10-708: Probabilistic Graphical Models 10-708, Spring 2016

18 : Dirichlet Process and Dirichlet Process Mixtures

Lecturer: Matt Gormley *Scribes: Chiqun Zhang, Hsu-Chieh Hu*

1 Wrap up topic modeling

1.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. The plate diagram of LDA model is given in Figure 1.

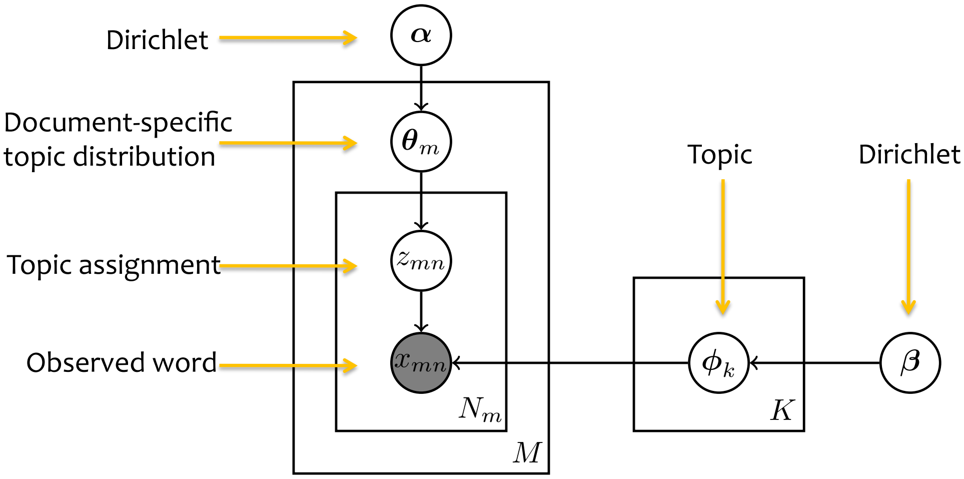


Figure 1: Plate diagram for LDA model

The generative story of LDA begins with only a Dirichlet prior over topics. Each topic is defined as a Multinomial distribution over the vocabulary, parameterized by ϕ_k. Since LDA is an unsupervised learning, a topic in LDA is visualized as its high probability words and a pedagogical label is used to identify the topic. The LDA can be decomposed into two part, one is the distributions over words and the other one is the distributions over topics. Before we step into the inference, it is natural to ask: "Is this a believable story for the generation of a corpus of documents?" or "Why might it work well anyway?"

The answer for this question is that LDA is a trading off two goals:

- For each document, allocate its words to as few topics as possible.
- For each topic, assign high probability to as few terms as possible.

Because putting a document in a single topic will require all of its words have probability under that topic, it will make the second goal hard. On the other hand, putting very few words in each topic will assign many topics to it to cover a document's words, which will make the first goal hard. LDA actually trades off there goals to find groups of tightly co-occurring words.

1.2 LDA Inference

The standard EM cannot be applied to LDA because there are α and β in LDA which are also latent. In addition, it is intractable to do the exact inference for all z, θ and ϕ. Because exact MAP inference in LDA is NP-hard for a large number of topics. For example, to compute the posterior in LDA, we can apply Junction tree algorithm. The Junction tree algorithm can be generalized into three steps:

- ‘Moralization” coverts the directed graph to undirected graph.
- “Triangulation” breaks the 4-cycles by adding edges.
- Cliques are arranged into a junction tree

For this algorithm, the time complexity is exponential in size of cliques, which will be the topics in LDA. Therefore, since LDA cliques will be large, at least $O(topics)$, the complexity is $O(2^{topics})$. Also, since the parameters are highly coupled, the sample method for all z, θ and ϕ is intractable either.

To handle this problem, we apply Collapsed Gibbs sampler method, whose general idea is given in Figure 2. In this method, the θ and ϕ are integrated out and the partially marginal distribution from α and β to z becomes Dirichlet Multinomial.

