Large Sample Asymptotic for Nonparametric Mixture Model with Count Data

Vu Nguyen†, Dinh Phung†, Trung Le‡, Svetha Venkatesh†
†Deakin University, Australia
{v.nguyen,dinh.phung,svetha.venkatesh}@deakin.edu.au
‡HCMc University of Pedagogy, Vietnam
trunglm@hcmup.edu.vn

Abstract

Bayesian nonparametric models have become popular recently due to its flexibility in identifying the unknown number of clusters. However, the flexibility comes at a cost for learning. Thus, Small Variance Asymptotic (SVA) is one of the promising approach for scalability in Bayesian nonparametric models. SVA approach for count data is also developed in which the likelihood function is replaced by the Kullback–Leibler divergence. In this paper, we present the Large Sample Asymptotic for count data when the number of sample in Multinomial distribution goes to infinity, we derive the similar result to SVA for scalable clustering.

1 Introduction

Traditional clustering algorithms often require an explicit choice of the model size in advance. For example, in K-means clustering, the number of clusters must be selected a priori even though this quantity is not known for many practical applications. Therefore, instead of performing model selection, Bayesian nonparametric (BNP) models [5,1,17] emerged as a promising approach to infer the complexity from the data directly. This flexibility allows Bayesian nonparametric models being able to identify the suitable number of clusters. However, the flexibility of BNP comes at a cost: training BNP models on massive data sets is notoriously challenging.

To address the scalability problem of BNP, a recent thread of research, namely Small Variance Asymptotic (SVA) of BNP model has gained much attention [9,8]. Small Variance Asymptotic for BNP [9] provide scalability, but still maintain the main properties of Bayesian nonparametric modeling. For generic case of distributions (non Gaussian case), the asymptotic extension is proposed in [8]. Then, the recent work has exploited this scalable approach for various tasks [15,18,19,7,11].

In this paper, we present the Large Sample Asymptotic (LSA) for count data in which we let the number of samples (or the number of time tossing a dice in Multinomial distribution) goes to infinity, then we derive the hard clustering assignment for Bayesian nonparametric model. Our proposed analysis can be seen the alternative view to the small variance asymptotic for count data [8].

2 Small Variance Asymptotic

Recent works of Small Variance Asymptotic (SVA) are motivated by the connection between K-means and Gaussian Mixture Model (GMM): as the variances of Gaussian goes to zero, the GMM becomes K-means [2,9]. The asymptotic derivation to DPM and HDP are introduced in [9], opening the line of work in BNP for scalability.

The data type is not restricted to Gaussian case, but it can be generic, such as count data. For non-Gaussian cases of distributions, SVA derivation is proposed in [8] which is suitable for discrete-data.
Specifically, the SVA for exponential family distributions is presented [8] as follows:

\[ p(z_t = k |) = \left\{ \begin{array}{ll}
D_\phi (x_t, \mu_k) & \text{used } k \\
\lambda & \text{new } k
\end{array} \right. \]

where \( D_\phi \) denotes for the Bregman divergence for the likelihood function, such as KL divergence for Multinomial distribution.


### 3 Large Sample Asymptotic for Count Data

We manipulate the Multinomial likelihood in the Kullback–Leibler divergence form which also involve the number of sample \( n \), then we derive the hard assignment in the limit as \( n \rightarrow \infty \).

#### 3.1 Large Sample Asymptotic

We have a data point \( x = (x_1, ..., x_D) \in \mathcal{R}^D \), where each element \( x_i \) is a count in \( D \) bins, and the Multinomial parameter \( \phi = (\phi_1, ..., \phi_D) \in \mathcal{R}^D \) such that \( \sum_{d=1}^{D} \phi_d = 1 \). To express the probability of observation (or a histogram) \( x \) given a parameter \( \phi \), the Multinomial likelihood is defined as:

\[ p(x | \phi) = \frac{n!}{\prod_{d=1}^{D} x_d! \prod_{d=1}^{D} \phi_d} \]

where \( n = \sum_{d=1}^{D} x_d \) is a number of samples or trials.

As discussed in [16] that independent observations constituting a histogram are multiplied together to recover the joint probability of all measurements. Thus, an invariant likelihood across histogram counts is the geometric mean of the Multinomial likelihood \( p(x | \phi)^\frac{1}{n} \). We term this quantity as the average Multinomial likelihood and define the average log-likelihood \( \bar{L} \equiv \log p(x | \phi)^\frac{1}{n} \).

\[ \bar{L} = \frac{1}{n} \log n! - \frac{1}{n} \sum_{d=1}^{D} \log x_d! + \frac{1}{n} \sum_{d=1}^{D} \frac{x_d}{n} \log \phi_d. \]

Using Stirling’s approximation [4], we have the expansion of \( \log n! = n \log n - n + O(\log n) \). We apply Stirling’s approximation into Eq. [2]

\[ L = \log n - \sum_{d=1}^{D} \frac{x_d}{n} \log x_d + \sum_{d=1}^{D} \frac{x_d}{n} \log \phi_d + \frac{1}{n} \left( O(\log n) + \sum_{d=1}^{D} O(\log x_d) \right). \]

