Computer Modeling in Engineering & Sciences CMES, vol.121, no.1, pp.49-82, 2019

The Exact Inference of Beta Process and Beta Bernoulli Process From Finite Observations

Yang Cheng[1], Dehua Li[1,*] and Wenbin Jiang[2]

Abstract: Beta Process is a typical nonparametric Bayesian model. and the Beta Bernoulli Process provides a Bayesian nonparametric prior for models involving collections of binary valued features. Some previous studies considered the Beta Process inference problem by giving the Stick-Breaking sampling method. This paper focuses on analyzing the form of precise probability distribution based on a Stick-Breaking approach, that is, the joint probability distribution is derived from any finite number of observable samples: It not only determines the probability distribution function of the Beta Process with finite observation (represented as a group of number between $[0,1]$), but also gives the distribution function of the Beta Bernoulli Process with the same finite dimension (represented as a matrix with element value of 0 or, 1) by using this distribution as a prior.

Keywords: Beta process, joint distribution, beta Bernoulli process, exact inference.

1 Introduction

Non-parametric Bayesian model is a kind of probability model, and the number of parameters of its probability distribution can increase with the increase of the number of samples [Alqifari and Coolen (2019)]. It is one of the most important and complex types of Probability Graph models. Therefore, the inference of Non-parametric Bayesian model has always been an important research direction of probability model [Griffin, Kalli and Steel (2018)], such as variational inference [Yao, Vehtari, Simpson et al. (2018)] and regression analysis [Seo, Wallat, Graepel et al. (2000)].

The Beta Process is a Non-parametric Bayesian model. It is mostly used for Bayesian Nonparametric prior of binary sparse characteristic matrix [Andrea, Stefano and Pietro (2018)]. It is widely used in various fields, such as Dictionary learning [Liu, Yu and Sun (2016)], Factor analysis [Andrew, Pu and Sun (2017)], Boltzmann machine learning [Lee and Hong (2016)] and so on. As a Non-parametric Bayesian model, it is almost the preferred prior distribution [Romain, Thibaux and Michael (2007)] for a sequence of any length whose element values are within the interval of $(0,1)$. When the Beta Process is

[1] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, 430074, China.

[2] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, China.

[*] Corresponding Author: Dehua Li. Email: lidehua1946@sina.com.

taken as the prior distribution of the Bernoulli Process, the Beta Process is marginalized, and the Beta Bernoulli Process will be obtained. When the second parameter of the marginalized Beta Process is set to 1 , it will become an Indian Buffet Process (IBP) [Griffiths and Ghahramani (2011)].

At present, Paislry has derived a method of Stick Breaking Construction for complete Beta Process, which has been widely used in Beta Process Factor Analysis [John and Lawrence (2009)]. It is useful to study the inference method of Beta Process, which depends on the infinite Bernoulli Process tends to the Poisson Process. Similar methods are used to deduce the infinite sequence of IBP. Teh et al. made new progress in this regard (2007) [Teh, Görür and Ghahramani (2007)], and they derived a Stick Breaking Construction method for the special case of "marginalizing the Beta Process in the Beta Bernoulli Process to generate a single parameter IBP". The Stick Breaking construction method is an important distribution fitting tool of Non-parametric Bayesian model [Eric and Padhraic (2017)], which is widely used in Dirichlet Process [Antoniak (1974)] and Gamma Process [Acharya, Teffer, Henderson et al. (2015)], etc.

Since it is a supervised learning, the task of this machine learning is based on the observable sample, to reversely deduce the model likelihood function contained in the sample after the sampling method is given [Finale and Shakir (2009)]. However, different tasks and different conditions will lead to different observable variables in specific tasks. If the observation variable is not the initial variable but the intermediate variable after the operation of the initial variable, the form of the likelihood function itself will change. Therefore, in practice, it is our core processing task to analyze the likelihood function of variables that are more likely to be the final observation variables in some tasks.

The variational inference [John and Lawrence (2011)] of the Beta Process and the likelihood function [Teh, Görür and Ghahramani (2007)] of the Beta Bernoulli Process in the past were mainly used to build the probability distribution function for the intermediate variables needed to generate the sample algorithm. This way of the construction of a probability distribution function is based on the following three basic hypothesis as constraints.

Above all, because in the Stick Breaking Construction method of the Beta Process, the final observed variables sampled from the Beta Process are generated by function mapping and arithmetic operations on two random variables that obey the other two distributions. Therefore, most of the inference work in the past was to directly establish the joint distribution function on the other two intermediate variables, and the result was that the joint probability distribution function itself did not contain the beta random variables [Teh, Görür and Ghahramani (2007); John and Lawrence (2011)].

Afterwards, when the Beta Process samples are generated by using the Stick Breaking Construction method, it is necessary to model the number of rounds in each sample, and the indicator function is adopted, but it is tedious to directly establish the probability distribution for the indicator function of all samples [John and Lawrence (2011)].

Finally, when samples are generated from the Beta Process through the Stick Breaking Construction method of Beta Process, and as a Bernoulli Process prior to the Bernoulli Process modeling, because it is a list of the product of the observed variables as the parameters of the Bernoulli distribution [John and Aimee (2010)], makes it hard for

subsequent integral treatment, need through the sampling method of approximate integral operation, Sampling is also a tedious step.

In this work, we intercept a finite number of random variable observations sampled from a Beta Process, calculate a posteriori Bernoulli Process, and make inferences. Here, we mainly do two things. First, for a high-dimensional sequence consisting of any finite number of real number observations with values between $[0,1]$, only the hypothesis generated by sampling from a Beta Process is made, and its probability distribution is directly analyzed and inferred. The second is that for a $0/1$ matrix that can have any finite row, and each row can have any finite column, you just make the assumption that you sampled from a Beta Bernoulli process and infer its distribution function. The definition of relevant parameters is similar to that given in [John and Lawrence (2011)]. The inference process here can be made without any additional assumptions. The above three restrictions can be relaxed in turn:

Above all, we set up the joint probability distribution function of the Beta Process by taking the Beta random variable itself as the observation variable, and the other random variables as the intermediate variables. In this way, the distribution function of the Beta Process can be directly generated by marginalization.

Furthermore, when we construct the likelihood function, instead of recursing the number of rounds of each sample sequentially, we only focus on the number of rounds of the last sample, so that we can directly construct the joint distribution function of the number of samples in each round at one time.

Finally, we analyze the Beta Process of a finite number of observation samples and use it as a prior distribution to directly calculate the posterior probability of the occurrence of any finite dimensional binary valued matrix, so that the posterior probability can be directly analyzed and calculated. Thus, the possibility of any finite dimensional binary matrix is analyzed directly.