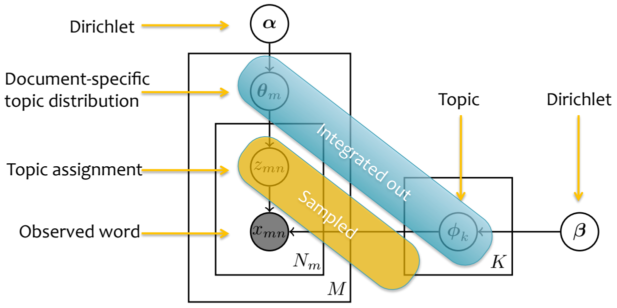


Figure 2: General illustration of Collapsed Gibbs sampler method

1.3 Gibbs sampling for LDA

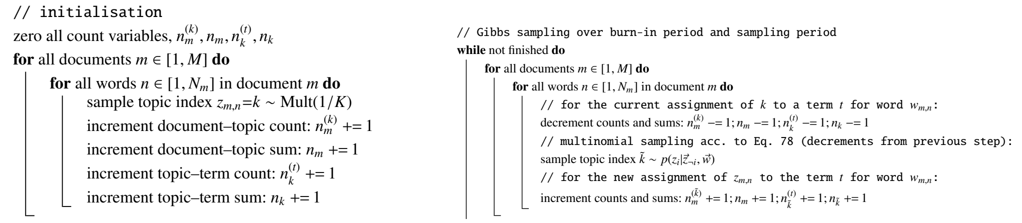
First, we need to derive the full conditionals:

$$
\begin{aligned}
p(z_i = k|Z^{-i}, X, \alpha, \beta) = \frac{p(X, Z|\alpha, \beta)}{p(X, Z^{-i}|\alpha, \beta)} \propto p(X, Z|\alpha, \beta) \\
= p(X|Z, \beta)p(Z|\alpha) = \int_{\Phi} p(X|Z, \Phi)p(\Phi|\beta)d\Phi \int_{\Theta} p(Z|\Theta)p(\Theta|\alpha)d\Theta \\
= \left(\prod_{k=1}^{K} \frac{B(n_k + \beta)}{B(\beta)}\right)\left(\prod_{m=1}^{M} \frac{B(n_m + \alpha)}{B(\alpha)}\right) \\
= \frac{n_{kt}^{-i} + \beta_t}{\sum_{v=1}^{T} n_{kv}^{-i} + \beta_v} \cdot \frac{n_m^{-i}k + \alpha_k}{\sum_{j=1}^{K} n_{mj}^{-i} + \alpha_j}
\end{aligned}
$$

where t, m are given by i. n_{kt} is the number of times topic k appears with type t and n_{mk} is the number of times topic k appears in document m.

A property for Gibbs sampling for LDA is that the Dirichlet is conjugate to the Multinomial. In LDA, we draw distribution over words from the Dirichlet distribution. Then we draw every word from the Multinomial distribution. Then the posterior of ϕ can be written as $p(\phi|X) = \frac{P(X|phi)p(\phi)}{p(X)}$, which turns out to be a Dirichlet distribution $Dir(\beta + n)$. Here the count vector n denoted the number of times every word appears.

The Gibbs sampling for LDA algorithm can be divided into two part, the initialization and the sampling.



```
// initialisation
zero all count variables, $n_m^{(k)}, n_m, n_k^{(t)}, n_k$
for all documents $m \in [1, M]$ do
    for all words $n \in [1, N_m]$ in document $m$ do
        sample topic index $z_{m,n}=k \sim \mathrm{Mult}(1/K)$
        increment document–topic count: $n_m^{(k)}$ += 1
        increment document–topic sum: $n_m$ += 1
        increment topic–term count: $n_k^{(t)}$ += 1
        increment topic–term sum: $n_k$ += 1
```

