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:

$$p(z_i = k \mid) = \begin{cases} D_\phi(x_i, \mu_k) & \text{used } k \\ \lambda & \text{new } k \end{cases}$$

where D_ϕ denotes for the Bregman divergence for the likelihood function, such as KL divergence for Multinomial distribution.

Machine learning practitioners have widely applied SVA to approximately estimate Multinomial distribution using KL divergence [10] for scalability. The asymptotic work of (infinite) HMM [15] and Dependent DPM [3] offer scalable analysis for sequential data. DP-space [19] is for scalable subspace clustering, JUMP-means [7] is for scalable Markov Jump Processes, and Lee et al [11] is for Bayesian hierarchical clustering.

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 $n \to \infty$.

3.1 Large Sample Asymptotic

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

$$p(x \mid \phi) = \frac{n!}{\prod_{d=1}^D x_d!} \prod_{d=1}^D \phi_d^{x_d} \tag{1}$$

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 $p(x \mid \phi)^{\frac{1}{n}}$. We term this quantity as the average Multinomial likelihood and define the average log-likelihood $\bar{L} \equiv \log p(x \mid \phi)^{\frac{1}{n}}$.

$$\bar{L} = \frac{1}{n} \log n! - \frac{1}{n} \sum_{d=1}^D \log x_d! + \sum_{d=1}^D \frac{x_d}{n} \log \phi_d. \tag{2}$$

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.

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

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

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

The normalized histogram can be viewed as a probability distribution $\hat{x} = (\hat{x}_1, ...\hat{x}_D) = \left(\frac{x_1}{n}, ...\frac{x_D}{n}\right)$ 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 $\bar{L} = \log p(x \mid \phi)^{\frac{1}{n}}$, thus equivalently we obtain:

$$p(x \mid \phi) = \exp \left\{ -n D_{KL}(\hat{x} || \phi) + O(\log n) + \left[\sum_{d=1}^D O(\log x_d) \right] \right\}. \tag{3}$$

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 [1] 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:

$$p(z_i = k \mid z_{-i}, x_i, \phi, \alpha, G) \propto \begin{cases} N_k \times p\left(x_i \mid \phi_{z_i}\right) & \text{used } k \\ \alpha \times \int_\phi p(x_i \mid \phi)\, dG(\phi) & \text{new } k \end{cases} \tag{4}$$

where N_k is the number of data point in component k, $N = \sum_{k=1}^K N_k$ is the total number of data points, and α 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 $p\left(x_i \mid \phi_{z_i}\right)$ is described in Eq. 3:

$$p\left(x_i \mid \phi_{z_i}\right) = \exp\left\{-nD_{KL}\left(\hat{x}_i || \phi_{z_i}\right) + T\right\} \tag{5}$$

where $\hat{x}_i = [\frac{x_{i1}}{n}, ..., \frac{x_{id}}{n}]$ 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.

$$\begin{aligned} p(z_i = k^{\text{new}} \mid x_i, G) \propto & p(z_i = k^{\text{new}} \mid z_{-i}, \alpha) \times p(x_i \mid z_i = k^{\text{new}}, G) \\ = & \alpha \times \int_\phi p(x_i \mid \phi)\, dG(\phi) \\ = & \alpha \times \frac{n!}{\prod_{d=1}^D x_{id}!} \times \frac{\Gamma\left(\sum_{d=1}^D \gamma_d\right)}{\Gamma\left(\sum_{d=1}^D [\gamma_d + x_{id}]\right)} \times \prod_{d=1}^D \frac{\Gamma(x_{id} + \gamma_d)}{\Gamma(\gamma_d)}. \end{aligned} \tag{6}$$

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

The geometric mean is expressed $\bar{L} = \log p(x_i \mid z_i = k^{\text{new}}, G)^{\frac{1}{n}}$ (as in Sec. 3.1):

$$\bar{L} = \frac{1}{n}\log n! + \frac{1}{n} \underbrace{\left\{ \log\Gamma\left(\sum_{d=1}^D \gamma_d\right) - \log\Gamma\left(\sum_{d=1}^D [\gamma_d + x_{id}]\right) + \sum_{d=1}^D \log\frac{x_{id} + \gamma_d}{x_{id}\gamma_d} \right\}}_{\log C(x_i)} \tag{7}$$

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 γ_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 [4] $\log n! = n\log n - n + O(\log n)$ and canceling the common factors, we obtain

$$\bar{L} = \log n + \frac{1}{n}\log C(x_i) - 1 + \frac{1}{n} O(\log n). \tag{8}$$

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

$$p(x_i \mid z_i = k^{\text{new}}, G) = C(x_i) \times \exp(-n + n\log n) \times \exp[O(\log n)]. \tag{9}$$

Table 1: Image clustering comparison on NUS WIDE dataset. Number of cluster in K-means ranges from $K = 5$ to 30, then we report the mean and standard deviation. Time is recorded in second unit.