Finite dimensional binary matrices can be used to select factors, such as modeling radar signal data. The radar transmits a set of full-bandwidth spectrum data $X = \{x_n\}_{n=1}^{N}$, the n_{th} sample is $x_n = [x_{n1}, \cdots, x_{nL}]$, L is the sample dimension. The factor analysis model can be used to model the full bandwidth spectrum data. Beta Bernoulli priori is used for the model, the n_{th} full-bandwidth spectrum data is expressed as $x_n = \Phi\hat{\omega}_n + \varepsilon_n$, and $\hat{\omega}_n = \omega_n \odot z_n$. Where, Φ represents the Shared factor loading matrix of this set of full-bandwidth spectrum data. K is the number of factors. The sparse weight $\hat{\omega}_n = \left[\hat{\omega}_{n1}, \cdots, \hat{\omega}_{nK}\right]^T \in R^K$ is composed of weight $\omega_n = [\omega_{n1}, \cdots, \omega_{nK}]^T$ and binary allocation variable $z_n = [z_{n1}, \cdots, z_{nK}]^T$. $\varepsilon_n = [\varepsilon_{n1}, \cdots, \varepsilon_{nL}]^T \in R^L$ is the noise variable. Where, it can be seen that ω_n is used to represent the weight of $\{\Phi_k\}_{k=1}^{K}$, while the binary variable $z_{nj} \in \{0,1\}$ is used to achieve sparse, non-zero only in the position of some

column vectors of Φ . Here, for the binary variable z_n , we use the Beta Bernoulli Process priors. For the Beta Bernoulli Process priors, we usually use the finite approximation of the Beta Bernoulli Process. Here, however, you can directly use the priors of the Beta Bernoulli Process without approximations.

The rest of this article is organized as follows. The second part through the analysis, provides a preliminary knowledge of the Beta Process and the probability distribution function of the observed variables generated by the Stick Breaking Construction method. In the third part, the inference method of probability distribution function of intermediate variable is given. In the last part, the final likelihood function of the observed variable is given by deduction. Describing likelihood function as efficiently as possible is an important step in machine learning with probability model.

2 The definition of the beta process and stick breaking construction

The Beta Process is a nonparametric Bayesian method, which is used to describe a sequence composed of an infinite number of atoms, in which each atom has a weight, and the weight is subject to a degenerate Beta distribution.

2.1 Beta process definition

Let H_0 be a non-atomic continuous measure on the space (Ω, B) , and $H_0(\Omega) = \gamma$. γ finite. Let a, b be two positive scalars. Define a process H_K as

$$
\begin{aligned}
&H_K(\theta) = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}(\theta) \\
&\pi_k \left| a, b, \gamma \overset{iid}{\sim} Beta\left(\frac{a\gamma}{K}, \frac{b(K-\gamma)}{K} \right) \right. \\
&\theta_k \overset{iid}{\sim} \frac{1}{\gamma} H_0
\end{aligned}
\tag{1}
$$

When $K \to \infty$, $H_K \to H$, where H is a beta process, namely $H \sim BP(a, b, H_0)$.

2.2 Stick breaking construction of beta process

By defining a concept called stick breaking, Paisley et al. [John and Aimee (2010)] clearly proposed a method to build the Beta Process. Stick breaking is a method used to generate discrete probability measure [Ishwaran and James (2001)], which plays an important role in the inference of Non-parametric Bayesian model. For the Beta Process, Paisley et al. [John and Aimee (2010); John, David and Michael (2012)] proposed the following expression method:

$$
\begin{aligned}
H &= \sum_{j=1}^{C_1} V_{1j}\delta_{\theta_{1j}} + \sum_{i=2}^{\infty}\sum_{j=1}^{C_i} V_{ij}e^{-T_{ij}}\delta_{\theta_{ij}} \\
C_i &\overset{iid}{\sim} Poisson\left(\frac{a\gamma}{b}\right) \\
\theta_{ij} &\overset{iid}{\sim} \frac{1}{\gamma}H_0 \\
T_{ij} &\overset{iid}{\sim} Gamma\left(i-1,b\right) \\
V_{ij} &\overset{iid}{\sim} Beta\left(1,b\right)
\end{aligned}
\tag{2}
$$

2.3 Calculation of edge distribution of π_k

Through the construction of stick breaking concept, Paisley et al. [John and Aimee (2010); John and Michael (2016)] proposed a method that can clearly show the process of the construction of the Beta Process. They divided the probability distribution of elements in the Beta Process into two groups: that is π_j is generated in the $1th$ round, and π_j is not generated in the first round of cycles. Paisley et al. [John and Aimee (2010)] calculated the marginal probability distribution of these two types of observations respectively.

Here, when π_k is specified to be generated in the i_{th} round, the conditional probability density function of π_k can be defined as follows:

When $i=1$, the corresponding weight of the atoms in this round follows a Beta distribution, with the first parameter as 1 , and the second parameter as b . That is $V_k \sim Beta\left(1,b\right)$ and $\pi_k = V_k$, Its probability density function is given as,

$$p\left(\pi_k \middle| 1,b\right) = b\left(1-\pi_k\right)^{b-1} \tag{3}$$

For other case $i>1$, we have $\pi_k = V_k e^{-T_k}$, and $V_k \sim Beta\left(1,b\right)$, at the same time $T_k \sim Gamma\left(i-1,\mathrm{b}\right)$. Another probability density function can be obtained by calculating the probability distribution of the function of random variables and the probability distribution function of the product of random variables:

$$p\left(\pi_k \mid i,b\right) = \left(-1\right)^{i-2}\frac{b^i}{\Gamma\left(i-1\right)}\int_{\pi_k}^{1} w^{b-2}\left(\ln w\right)^{d_k-2}\left(1-\frac{\pi_k}{w}\right)^{b-1} dw \tag{4}$$

where, the intermediate variable $w = e^{-T_k}$ is defined in Eq. $\left(4\right)$, and at the same time, i indicates the number of rounds of variable π_k occurrence.

3 Calculation of the joint probability density function of the final observation variables $\vec{\pi}$

Previous studies have shown that the number of atoms generated in each round of the stick-breaking construction of Beta Process obeys a Poisson distribution. The stick-breaking construction represented as the superposition of a countable infinite set of independent Poisson processes is useful for further representing the Beta Process.

In order to facilitate the inference, Paisley et al. [John and Aimee (2010)] proposed to use an indicator variable d_k to mark the number of rounds in which the k_{th} atom appeared, so as to obtain the formula:

$$d_k = 1 + \sum_{i=1}^{\infty} \mathrm{I}\left(\sum_{j=1}^{i} C_i < k \right)$$

The equation $d_k = i$ indicates that the k_{th} atom occurred in the i_{th} round.

3.1 Inference for d_k

For given d_k , we can reconstruct $\{C_i\}_{i=1}$.

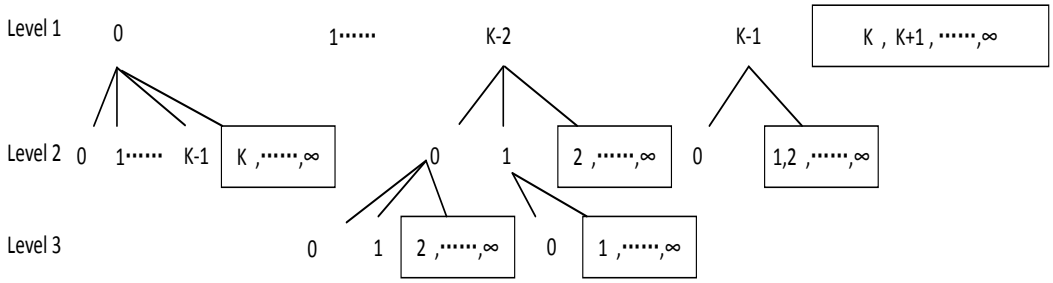

Figure 1: The relationship between the number of samples contained in each round and variable d_k

Given these latent indicator variables, the observation generation process can be rewritten as $\vec{\pi}\left|\{d_k\}_{k=1}^{\infty}\right.$. By changing the variable $\{d_k\}_{k=1}^{\infty}$ to $\left\{\{C_i\}_{i=1}^{d_k-1}, d_k\right\}$, the expression can be redescribed.