(a) The initialization part of Gibbs sampling algorithm.

```
// Gibbs sampling over burn-in period and sampling period
while not finished do
    for all documents $m \in [1, M]$ do
        for all words $n \in [1, N_m]$ in document $m$ do
            // for the current assignment of $k$ to a term $t$ for word $w_{m,n}$:
            decrement counts and sums: $n_m^{(k)}$ −= 1; $n_m$ −= 1; $n_k^{(t)}$ −= 1; $n_k$ −= 1
            // multinomial sampling acc. to Eq. 78 (decrements from previous step):
            sample topic index $\tilde{k} \sim p(z_i|\vec{z}_{\neg i}, \vec{w})$
            // for the new assignment of $z_{m,n}$ to the term $t$ for word $w_{m,n}$:
            increment counts and sums: $n_m^{(\tilde{k})}$ += 1; $n_m$ += 1; $n_{\tilde{k}}^{(t)}$ += 1; $n_{\tilde{k}}$ += 1
```

(b) The sampling part of Gibbs sampling algorithm.

Figure 3: The Gibbs sampling for LDA algorithm.

Also, we should recall that Gibbs sampling is a special case of Metropolis-Hastings method with a special proposal distribution, which ensures the hasting ratio is always 1.0.

1.4 Extensions of LDA

1.4.1 Correlated topic models

The Dirichlet is a distribution on the simplex, positive vectors that sum to 1. And it assumes that the components are nearly independent. However, in real data, an article about fossil fuels is more likely to also be about geology than about genetics. In correlated topic models, we apply the logistic normal distribution, which can model dependence between components (Aitchsion, 1980). In this distribution, the log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution.

$$X \sim N_k(\mu, \Sigma)$$
$$\theta_i \propto exp\{x_i\}$$

Figure 4 shows a plate diagram for the correlated topic model. The basic properties are: 1. Draw topic proportions from a logistic normal. This allows topic occurrences to exhibit correlation. 2. Provides a "map" of topics and how they are related. 3. Provides a better fit to text data, but computation is more complex. And we should notice that the correlated topic model doesn't use Dirichlet distribution, so it is not conjugate any more.

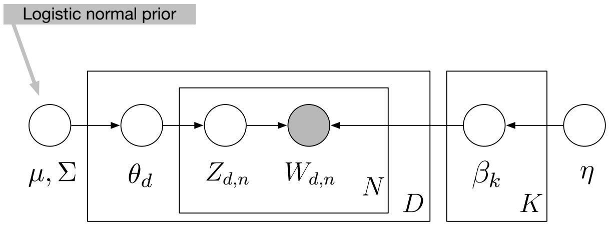


Figure 4: Plate diagram for correlated topic models.

1.4.2 Dynamic topic models

In a dynamic topic models, we also consider the time evolution effect on the learning model. For example, in a document classification learning problem, in dynamic topic model, the documents are divided up by year. For each year, we start with a separate topic model and then add a dependence of each year on the previous one. Figure 5 shows the plate diagram of dynamic topic model. Recall that LDA assumes that the order of the documents does not matter, but this assumption is not appropriate for sequential corpora. In addition, we may also want to track how language changes over time. In dynamic topic model, the topics are allowed to drift in a sequence.

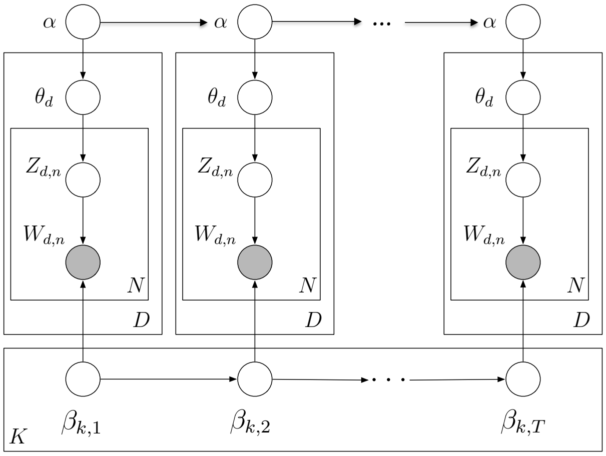


Figure 5: Plate diagram for dynamic topic models.