As we have \( \log n = \sum_{d=1}^{D} \frac{x_d}{n} \log n \), then the above formula becomes

\[ \bar{L} = -\sum_{d=1}^{D} \frac{x_d}{n} \log \frac{x_d}{n} + \frac{1}{n} \left( O(\log n) + \sum_{d=1}^{D} O(\log x_d) \right). \]

The normalized histogram can be viewed as a probability distribution \( \hat{x} = (\hat{x}_1, ..., \hat{x}_D) = (\frac{x_1}{n}, ..., \frac{x_D}{n}) \) and substituted accordingly.

\[ \bar{L} = -D_{KL}(\hat{x}||\phi) + \frac{1}{n} \left( O(\log n) + \sum_{d=1}^{D} O(\log x_d) \right). \]

We have defined in Eq. [2] as \( \bar{L} = \log p(x | \phi)^\frac{1}{n} \), thus equivalently we obtain:

\[ p(x | \phi) = \exp \left\{ -nD_{KL}(\hat{x}||\phi) + O(\log n) + \sum_{d=1}^{D} O(\log x_d) \right\}. \]

(3)
We have presented the Multinomial likelihood using the KL divergence and the number of samples (or trials) \( n \). In the following section, we utilize the above form into Gibbs sampling form for Dirichlet Process Mixture. Then we let the number of samples \( n \) goes to infinity to obtain the Large Sample Asymptotic in which the probabilistic inference becomes the hard clustering for scalability.

### 3.2 Scalable Clustering for Dirichlet Process Mixture using Large Sample Asymptotic

We now present the hard clustering algorithm for Dirichlet Process Mixture (11) using the Large Sample Asymptotic on count data. Let \( \{ x_i, z_i \}_{i=1}^{N} \) be a collection of data observation \( x_i \) and the corresponding latent assignment \( z_i \). Let \( K \) be the number of active clusters, \( \phi = \{ \phi_1, ..., \phi_K \} \) be the parameter representing for each cluster in DPM. Let \( G \) be the prior distribution for generating the observation \( x \). We consider the sampling \( z_i \) conditional on other variables:

\[
p(z_i = k \mid z_{-i}, x_i, \phi, \alpha, G) \propto \left\{ \frac{N_k \times p(x_i \mid \phi_{z_i})}{\alpha \times \int \phi \, dG(\phi)} \right\} \quad \text{used } k
\]

where \( N_k \) is the number of data points in component \( k \), \( N = \sum_{k=1}^{K} N_k \) is the total number of data points, and \( \alpha \) is the concentration parameter.

By an abuse of notation, we are using \( N_k \) to denote the number of data points in cluster \( k \) in DPM and \( n \) (lower-case) is the number of trial in Multinomial distribution that will be later assumed to go to infinity.

**Assigning \( z_i \) to used cluster \( k \).** The Multinomial probability of \( p(x_i \mid \phi_{z_i}) \) is described in Eq. 3

\[
p(x_i \mid \phi_{z_i}) = \exp \left\{ -n D_{KL}(\hat{x}_i || \phi_{z_i}) + T \right\}
\]

where \( \hat{x}_i = \frac{x_{id}}{n} \) as a normalized histogram for \( x_i \) and \( n = \sum_{d=1}^{D} x_{id} \) is the number of trial, assumed to be the same for all data points. \( n \) will later go to infinity in our analysis.

**Assigning \( z_i \) to new cluster \( k \).**

\[
p(z_i = k^{\text{new}} \mid x_i, G) \propto p(z_i = k^{\text{new}} \mid z_{-i}, \alpha) \times p(x_i \mid z_i = k^{\text{new}}, G)
\]

\[
= \alpha \times \int \phi \, dG(\phi)
\]

\[
= \alpha \times \frac{n!}{\prod_{d=1}^{D} x_{id}!} \times \frac{\Gamma \left( \sum_{d=1}^{D} \gamma_d \right)}{\Gamma \left( \sum_{d=1}^{D} [\gamma_d + x_{id}] \right)} \times \prod_{d=1}^{D} \frac{\Gamma(x_{id} + \gamma_d)}{\Gamma(\gamma_d)}.
\]

where we compute the term \( p(x_i \mid z_i = k^{\text{new}}, G) = \int \phi \, dG(\phi) \) using Multinomial Dirichlet conjugacy.

The geometric mean is expressed \( \bar{L} = \log p(x_i \mid z_i = k^{\text{new}}, G) \) (as in Sec. 3.1):

\[
\bar{L} = \frac{1}{n} \log n! + \frac{1}{n} \left\{ \log \Gamma \left( \sum_{d=1}^{D} \gamma_d \right) - \log \Gamma \left( \sum_{d=1}^{D} [\gamma_d + x_{id}] \right) + \sum_{d=1}^{D} \log \frac{x_{id} + \gamma_d}{x_{id} \gamma_d} \right\}
\]

where we have defined that \( C(x_i) \) is a function of \( x_i \) such that \( \log C(x_i) = \log \Gamma \left( \sum_{d=1}^{D} \gamma_d \right) - \log \Gamma \left( \sum_{d=1}^{D} [\gamma_d + x_{id}] \right) + \sum_{d=1}^{D} \log \frac{x_{id} + \gamma_d}{x_{id} \gamma_d} \) where \( \gamma_d \) is a Dirichlet symmetric (provided and fixed). We note that \( C(x_i) \) will result in a finite constant given \( x_i \).