When given variable $d_k \geq 2$, the variable $\{C_i\}_{i=1}^{d_k-1}$ needs to be introduced to represent the size of each round. When the corresponding indicator variable of all samples generated in round j is $d_k = j$, the total quantity is C_j . Here means, for example, when all samples generated in round 2 are expressed as $d_k = 2$, then the total number of samples in round 2 is C_2 , and so on.

For given variables $d_k \geq 2$, we use variable $\{C_i\}_{i=1}^{d_k-1}$ to represent the size of each group. All samples for round j corresponding indicator variable $d_k = j$, the total number is C_j , this means: All samples for round 2 are represented as $d_k = 2$, total quantity of samples in round 2 is C_2 , and so on. Accordingly, j can be used to represent d_k for sequence analysis of all rounds.

Thus, as shown in Fig. 1, the probability of the k^{th} atom being observed in round 1 is:

$$p(d_k = 1) = 1 - \sum_{C_1=0}^{k-1} Poi(C_1)$$

This is what is shown in the first line frame section of Fig. 1. By analogy, the probability of k^{th} atom being observed in round 2 is:

$$p(d_k = 2) = \sum_{C_1=0}^{k-1} Poi(C_1)\left[1 - \sum_{C_2=0}^{k-C_1-1} Poi(C_2)\right]$$

According to the same reason continue to deduce, can get the final result about the probability of the k^{th} atom appearing in the round d_k . When $d_k \geq 3$, the final result is:

$$p(d_k) = \sum_{C_1=0}^{k-1} \cdots \sum_{C_v=0}^{k-\sum_{s=1}^{v-1} C_s - 1} \cdots \sum_{C_{d_k-1}=0}^{k-\sum_{m=1}^{d_k-2} C_m - 1} \prod_{w=1}^{d_k-1} Poi(C_w)\left[1 - \sum_{C_{d_k}=0}^{k-\sum_{l=1}^{d_k-1} C_l - 1} Poi(C_{d_k})\right]$$

On the other hand, the marginal probability can be viewed as obtained by marginalizing other variables of the joint probability distribution. In this way, the form of joint probability distribution can be obtained through the marginal probability distribution:

$$p(C_1, \cdots, C_{d_k-1}, d_k) = \prod_{w=1}^{d_k-1} Poi(C_w)\left[1 - \sum_{C_{d_k}=0}^{k-\sum_{l=1}^{d_k-1} C_l - 1} Poi(C_{d_k})\right] \tag{5}$$

Below, we discuss the likelihood terms and prior terms.

The inference processing process of the joint probability distribution of the number of samples generated in each round can be described as follows:

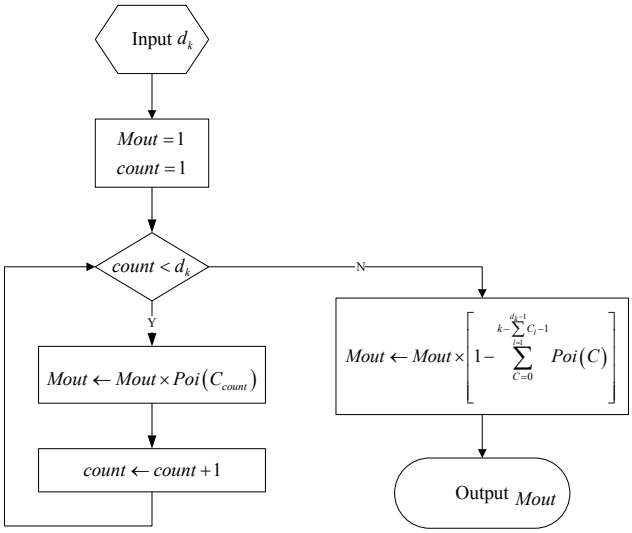

Figure 2: Calculation procedure of probability $p\left(C_1,\cdots,C_{d_k-1},d_k\right)$

Here, $count$ is the loop variable. When the loop is complete, the output variable value $Mout$ is the result $p\left(C_1,\cdots,C_{d_k-1},d_k\right)$.

Here, the loop only executes d_k times, and the time complexity is $\mathrm{O}\left(d_k\right)$. Next, we discuss the likelihood term and the prior term.

3.2 Likelihood term

To solve the problem of likelihood term, the integration problem of random variable d_k should be solved first. Here, the conditional probability formula is adopted, and the joint probability can be expanded by the value of variable d_k. And then we introduce the joint probability of $\vec{C}$. From the conditional probability formula, there is:

$$\begin{aligned} &p(\pi_1,\cdots,\pi_k)= \\ &p\left(\pi_1,\cdots,\pi_k\left|d_k=1\right.\right)\cdot p\left(d_k=1\right)+\sum_{C_1} p\left(\pi_1,\cdots,\pi_k\left|C_1,d_k=2\right.\right)\cdot p\left(C_1,d_k=2\right) \\ &+\sum_{d_k=3}^{\infty}\sum_{C_1\cdots C_{d_k-1}} p\left(\pi_1,\cdots,\pi_k\left|C_1,\cdots \mathrm{C}_{d_k-1},d_k\right.\right)\cdot p\left(C_1,\cdots \mathrm{C}_{d_k-1},d_k\right) \end{aligned} \tag{6}$$

Eq. (6) proposes a method to realize the joint probability distribution, which represents the observation sequence generated by the Beta Process. We use limited observation (here k) to generate the observation likelihood.

$p\left(\vec{C}, d_k\right)$ here is the joint probability distribution generated by Eq. (5). If we expand and analyze this formula, we can get the integral of d_k . Substitute Eq. (5) into Eq. (6), will get:

$$
\begin{aligned}
& p(\pi_1, \cdots, \pi_k) \\
& = \left[\prod_{j=1}^{k} b\left(1-\pi_j\right)^b\right] \cdot \left[1-\sum_{C_1=0}^{k-1} Poi\left(C_1\right)\right] \\
& + \sum_{C_1=0}^{k-1} Poi\left(C_1\right)\left[1-\sum_{C_2=0}^{k-C_1-1} Poi\left(C_2\right)\right] \cdot p\left(\pi_1, \cdots, \pi_k \middle| C_1, d_k=2\right) \\
& + \sum_{d_k=3}^{\infty} \sum_{C_1=0}^{k-1} \cdots \sum_{C_v=0}^{k-\sum_{s=1}^{v-1} C_s-1} \cdots \sum_{C_{d_k-1}=0}^{k-\sum_{m=1}^{d_k-2} C_m-1} \prod_{w=1}^{d_k-1} Poi\left(C_w\right)\left[1-\sum_{C_{d_k}=0}^{k-\sum_{l=1}^{d_k-1} C_l-1} Poi\left(C_{d_k}\right)\right] \\
& \cdot p\left(\pi_1, \cdots, \pi_k \middle| C_1, \cdots \mathrm{C}_{d_k-1}, d_k\right)
\end{aligned}
\tag{7}
$$

Here item $p\left(\pi_1, \cdots, \pi_k \middle| d_k=1\right)$ in Eq. (6) has been replaced by Eq. (3).

3.3 Derivation of conditional probability term

The data is generated by H through the Beta Process and expressed in the form of an infinite dimensional vector, with each element between $(0,1)$. The probability distribution $p\left(\vec{\pi} \middle| \vec{C}, d_k\right)$ is analyzed as follows: this formula is equivalent to the posterior distribution of $\pi_1, \pi_2, \cdots, \pi_k$ after the indicator sequence is given.