1.4.3 Polylingual topic model

Polylingual topic model is an extension of latent Dirichlet allocation (LDA) for modeling polylingual document tuples. Each tuple is a set of documents that are loosely equivalent to each other, but written in different languages. In this model, the data is comparable versions of each document exist in multiple languages, for example, the Wikipedia article for "Barak Obama" in twelve languages. The polyingual topic model is very similar to LDA, except that the topic assignments, z, and words, w, are sampled separately for each language. Figure 6 shows the plate diagram of the Polylingual topic model.

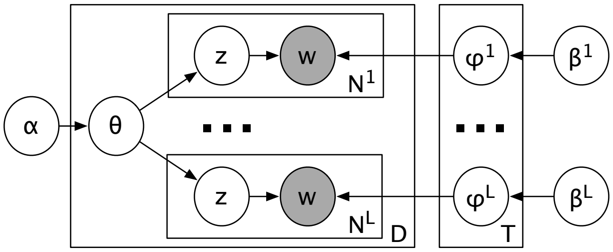


Figure 6: Plate diagram for Polylingual topic models.

1.4.4 Supervised LDA

LDA is an unsupervised model, but many data are paired with response variables. For example, user reviews can be paired with a number of stars; web pages can be paired with a number of "likes"; documents can be paired with links to other documents; or images can be paired with a category. The supervised LDA are topic models of documents and responses. It can fit to find topics predictive of the response. In supervised LDA, we add to LDA a response variable associated with each document. Then we jointly model the documents and the responses, in order to find latent topics that will best predict the response variables for future unlabeled documents. Figure 7 shows the plate diagram for this model.

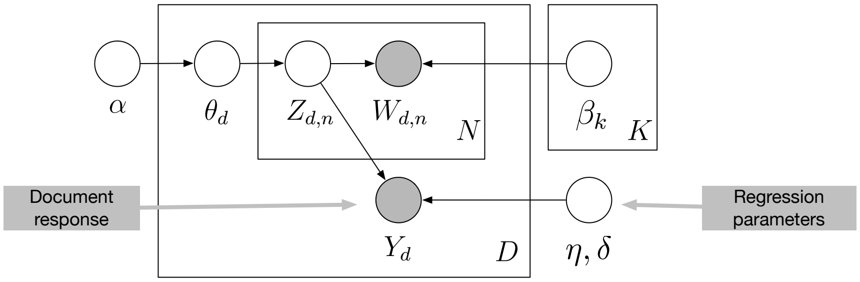


Figure 7: Plate diagram for Supervised LDA.

2 Dirichlet Process and Dirichlet Process Mixtures

2.1 Introduction

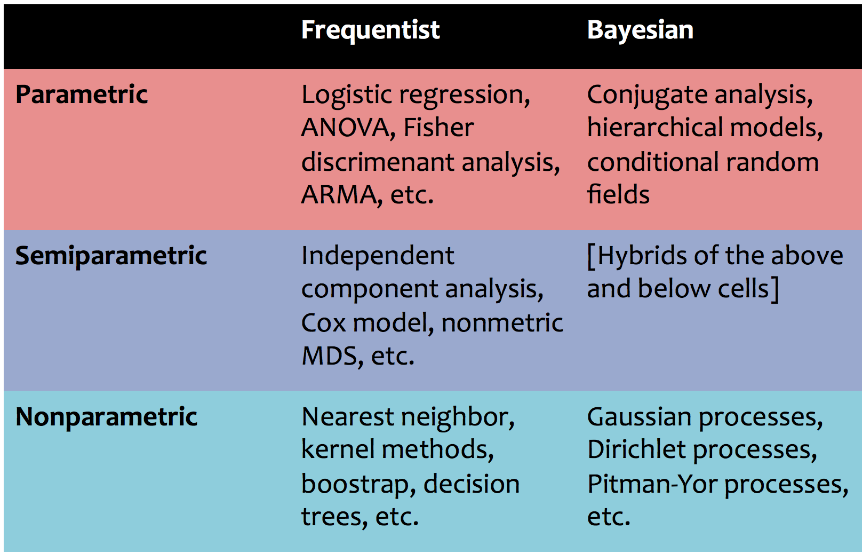
In parametric modeling, it is assumed that data can be represented by models using a fixed, finite number of parameters. Examples of parametric models include clusters of K Gaussians and polynomial regression models. In many problems, determining the number of parameters a priori is difficult; for example, selecting the number of clusters in a cluster model, the number of segments in an image segmentation problem, the number of chains in a hidden Markov model, or the number of topics in a topic modelling problem before the data is seen can be problematic.

