{dk}) - \Psi(\eta_{d\cdot})). \tag{22}$$

$$\mathrm{E}_q[\log(p(\phi|\beta))] = \sum_{k=1}^{K}\sum_{v=1}^{V}\mathrm{E}_q[n_{\cdot kv}]\mathrm{E}_q[\log(\phi_{kv})] = \sum_{k=1}^{K}\sum_{v=1}^{V}(\sum_{d=1}^{D}\sum_{i=1}^{N_d}\delta(w_{id}=v)\gamma_{idk})(\Psi(\varphi_{kv})-\Psi(\varphi_{k\cdot})). \tag{23}$$

$$\mathrm{E}_q[\log(q(\theta|\eta))] = \sum_{d=1}^{D}[\log(\Gamma(\eta_{d\cdot}))-\sum_{k=1}^{K}\log(\Gamma(\eta_{dk}))+\sum_{k=1}^{K}(\eta_{dk}-1)(\Psi(\eta_{dk})-\Psi(\eta_{d\cdot}))]. \tag{24}$$

$$\mathrm{E}_q[\log(q(\mathbf{z}|\gamma))] = \sum_{d=1}^{D}\sum_{i=1}^{N_d}\mathrm{E}_q[\log(q(z_{id}|\gamma_i d))] = \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K}\gamma_{idk}\log(\gamma_{idk}). \tag{25}$$

$$\mathrm{E}_q[\log(q(\phi|\varphi))] = \sum_{k=1}^{K}\mathrm{E}_q[\log(q(\phi_k|\varphi_k))] = \sum_{k=1}^{K}[\log(\Gamma(\varphi_{k\cdot}))-\sum_{v=1}^{V}\log(\Gamma(\varphi_{kv}))+\sum_{v=1}^{V}(\varphi_{kv}-1)(\Psi(\varphi_{kv})-\Psi(\varphi_{k\cdot}))]. \tag{26}$$

After we decompose the expectation terms in (12) we can then pull out the terms in (12) that are related to certain variational parameter. More precisely, we are interested in the following three types of functions:

$$\ell_{[\eta_{dk}]} = (\Psi(\eta_{dk})-\Psi(\eta_{d\cdot}))(\alpha_k+\gamma_{\cdot dk}-\eta_{dk})-\log(\Gamma(\eta_{d\cdot}))+\log(\Gamma(\eta_{dk})). \tag{27}$$

$$\ell_{[\varphi_{kv}]} = (\Psi(\varphi_{kv})-\Psi(\varphi_{k\cdot}))(\beta_{kv}+(\sum_{d=1}^{D}\sum_{i=1}^{N_d}\delta(w_{id}=v)\gamma_{idk})-\varphi_{kv})-\log(\Gamma(\varphi_{k\cdot}))+\log(\Gamma(\varphi_{kv})). \tag{28}$$

$$\ell_{[\gamma_{idk}]} = \gamma_{idk}(\Psi(\eta_{dk})-\Psi(\eta_{d\cdot}))+\sum_{v=1}^{V}(\delta(w_{id}=v)\gamma_{idk}(\Psi(\varphi_{kv})-\Psi(\varphi_{k\cdot})))-\gamma_{idk}\log(\gamma_{idk}). \tag{29}$$

Note that the values of η_{dk} and φ_{kv} are unconstrained, while the value of γ_{idk} have to follow that $\sum_{k=1}^{K}\gamma_{idk}=1$. Therefore, when deriving the optimal values for η_{dk} and φ_{kv}, we can just set their partial derivatives to zero and solve the equations. For η_{dk}, we can have

$$\frac{\partial\ell_{[\eta_{dk}]}}{\partial\eta_{dk}} = (\Psi'(\eta_{dk})-\Psi'(\eta_{d\cdot}))(\alpha_k+\gamma_{\cdot dk}-\eta_{dk})-(\Psi(\eta_{dk})-\Psi(\eta_{d\cdot}))-\Psi(\eta_{d\cdot})+\Psi(\eta_{dk})=0. \tag{30}$$

Similarly, for φ_{kv}, we can also have

$$\frac{\partial\ell_{[\varphi_{kv}]}}{\partial\varphi_{kv}} = (\Psi'(\varphi_{kv})-\Psi'(\varphi_{k\cdot}))(\beta_{kv}+(\sum_{d=1}^{D}\sum_{i=1}^{N_d}\delta(w_{id}=v)\gamma_{idk})-\varphi_{kv})-(\Psi(\varphi_{kv})-\Psi(\varphi_{k\cdot}))-\Psi(\varphi_{k\cdot})+\Psi(\varphi_{kv})=0. \tag{31}$$

Solving the equations above, we can obtain (13) and (14). For γ_{idk}, we have to plug in a Lagurange Multiplier term, resulting in

$$\frac{\partial}{\partial\gamma_{idk}}(\ell_{[\gamma_{idk}]}+\lambda(\sum_{k=1}^{K}\gamma_{idk}-1)) = (\Psi(\eta_{dk})-\Psi(\eta_{d\cdot}))+(\Psi(\varphi_{kw_{id}})-\Psi(\varphi_{k\cdot}))-\log(\gamma_{idk})-1+\lambda=0. \tag{32}$$

Solving the above equation for different k, we can derive (15). Note that the $\Psi(\eta_{d\cdot})$ and term in (32) is common for all values of k, and therefore we don't have to write it explicitly in (15).