3.3.1 Inference for $d_k = 2$

Given d_k value, can reconstruct the $\{\pi_k\}_{k=1}$. In other words, k samples were generated in the first round and the second round, and the sum of the number of samples in the two rounds was k .

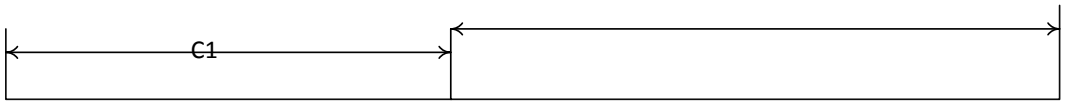

Figure 3: the relationship between the value of C_1 and variables $d_k = 2$

Given C_1 and k , The joint distribution of variables $\vec{\pi}$ is:

$$\left[\frac{\prod_{q=1}^{k} p\left(\pi_q \middle| 1,b\right)}{\prod_{r=C_1+1}^{k} p\left(\pi_r \middle| 1,b\right)}\right]\cdot\left[\prod_{f=C_1+1}^{k} p\left(\pi_f \middle| 2,b\right)\right] \tag{8}$$

Here, the $p\left(\pi_i \middle| d_j,b\right)$ item in Eq. (8) will be replaced by Eqs. (3) and (4).

3.3.2 Inference at $d_k \geq 3$

Given d_k , we can reconstruct $\left\{\pi_k\right\}_{k=1}$. This means that k samples are generated in the initial d_k rounds, and the sum of the number of samples in d_k rounds is k , which can be understood as that all samples generated in the first d_k-1 rounds have been completely observed, while only partial samples have been observed in the last round.

For each round of samples, the corresponding joint probability distribution needs to be defined. So it need to separate the samples from each round.

From the definition of round, we can obtain the elements from the j_{th} round:

$$C_j=\left\{C_j,\cdots \mathrm{C}_{d_k-1},\cdots d_k\right\}\setminus\left\{C_{j+1},\cdots \mathrm{C}_{d_k-1},\cdots d_k\right\}$$

Then, given the values of $\vec{C}$ and k , the joint distribution of variables $\vec{\pi}$ can be obtained:

$$\left[\frac{\prod_{q=1}^{k} p\left(\pi_q \middle| 1,b\right)}{\prod_{r=C_1+1}^{k} p\left(\pi_r \middle| 1,b\right)}\right]\times\left\{\prod_{t=2}^{d_k-1}\left[\frac{\prod_{l=\sum_{j=1}^{t-1} C_j+1}^{k} p\left(\pi_l \middle| \mathrm{t},b\right)}{\prod_{s=\sum_{i=1}^{t} C_i+1}^{k} p\left(\pi_s \middle| \mathrm{t},b\right)}\right]\right\}\times\left[\prod_{f=\sum_{m=1}^{d_k-1} C_m+1}^{k} p\left(\pi_f \middle| d_k,b\right)\right] \tag{9}$$

The $p\left(\pi_i \middle| d_j,b\right)$ term in Eq. (9) will also be replaced by Eqs. (3) and (4).

The calculation process of the conditional probability $p\left(\vec{\pi}\middle|\vec{C},d_k\right)$ can be described as follows:

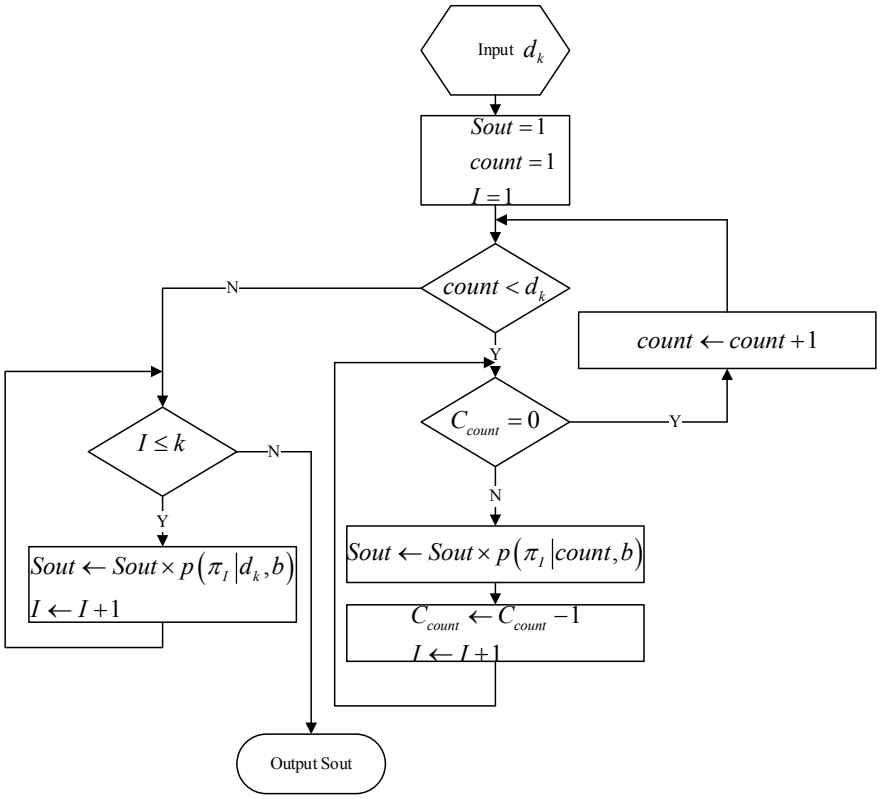

Figure 4: Calculation process of conditional probability $p\left(\pi_1,\cdots,\pi_k \middle| C_1,\cdots,C_{d_k-1},d_k\right)$

Here, $count, I$ is the loop variable. When the loop is complete, the output variable value $Sout$ is the result $p\left(\pi_1,\cdots,\pi_k \middle| C_1,\cdots,C_{d_k-1},d_k\right)$.

For $s = 1,2,\cdots,k$ samples $\left\{\pi_s\right\}_{s=1}^{k}$ can be drawn *i.i.d.* from the Beta Process $BP(a,b,H)$.

In this way, the joint probability distribution of the final observed variables can be calculated,

$$
\begin{aligned}
&p(\pi_1,\cdots,\pi_k\,|a,b)= \\
&=\left[\prod_{j=1}^{k} b\left(1-\pi_j\right)^b\right]\cdot\left[1-\sum_{C_1=0}^{k-1} Poi\left(C_1\right)\right] \\
&+\sum_{C_1=0}^{k-1} Poi\left(C_1\right)\left[1-\sum_{C_2=0}^{k-C_1-1} Poi\left(C_1\right)\right]\cdot\left[\frac{\prod_{q=1}^{k} p\left(\pi_q\,|1,b\right)}{\prod_{r=C_1+1}^{k} p\left(\pi_r\,|1,b\right)}\right]\cdot\left[\prod_{f=C_1+1}^{k} p\left(\pi_f\,|2,b\right)\right] \\
&+\sum_{d_k=3}^{\infty}\sum_{C_1=0}^{k-1}\cdots\sum_{C_v=0}^{k-\sum_{s=1}^{v-1}C_s-1}\cdots\sum_{C_{d_k-1}=0}^{k-\sum_{m=1}^{d_k-2}C_m-1}\prod_{w=1}^{d_k-1} Poi\left(C_w\right)\left[1-\sum_{C_{d_k}=0}^{k-\sum_{l=1}^{d_k-1}C_l-1} Poi\left(C_{d_k}\right)\right]\cdot\left[\frac{\prod_{q=1}^{k} p\left(\pi_q\,|1,b\right)}{\prod_{r=C_1+1}^{k} p\left(\pi_r\,|1,b\right)}\right] \\
&\times\left\{\prod_{t=2}^{d_k-1}\left[\frac{\prod_{l=\sum_{j=1}^{t-1}C_j+1}^{k} p\left(\pi_l\,|\mathrm{t},b\right)}{\prod_{s=\sum_{i=1}^{t}C_i+1}^{k} p\left(\pi_s\,|\mathrm{t},b\right)}\right]\right\}\times\left[\prod_{f=\sum_{m=1}^{d_k-1}C_m+1}^{k} p\left(\pi_f\,|d_k,b\right)\right]
\end{aligned}
\tag{10}
$$

Substituting Eqs. (3) and (4) into Eq. (10), the final precise joint probability distribution function for a finite number of observations of Beta Process is obtained.