In nonparametric modeling, the number of parameters is not fixed, and often grows with the sample size. Kernel density estimation is an example of a nonparametric model. In Bayesian nonparametrics, the number of parameters is itself considered to be a random variable. One example is to do clustering with k-means (or mixture of Gassuians) while the number of clusters k is unknown. Bayesian inference addresses this problem by treating k itself as a random variable. A prior is defined over an infinite dimensional model space, and inference is done to select the number of parameters. Such models have infinite capacity, in that they include an infinite number of parameters a priori; however, given finite data, only a finite set of these parameters will be used. Unused parameters will be integrated out.

2.2 Parametric vs. Nonparametric

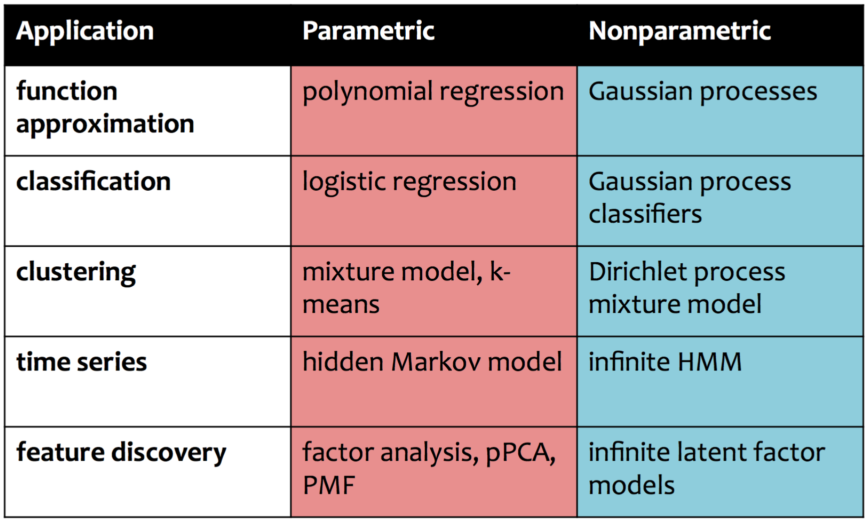
- Parametric models:
  - Finite and fixed number of parameters
  - Number of parameters is independent of the dataset
- Nonparametric models:
  - Have parameters (”infinite dimensional” would be a better name)
  - Can be understood as having an infinite number of parameters
  - Can be understood as having a random number of parameters

  - Number of parameters can grow with the dataset
- Semiparametric models:
  - Have a parametric component and a nonparametric component



	Frequentist	Bayesian
Parametric	Logistic regression, ANOVA, Fisher discrimenant analysis, ARMA, etc.	Conjugate analysis, hierarchical models, conditional random fields
Semiparametric	Independent component analysis, Cox model, nonmetric MDS, etc.	[Hybrids of the above and below cells]
Nonparametric	Nearest neighbor, kernel methods, boostrap, decision trees, etc.	Gaussian processes, Dirichlet processes, Pitman-Yor processes, etc.