Using Stirling approximation \( \log n! = n \log n - n + O(\log n) \) and canceling the common factors, we obtain

\[
\bar{L} = \log n + \frac{1}{n} \log C(x_i) - 1 + \frac{1}{n} O(\log n).
\]

Substituting Eq. 8 back to \( \frac{1}{n} \log p(x_i \mid z_i = k^{\text{new}}, G) = \bar{L} \), we obtain:

\[
p(x_i \mid z_i = k^{\text{new}}, G) = C(x_i) \times \exp \left( -n + n \log n \right) \times \exp \left[ O(\log n) \right].
\]
Then, the Eq. 6 becomes
\[ p(z_i = k_{\text{new}} | x_i, G) \propto \alpha \times C(x_i) \times \exp(-n + n \log n) \times \exp[O(\log n)]. \] (10)

In order to obtain non-trivial assignments, we must let \( \alpha \) be a function of \( n \) (but independent from the data) as \( \alpha = \exp(n - n \log n - n \lambda) \) for some \( \lambda \). Then, we plug \( \alpha \) with the new term into Eq. 10, we get the following:
\[ p(z_i = k_{\text{new}} | x_i, G) = C(x_i) \times \exp(-n \lambda) \times \exp[O(\log n)]. \] (11)

Substituting Eq. 5 and Eq. 11 into the Eq. 4 for sampling \( z_i \), we obtain the following probabilities to be used during Gibbs sampling.
\[ \hat{p}(z_i = k) = \frac{N_i \exp \left\{ -nD_{KL}(\hat{x}_i||\pi_k) + \sum_{d=1}^{D} O(\log x_{id}) \right\}}{C(x_i) \exp(-n\lambda \log K) + \sum_{k=1}^{K} \exp[O(\log n)]} 1 \leq k \leq K \] (12)
\[ \hat{p}(z_i = k_{\text{new}}) = \frac{C(x_i) \exp(-n\lambda)}{C(x_i) \exp(-n\lambda \log K) + \sum_{k=1}^{K} \exp[O(\log n)]} \] (13)

where we denote \( A_k = N_i \times \exp \left\{ -nD_{KL}(\hat{x}_i||\phi_k) + \sum_{d=1}^{D} O(\log x_{id}) \right\} \) and \( C(x_i) \) will result in a finite constant given \( x_i \) that \( \log C(x_i) = \log \Gamma(\sum_{d=1}^{D} \gamma_d) - \log \Gamma(\sum_{d=1}^{D} \gamma_d + x_{id}) + \sum_{d=1}^{D} \log x_{id} \).

All of the above probabilities will become binary when \( n \to \infty \). More specifically, all of the \( k + 1 \) values will be increasingly dominated by the smallest value of \( \{D_{KL}(\hat{x}_i||\phi_1), ...D_{KL}(\hat{x}_i||\phi_K), \lambda \} \). In other word, as \( n \to \infty \), only the smallest of these value will receive a non-zero probability. The data point \( x_i \) will be assigned to the nearest cluster with a divergence at most \( \lambda \). If the closest mean has a divergence greater than \( \lambda \), we will start a new cluster containing only \( x_i \).

Finally, we obtain the scalable inference in Dirichlet Process Mixture using Large Sample Asymptotic. When the number of samples (or trials) in Multinomial distribution goes to infinity, we obtain the hard assignment as follows:
\[ \lim_{n \to \infty} \hat{p}(z_i = k) = \begin{cases} D_{KL}(\hat{x}_i||\phi_k) & \text{used } k \\ \lambda & \text{new } k. \end{cases} \]

We note that our hard assignment of LSA for DPM (when \( n \to \infty \)) is interestingly similar to the SVA for DPM (when \( \sigma \to 0 \)\). To select \( \lambda \), we can use the farthest-first heuristic \( \bar{9} \) (given the expected number of cluster \( K \)) in a cross validation set.

### 4 Experiments

We present an evaluation of DPM using Large Sample Asymptotic (DPM-LSA) on image clustering using NUS WIDE 13 animal dataset (obtained from http://www.ml-thu.net/~jun/data/). There are 13 categories of the animals with 3411 images. We use SIFT \( \bar{13} \) as a feature. Each SIFT vector (500 dimensions) is assumed to follow Multinomial distribution \( \bar{14} \).

We compare our DPM-LSA (\( \lambda = 1.73 \)) with Affinity Propagation (AP) \( \bar{6} \), K-means, Gaussian Mixture Model, Dirichlet Process Mixture \( \bar{11} \) using collapsed Gibbs inference \( \bar{12} \), and DPmeans \( \bar{9} \) (\( \lambda = 5960 \)) on image clustering task using Matlab environment.
References