Thus, the overall calculation process of the likelihood term can be described as follows:

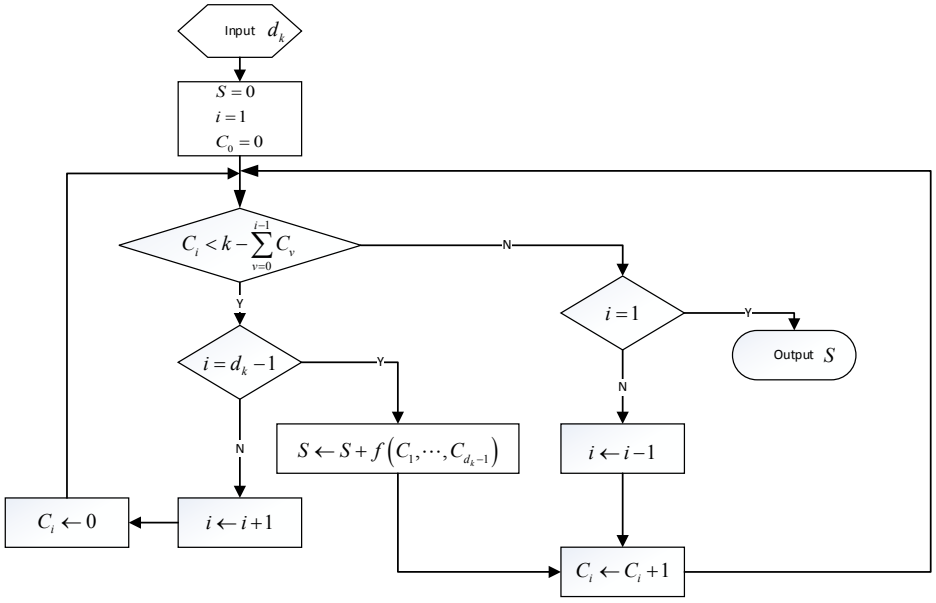

Figure 5: Calculation process of probability $p\left(\pi_1,\cdots,\pi_k,d_k\right)$

In the calculation process shown in Fig. 5, $f\left(C_1,\cdots,C_{d_k-1}\right)$ is used to represent the calculation result of multiplying *Mout* in Fig. 2 and *Sout* in Fig. 4, that is, the joint probability distribution of the observation variable $\vec{\pi}$, and the variable $\left(\vec{C},d_k\right)$ that needs marginalization. The distribution function is: $p\left(\pi_1,\cdots,\pi_k,C_1,\cdots,C_{d_k-1},d_k\right)$.

The computation time complexity is $\mathrm{O}\left(d_k\right)$.

4 Calculation of the final joint probability distribution function of the observed variables $\vec{z}$

Data $\vec{\pi}$ is drawn *i.i.d.* from a Beta Process and then via a Bernoulli Process can obtain finite dimensional binary vector form: $z_{ij}\in\left\{0,1\right\}^{k\times g_k}$, Where g_i represents row i of matrix Z containing g_i column, namely:

$$\begin{aligned}&\left(\pi_1,\cdots,\pi_k\right)\overset{iid}{\sim}BP\left(a,b,H\right)\\&z_{ij}\overset{iid}{\sim}Bernoulli\left(\pi_i\right)\end{aligned}\tag{11}$$

This can be represented as a $k\times g_k$ dimensional binary matrix [Finale, Kurt, Jurgen et al. (2009)].

$$\begin{pmatrix}\pi_1\\\vdots\\\pi_k\end{pmatrix}\begin{pmatrix}z_{11}&\cdots&z_{1g_1}\\\vdots&\ddots&\vdots\\z_{k1}&\cdots&z_{kg_k}\end{pmatrix}$$

Figure 6: the relationship between the variables $\vec{\pi}$ and variables $\left\{z_{ij}\right\}$

The sufficient statistics calculated from $\left\{z_{jh}\right\}_{h=1}^{g_j}$ are the counts along each dimension j, so that we have:

$$M_j=\sum_{h=1}^{g_j}z_{jh}\tag{12}$$

where z_{ij} has been specified to be drawn from a Bernoulli Process with a parameter π_i. Then the Joint probability distribution of $\left\{\vec{z}\right\}$ and $\left\{\vec{\pi}\right\}$ is subject to:

$$p\left(z_{11},\cdots,z_{kg_k},\pi_1,\cdots,\pi_k\right)=\prod_{j=1}^{k}\prod_{h=1}^{g_j}\pi_j^{z_{jh}}\left(1-\pi_j\right)^{1-z_{jh}}\ p\left(\pi_1,\cdots,\pi_k\right) \tag{13}$$

Here each π_j parameter follows a beta distribution. The joint probability distribution of $\vec{\pi}$ can be calculated by Eq. (10).

By substituting Eqs. (10), (12) into Eq. (13), the exact Joint Probability distribution function of Beta Bernoulli Process for finite observation can be obtained.

$$p\left(z_{11},\cdots,z_{kg_k}\left|a,b\right.\right)=\int_0^1\cdots\int_0^1\prod_{j=1}^{k}\pi_j^{M_j}\left(1-\pi_j\right)^{g_j-M_j}p\left(\pi_1,\cdots,\pi_k\right)d\pi_1\cdots d\pi_k \tag{14}$$

Using the above equation, the variable $\vec{\pi}$ in the intermediate step can be conveniently eliminated by integration, and the result is given in Eq. (15).