Figure 8: Frequentist and Bayesian methods for Parametric and Nonparametric



Application	Parametric	Nonparametric
function approximation	polynomial regression	Gaussian processes
classification	logistic regression	Gaussian process classifiers
clustering	mixture model, k-means	Dirichlet process mixture model
time series	hidden Markov model	infinite HMM
feature discovery	factor analysis, pPCA, PMF	infinite latent factor models

Figure 9: Different applications for Parametric and Nonparametric

Definition: a model of a collection of distributions

$$\{p_\theta : \theta \in \Theta\} \tag{1}$$

parametric model: the parameter vector is finite dimensional

$$\Theta \subset \mathcal{R}^k \tag{2}$$

nonparametric model: the parameters are from a possibly infinite dimensional space

$$\Theta \subset \mathcal{F} \tag{3}$$

2.3 Motivations

Model selection is an operation that is fraught with difficulties, whether we use cross validation or marginal probabilities as the basis for selection. The Bayesian nonparametric approach is an alternative to parametric modeling and selection. There are two motivations for Bayesian nonparametric models:

In clustering, the actual number of clusters used to model data is not fixed, and can be automatically inferred from data using the usual Bayesian posterior inference framework. The equivalent operation for finite mixture models would be model averaging or model selection for the appropriate number of components, an approach which is fraught with difficulties. Thus infinite mixture models as exemplified by DP mixture models provide a compelling alternative to the traditional finite mixture model paradigm.

In density estimation, we are interested in modeling the density from which a given set of data is drawn. To avoid limiting ourselves to any parametric class, we may again use a nonparametric prior over an infinite set of distributions.

2.4 Exchangability and de Finetti's Theorem

Definition1: a joint probability distribution is exchangeable if it is invariant to permutation.

Definition2: The possibly infinite sequence of random variables $(X_1, X_2, X_3, \dots)$ is exchangeable if for any finite permutations of the indices $(1, 2, \dots n)$:

$$p(X_1, X_2, \dots, X_n) = p(X_{s(1)}, X_{s(2)}, \dots, X_{s(n)}) \tag{4}$$

The meaning of exchangability is different from independent and identical distributed (i.i.d.). Exchangability means that it does not matter if data is reordered. The de Finetti's theorem states that if $(X_1, X_2, X_3, \dots)$ is infinitely exchangeable. The joint distribution has representation as mixture:

$$p(x_1, x_2, \dots, x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) dP(\theta) \tag{5}$$

2.5 Chinese Restaurant Process (CRP)

The distribution over partitions can be described in terms of the following restaurant metaphor of Figure 10. We assume that a Chinese restaurant has infinite tables, each of which can seat infinite customers. In addition, there is only one dish on each table.

The first customer enters the restaurant and sits at the first table. The second customer enters and decides either to sit with the first customer, or choose a new table. In the general case, the $n+1$st customer either choose to join an already occupied table k with the probability propotional to the number of customers n_k already sitting there, or sit at a new table with the probability propotional to α. In this metaphor, customers are identified with the intergers 1,2,3,... and tables as clusters. When all the n customers have sat down the tables, they are partitioned into clusters, which exhibits the clustering property of the Dirichlet process above.

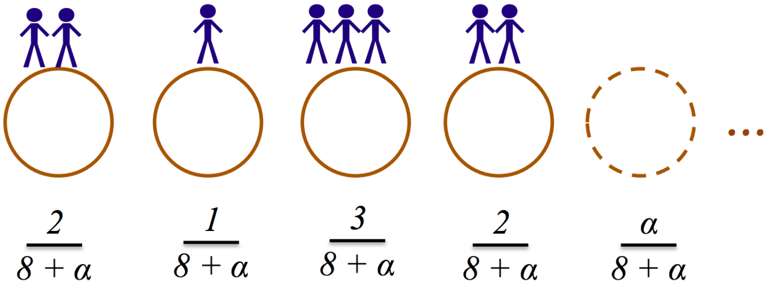


Figure 10: Chinese Restaurant Process