Approach	AP [6]	K-means	GMM	DPM [1]	DPmeans [9]	DPM-LSA
# Cluster	K=18	K=5-30	K=5-30	K=17	K=17	K=19
NMI	0.166	0.19(.01)	0.19(.01)	**0.188**	0.161	*0.174*
Fscore	0.145	0.16(.01)	0.16(.01)	0.173	0.166	**0.184**
Time	167.8	14	15	1200	38	39.8

Then, the Eq. 6 becomes

$$p\left(z_i = k^{\text{new}} \mid x_i, G\right) \propto \alpha \times C\left(x_i\right) \times \exp\left(-n + n\log n\right) \times \exp\left[O\left(\log n\right)\right]. \tag{10}$$

In order to obtain non-trivial assignments, we must let α be a function of n (but independent from the data) as $\alpha = \exp\left(n - n\log n - n\lambda\right)$ for some λ. Then, we plug α with the new term into Eq. 10, we get the following:

$$p\left(z_i = k^{\text{new}} \mid x_i, G\right) = C\left(x_i\right) \times \exp\left(-n\lambda\right) \times \exp\left[O\left(\log n\right)\right]. \tag{11}$$

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.

$$\hat{\gamma}(z_i = k) = \frac{N_k \exp\left\{-nD_{KL}\left(\hat{x}_i || \pi_{z_i}\right) + \sum_{d=1}^{D} O\left(\log x_{id}\right)\right\}}{C\left(x_i\right)\exp\left(-n\lambda\log K\right) + \sum_{u=1}^{K} A_u \exp\left[O\left(\log n\right)\right]} \quad 1 \le k \le K \tag{12}$$

$$\hat{\gamma}(z_i = k^{\text{new}}) = \frac{C\left(x_i\right) \times \exp\left(-n\lambda\right)}{C\left(x_i\right)\exp\left(-n\lambda\log K\right) + \sum_{u=1}^{K} A_u \exp\left[O\left(\log n\right)\right]} \tag{13}$$

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

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 $\left\{D_{KL}\left(\hat{x}_i || \phi_1\right), \ldots D_{KL}\left(\hat{x}_i || \phi_K\right), \lambda\right\}$. 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 λ. If the closest mean has a divergence greater than λ, 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:

$$\lim_{n\to\infty} \hat{\gamma}(z_i = k) = \begin{cases} D_{KL}\left(\hat{x}_i || \phi_k\right) & \text{used } k \\ \lambda & \text{new } k. \end{cases}$$

We note that our hard assignment of LSA for DPM (when $n \to \infty$) is interestingly similar to the SVA for DPM (when $\sigma \to 0$)[8]. To select λ, we can use the farthest-first heuristic [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 [13] as a feature. Each SIFT vector (500 dimensions) is assumed to follow Multinomial distribution [14].

We compare our DPM-LSA ($\lambda = 1.73$) with Affinity Propagation (AP) [6], K-means, Gaussian Mixture Model, Dirichlet Process Mixture [1] using collapsed Gibbs inference [12], and DPmeans [9] ($\lambda = 5960$) on image clustering task using Matlab environment.

References

[1] C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[2] Christopher M Bishop. *Pattern recognition and machine learning*. springer New York, 2006.

[3] Trevor Campbell, Miao Liu, Brian Kulis, Jonathan P How, and Lawrence Carin. Dynamic clustering via asymptotics of the dependent dirichlet process mixture. In *Advances in Neural Information Processing Systems*, 2013.

[4] Jacques Dutka. The early history of the factorial function. *Archive for history of exact sciences*, 43(3):225–249, 1991.

[5] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[6] B.J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, February 2007.

[7] Jonathan H Huggins, Karthik Narasimhan, Ardavan Saeedi, and Vikash K Mansinghka. Jump-means: Small-variance asymptotics for markov jump processes. In *International Conference on Machine Learning (ICML)*, 2015.

[8] Ke Jiang, Brian Kulis, and Michael I Jordan. Small-variance asymptotics for exponential family dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pages 3167–3175, 2012.

[9] B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, Edinburgh, UK, 2012.

[10] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[11] Juho Lee and Seungjin Choi. Bayesian hierarchical clustering with exponential family: Small-variance asymptotics and reducibility. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 581–589, 2015.

[12] J.S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

[13] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[14] T.V. Nguyen, D. Phung, X.L. Venkatesh, S. Nguyen, and H. Bui. Bayesian nonparametric multilevel clustering with group-level contexts. In *Proc. of International Conference on Machine Learning (ICML)*, pages 288–296, Beijing, China, 2014.

[15] Anirban Roychowdhury, Ke Jiang, and Brian Kulis. Small-variance asymptotics for hidden markov models. In *Advances in Neural Information Processing Systems*, pages 2103–2111, 2013.

[16] Jonathon Shlens. Notes on kullback-leibler divergence and likelihood. *arXiv preprint arXiv:1404.2000*, 2014.

[17] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[18] Yining Wang and Jun Zhu. Small-variance asymptotics for dirichlet process mixtures of svms. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[19] Yining Wang and Jun Zhu. Dp-space: Bayesian nonparametric subspace clustering with small-variance asymptotics. In *International Conference on Machine Learning (ICML)*, 2015.