$$\begin{aligned}
&p\left(z_{11},\cdots,z_{kg_k}\left|a,b\right.\right)=\\
&=\sum_{R=0}^{\infty}Poi\left(k+R\right)\prod_{j=1}^{k}\int_0^1 b\pi_j^{M_j}\left(1-\pi_j\right)^{b+g_j-M_j-1}d\pi_j+\\
&\sum_{C_1=0}^{k-1}\sum_{R=0}^{\infty}Poi\left(C_1\right)\cdot Poi\left(k-C_1+R\right)\cdot\left[\frac{\prod_{q=1}^{k}\int_0^1\pi_q^{M_q}\left(1-\pi_q\right)^{g_q-M_q}\ p\left(\pi_q\left|1,b\right.\right)d\pi_q}{\prod_{r=C_1+1}^{k}\int_0^1\pi_r^{M_r}\left(1-\pi_r\right)^{g_r-M_r}\ p\left(\pi_r\left|1,b\right.\right)d\pi_r}\right]\times\\
&\left[\prod_{f=C_1+1}^{k}\int_0^1\pi_f^{M_f}\left(1-\pi_f\right)^{g_f-M_f}\ p\left(\pi_f\left|2,b\right.\right)d\pi_f\right]+\\
&\sum_{d_k=3}^{\infty}\sum_{C_1=0}^{k-1}\cdots\sum_{C_v=0}^{k-\sum_{s=1}^{v-1}C_s-1}\cdots\sum_{C_{d_k-1}=0}^{k-\sum_{m=1}^{d_k-2}C_m-1}\sum_{R=0}^{\infty}\prod_{w=1}^{d_k-1}Poi\left(C_w\right)\cdot Poi\left(k-\sum_{l=1}^{d_k-1}C_l+R\right)\times\\
&\left[\frac{\prod_{q=1}^{k}\int_0^1\pi_q^{M_q}\left(1-\pi_q\right)^{g_q-M_q}\ p\left(\pi_q\left|1,b\right.\right)d\pi_q}{\prod_{r=C_1+1}^{k}\int_0^1\pi_r^{M_r}\left(1-\pi_r\right)^{g_r-M_r}\ p\left(\pi_r\left|1,b\right.\right)d\pi_r}\right]\times\\
&\left\{\prod_{t=2}^{d_k-1}\left[\frac{\prod_{l=\sum_{j=1}^{t-1}C_j+1}^{k}\int_0^1\pi_l^{M_l}\left(1-\pi_l\right)^{g_l-M_l}\ p\left(\pi_l\left|\mathrm{t},b\right.\right)d\pi_l}{\prod_{s=\sum_{j=1}^{t}C_j+1}^{k}\int_0^1\pi_s^{M_s}\left(1-\pi_s\right)^{g_s-M_s}\ p\left(\pi_s\left|\mathrm{t},b\right.\right)d\pi_s}\right]\right\}\times\left[\prod_{f=\sum_{m=1}^{d_k-1}C_m+1}^{k}\int_0^1\pi_f^{M_f}\left(1-\pi_f\right)^{g_f-M_f}\ p\left(\pi_f\left|\mathrm{d}_k,b\right.\right)d\pi_f\right]
\end{aligned} \tag{15}$$

Next, by replacing the distribution of the integral in Eq. (15) by using Eq. (3), we can obtain Eq. (16):

$$
\begin{aligned}
&p\left(z_{11},\cdots,z_{kg_k}\left|a,b\right.\right)= \\
&b^k\cdot\left(\sum_{R=0}^{\infty}Poi\left(k+R\right)\right)\cdot\prod_{j=1}^{k}\frac{\Gamma\left(M_j+1\right)\cdot\Gamma\left(b+g_j-M_j\right)}{\Gamma\left(b+g_j+1\right)} \\
&+\sum_{C_1=0}^{k-1}\sum_{R=0}^{\infty}Poi\left(C_1\right)\cdot Poi\left(k+R-C_1\right)\cdot\left[\frac{\prod_{q=1}^{k}\frac{\Gamma\left(M_q+1\right)\cdot\Gamma\left(b+g_q-M_q\right)}{\Gamma\left(b+g_q+1\right)}}{\prod_{r=C_1+1}^{k}\frac{\Gamma\left(M_r+1\right)\cdot\Gamma\left(b+g_r-M_r\right)}{\Gamma\left(b+g_r+1\right)}}\right]\cdot b^{C_1} \\
&\times\left[\prod_{f=C_1+1}^{k}\int_0^1\pi_f^{M_f}\left(1-\pi_f\right)^{g_f-M_f}\,p\left(\pi_f\left|2,b\right.\right)d\pi_f\right] \\
&+\sum_{d_k=3}^{\infty}\sum_{C_1=0}^{k-1}\cdots\sum_{C_v=0}^{k-\sum_{s=1}^{v-1}C_s-1}\cdots\sum_{C_{d_k-1}=0}^{k-\sum_{m=1}^{d_k-2}C_m-1}\sum_{R=0}^{\infty}\prod_{w=1}^{d_k-1}Poi\left(C_w\right)\cdot Poi\left(k-\sum_{l=1}^{d_k-1}C_l+R\right) \\
&\times\left[\frac{\prod_{q=1}^{k}\frac{\Gamma\left(M_q+1\right)\cdot\Gamma\left(b+g_q-M_q\right)}{\Gamma\left(b+g_q+1\right)}}{\prod_{r=C_1+1}^{k}\frac{\Gamma\left(M_r+1\right)\cdot\Gamma\left(b+g_r-M_r\right)}{\Gamma\left(b+g_r+1\right)}}\right]\cdot b^{C_1}\times\left\{\prod_{t=2}^{d_k-1}\left[\frac{\prod_{l=\sum_{j=1}^{t-1}C_j+1}^{k}\int_0^1\pi_l^{M_l}\left(1-\pi_l\right)^{g_l-M_l}\,p\left(\pi_l\left|\mathrm{t},b\right.\right)d\pi_l}{\prod_{s=\sum_{i=1}^{t}C_i+1}^{k}\int_0^1\pi_s^{M_s}\left(1-\pi_s\right)^{g_s-M_s}\,p\left(\pi_s\left|\mathrm{t},b\right.\right)d\pi_s}\right]\right\} \\
&\times\left[\prod_{f=\sum_{m=1}^{d_k-1}C_m+1}^{k}\int_0^1\pi_f^{M_f}\left(1-\pi_f\right)^{g_f-M_f}\,p\left(\pi_f\left|d_k,b\right.\right)d\pi_f\right]
\end{aligned}
\tag{16}
$$

Here, variable R is introduced through the properties of the Poisson distribution: $1-\sum_{C_i=0}^{k-1}Poi\left(C_i\right)=\sum_{R=0}^{\infty}Poi\left(k+R\right)$.

In order to simplify the following calculation, the calculation form can be appropriately simplified first:

$$\left\{\prod_{t=2}^{d_k-1}\left[\frac{\prod_{l=\sum_{j=1}^{t-1}C_j+1}^{k}\int_0^1\pi_l^{M_l}\left(1-\pi_l\right)^{g_l-M_l}\,p\left(\pi_l\left|\mathrm{t},b\right.\right)d\pi_l}{\prod_{s=\sum_{i=1}^{t}C_i+1}^{k}\int_0^1\pi_s^{M_s}\left(1-\pi_s\right)^{g_s-M_s}\,p\left(\pi_s\left|\mathrm{t},b\right.\right)d\pi_s}\right]\right\}\times\left[\prod_{f=\sum_{m=1}^{d_k-1}C_m+1}^{k}\int_0^1\pi_f^{M_f}\left(1-\pi_f\right)^{g_f-M_f}\,p\left(\pi_f\left|d_k,b\right.\right)d\pi_f\right]$$

$$=\prod_{t=2}^{d_k-1}\left[\prod_{l=\sum_{j=1}^{t}C_j+1}^{k}\left[\frac{\int_0^1\pi_l^{M_l}\left(1-\pi_l\right)^{g_l-M_l}\,p\left(\pi_l\left|\mathrm{t}+1,b\right.\right)d\pi_l}{\int_0^1\pi_l^{M_l}\left(1-\pi_l\right)^{g_l-M_l}\,p\left(\pi_l\left|\mathrm{t},b\right.\right)d\pi_l}\right]\right]\cdot\left[\prod_{s=C_1+1}^{k}\int_0^1\pi_s^{M_s}\left(1-\pi_s\right)^{g_s-M_s}\,p\left(\pi_s\left|2,b\right.\right)d\pi_s\right] \tag{17}$$

Eq. (17) is a simple shift of the last two terms in Eq. (15).

4.1 Likelihood term for $\left\{z_k\right\}$

Using the conjugate relationship between the Beta Process and the Bernoulli Process, through integral calculation, the following result can be obtained, without loss of generality. The Bernoulli samples produced in round d_k are analyzed here:

$$p\left(z_k\right)=\left(-1\right)^{d_k-2}\int_0^1\pi_k^{M_k}\left(1-\pi_k\right)^{g_k-M_k}\frac{b^{d_k}}{\Gamma\left(d_k-1\right)}\int_{\pi_k}^1 w^{b-2}\left(\ln w\right)^{d_k-2}\left(1-\frac{\pi_k}{w}\right)^{b-1}dwd\pi_k$$

$$=\left(-1\right)^{d_k-2}\frac{b^{d_k}}{\Gamma\left(d_k-1\right)}\int_0^1 dw\left[\int_0^w\pi_k^{M_k}\left(1-\pi_k\right)^{g_k-M_k}\left(w-\pi_k\right)^{b-1}d\pi_k\right]w^{-1}\left(\ln w\right)^{d_k-2} \tag{18}$$

In this case, z_k will be used to represent $\left\{z_{k1},\cdots,z_{kg_k}\right\}$. and the probability distribution of π_k has been represented by the Eq. (4).

By Taylor's expansion, we can obtain the integral of the variable π_k analytically.

Considering series analysis of the middle part of Eq. (18), i.e., the Taylor expansion of the term $\left(1-\pi_k\right)^{g_k-M_k}$, one can find that

$$\int_0^w\pi_k^{M_k}\left(1-\pi_k\right)^{g_k-M_k}\left(w-\pi_k\right)^{b-1}d\pi_k$$

$$=\sum_{s=0}^{g_k-M_k}\left(-1\right)^{g_k-M_k-s}\frac{\Gamma\left(g_k-M_k+1\right)}{\Gamma\left(s+1\right)\Gamma\left(g_k-M_k-s+1\right)}w^{g_k+b-s-1}\int_0^w\left(\frac{\pi_k}{w}\right)^{g_k-s}\left(1-\frac{\pi_k}{w}\right)^{b-1}d\pi_k \tag{19}$$

Through variable substitution, set $R=\frac{\pi_k}{w}$, then $R\in(0,1)$. At the same time have $dR=\frac{d\pi_k}{w}$. There are:

$$\int_0^w\left(\frac{\pi_k}{w}\right)^{g_k-s}\left(1-\frac{\pi_k}{w}\right)^{b-1}d\pi_k=w\int_0^1(R)^{g_k-s}(1-R)^{b-1}dR \tag{20}$$

The integral result in Eq. (19) can be obtained:

$$\begin{aligned}&\int_0^w\pi_k^{M_k}(1-\pi_k)^{g_k-M_k}(w-\pi_k)^{b-1}d\pi_k\\&=\sum_{s=0}^{g_k-M_k}(-1)^{g_k-M_k-s}\frac{\Gamma(g_k-M_k+1)}{\Gamma(s+1)\Gamma(g_k-M_k-s+1)}\cdot\frac{\Gamma(g_k-s+1)\Gamma(b)}{\Gamma(g_k-s+b+1)}w^{g_k+b-s}\end{aligned} \tag{21}$$

Substituting the calculation results of Eq. (21) into Eq. (18), we can obtain:

$$\begin{aligned}&p(z_k)=(-1)^{g_k-M_k+d_k-2}\frac{b^{d_k}\Gamma(g_k-M_k+1)\Gamma(b)}{\Gamma(d_k-1)}\\&\sum_{s=0}^{g_k-M_k}(-1)^s\frac{\Gamma(g_k-s+1)}{\Gamma(s+1)\Gamma(g_k-M_k-s+1)\Gamma(g_k-s+b+1)}\int_0^1w^{g_k+b-s-1}(\ln w)^{d_k-2}dw\end{aligned} \tag{22}$$

This is the likelihood term form of $\{z_k\}$ obtained by inference.

4.1.1 Likelihood term for $d_k=2$

From the above analysis, we now know that since $0\le s\le g_k-M_k$, $s\ne g_k+b$, so that have $g_k+b-s\ne 0$. Substituting $d_k=2$ into Eq. (22), and the following results can be obtained:

$$\begin{aligned}&p(z_k)=(-1)^{g_k-M_k}b^2\Gamma(g_k-M_k+1)\Gamma(b)\\&\sum_{s=0}^{g_k-M_k}(-1)^s\cdot\frac{1}{(g_k-s+b)}\cdot\frac{\Gamma(g_k-s+1)}{\Gamma(s+1)\Gamma(g_k-M_k-s+1)\Gamma(g_k-s+b+1)}\end{aligned} \tag{23}$$

4.1.2 Likelihood term for $d_k\ge 3$

We can use Eq. (22) to calculate the posterior distribution of z by integrating out the random variable w in Eq. (4) again. In this way, the last integral term of distribution function of z in Eq. (22) is:

$$\int_0^1 w^{g_k+b-s-1}\left(\ln w\right)^{d_k-2}\,dw \tag{24}$$

In order to calculate the posterior distribution of a given binary indicator variables, a prior distribution is required. In order to obtain the prior distribution, the following two steps can be performed:

First, variable substitution can be used to calculate the integral. Let $u=\ln w$, and then $w=e^u$. Therefore, Eq. (24) can be replaced by:

$$\int_0^1 w^{g_k+b-s-1}\left(\ln w\right)^{d_k-2}\,dw=\int_{-\infty}^0 e^{u\left(g_k+b-s\right)}u^{d_k-2}du \tag{25}$$

Next we let $t=-u\left(g_k+b-s\right)$, and then $u=-\dfrac{t}{g_k+b-s}$. Substituting these variables into Eq. (25), we can obtain:

$$\begin{aligned}\int_{-\infty}^0 e^{u\left(g_k+b-s\right)}u^{d_k-2}du&=\int_{\infty}^0 e^{-t}\left(-\frac{t}{g_k+b-s}\right)^{d_k-2}\left(-\frac{1}{g_k+b-s}\right)dt\\&=\left(-\frac{1}{g_k+b-s}\right)^{d_k-2}\left(\frac{1}{g_k+b-s}\right)\Gamma\left(d_k-1\right)\end{aligned} \tag{26}$$

4.2 Calculate the proportional term in equation

To calculate the proportional term in Eq. (17), we need to substitute the result of Eq. (23) into Eq. (17), in this way, we can obtain,

$$\begin{aligned}&\frac{\int_0^1 \pi_l^{M_l}\left(1-\pi_l\right)^{g_l-M_l}\,p\left(\pi_l\left|\mathrm{t}+1,b\right.\right)d\pi_l}{\int_0^1 \pi_l^{M_l}\left(1-\pi_l\right)^{g_l-M_l}\,p\left(\pi_l\left|\mathrm{t},b\right.\right)d\pi_l}\\&=-\frac{b}{t-1}\sum_{c=0}^{g_l-M_l}\left(-1\right)^c\frac{\Gamma\left(g_l-c+1\right)}{\Gamma\left(c+1\right)\Gamma\left(g_l-c+b+1\right)\Gamma\left(g_l-M_l-c+1\right)}\\&\times\frac{1}{\sum_{s=0}^{g_l-M_l}\left(-1\right)^s\frac{\Gamma\left(g_l-s+1\right)}{\Gamma\left(s+1\right)\Gamma\left(g_l-s+b+1\right)\Gamma\left(g_l-M_l-s+1\right)}\frac{\int_0^1 w^{g_l+b-s-1}\left(\ln w\right)^{t-2}\,dw}{\int_0^1 w^{g_l+b-c-1}\left(\ln w\right)^{t-1}\,dw}}\end{aligned} \tag{27}$$

Regarding the integral proportion term in the denominator of Eq. (27), by using the result of Eq. (26), it can be calculated as follows:

$$\frac{\int_0^1 w^{g_l+b-s-1}\left(\ln w\right)^{t-2}dw}{\int_0^1 w^{g_l+b-c-1}\left(\ln w\right)^{t-1}dw}=\frac{\left(-\frac{1}{g_l+b-s}\right)^{t-2}\left(\frac{1}{g_l+b-s}\right)\Gamma\left(t-1\right)}{\left(-\frac{1}{g_l+b-c}\right)^{t-1}\left(\frac{1}{g_l+b-c}\right)\Gamma\left(t\right)}=-\frac{1}{\left(t-1\right)}\frac{\left(g_l+b-c\right)^t}{\left(g_l+b-s\right)^{t-1}} \tag{28}$$

Substituting the result of Eq. (28) into Eq. (27), we can produce result:

$$\begin{aligned}&\frac{\int_0^1 \pi_l^{M_l}\left(1-\pi_l\right)^{g_l-M_l}\,p\left(\pi_l\left|\mathrm{t}+1,b\right.\right)d\pi_l}{\int_0^1 \pi_l^{M_l}\left(1-\pi_l\right)^{g_l-M_l}\,p\left(\pi_l\left|\mathrm{t},b\right.\right)d\pi_l}\\&=b\frac{\sum_{c=0}^{g_l-M_l}\left(-1\right)^c\frac{\Gamma\left(g_l-c+1\right)}{\Gamma\left(c+1\right)\Gamma\left(g_l-c+b+1\right)\Gamma\left(g_l-M_l-c+1\right)\left(g_l+b-c\right)^t}}{\sum_{s=0}^{g_l-M_l}\left(-1\right)^s\frac{\Gamma\left(g_l-s+1\right)}{\Gamma\left(s+1\right)\Gamma\left(g_l-s+b+1\right)\Gamma\left(g_l-M_l-s+1\right)\left(g_l+b-s\right)^{t-1}}}\end{aligned} \tag{29}$$

The Eq. (29) is the final result that we want to get here.

4.3 The final result of the joint probability distribution of $\vec{z}$

Substitute Eq. (29) into Eq. (17), then the power of constant b in Eq. (17) can be obtained by the following calculation result:

$$\prod_{t=2}^{d_k-1}\prod_{l=\sum_{j=1}^{t}C_j+1}^{k}b\cdot\prod_{l=C_1+1}^{k}b^2=b^{(d_k-1)k-2C_1-\sum_{t=2}^{d_k-1}\sum_{j=1}^{t}C_j} \tag{30}$$

At this time, substitute the results of Eqs. (17), (23), (29) and (30) into Eq. (16), the final calculation result of the variable $\vec{z}$ can be expressed as follows:

$$
\begin{aligned}
&p\left(z_{11},\cdots,z_{kg_k}\left|a,b\right.\right)= \\
&b^k\left(\sum_{R=0}^{\infty}Poi\left(k+R\right)\right)\cdot\prod_{j=1}^{k}\frac{\Gamma\left(M_j+1\right)\cdot\Gamma\left(b+g_j-M_j\right)}{\Gamma\left(b+g_j+1\right)} \\
&+\sum_{C_1=0}^{k-1}\sum_{R=0}^{\infty}Poi\left(C_1\right)\cdot Poi\left(k+R-C_1\right)\cdot\left[\frac{\prod_{q=1}^{k}\frac{\Gamma\left(M_q+1\right)\cdot\Gamma\left(b+g_q-M_q\right)}{\Gamma\left(b+g_q+1\right)}}{\prod_{r=C_1+1}^{k}\frac{\Gamma\left(M_r+1\right)\cdot\Gamma\left(b+g_r-M_r\right)}{\Gamma\left(b+g_r+1\right)}}\right] \\
&\cdot\Gamma\left(b\right)^{k-C_1}b^{2k-C_1}\cdot\prod_{f=C_1+1}^{k}\left\{\left(-1\right)^{g_f-M_f}\cdot\Gamma\left(g_f-M_f+1\right)\right. \\
&\left.\times\left[\sum_{s=0}^{g_f-M_f}\frac{\left(-1\right)^s\cdot\Gamma\left(g_f-s+1\right)}{\left(g_f-s+b\right)\cdot\Gamma\left(s+1\right)\cdot\Gamma\left(g_f-s+b+1\right)\cdot\Gamma\left(g_f-M_f-s+1\right)}\right]\right\} \\
&+\sum_{d_k=3}^{\infty}\sum_{C_1=0}^{k-1}\cdots\sum_{C_v=0}^{k-\sum_{s=1}^{v-1}C_s-1}\cdots\sum_{C_{d_k-1}=0}^{k-\sum_{m=1}^{d_k-2}C_m-1}\sum_{R=0}^{\infty}\prod_{w=1}^{d_k-1}Poi\left(C_w\right)\cdot Poi\left(k-\sum_{l=1}^{d_k-1}C_l+R\right) \\
&\times\left[\frac{\prod_{q=1}^{k}\frac{\Gamma\left(M_q+1\right)\cdot\Gamma\left(b+g_q-M_q\right)}{\Gamma\left(b+g_q+1\right)}}{\prod_{r=C_1+1}^{k}\frac{\Gamma\left(M_r+1\right)\cdot\Gamma\left(b+g_r-M_r\right)}{\Gamma\left(b+g_r+1\right)}}\right]\cdot b^{\left(d_k-1\right)k-\sum_{t=1}^{d_k-1}\sum_{j=1}^{t}C_j}\cdot\Gamma\left(b\right)^{k-C_1} \\
&\times\left\{\prod_{t=2}^{d_k-1}\left[\prod_{l=\sum_{j=1}^{t}C_j+1}^{k}\frac{\sum_{c=0}^{g_l-M_l}\left(-1\right)^c\frac{\Gamma\left(g_l-c+1\right)}{\left(g_l-c+b\right)^t\cdot\Gamma\left(c+1\right)\cdot\Gamma\left(g_l-c+b+1\right)\cdot\Gamma\left(g_l-M_l-c+1\right)}}{\sum_{s=0}^{g_l-M_l}\left(-1\right)^s\frac{\Gamma\left(g_l-s+1\right)}{\left(g_l-s+b\right)^{t-1}\cdot\Gamma\left(s+1\right)\cdot\Gamma\left(g_l-s+b+1\right)\cdot\Gamma\left(g_l-M_l-s+1\right)}}\right]\right\} \\
&\times\left\{\prod_{f=C_1+1}^{k}\left(-1\right)^{g_f-M_f}\cdot\Gamma\left(g_f-M_f+1\right)\right. \\
&\left.\cdot\left[\sum_{s=0}^{g_f-M_f}\frac{\left(-1\right)^s\cdot\Gamma\left(g_f-s+1\right)}{\left(g_f-s+b\right)\cdot\Gamma\left(s+1\right)\cdot\Gamma\left(g_f-s+b+1\right)\cdot\Gamma\left(g_f-M_f-s+1\right)}\right]\right\}
\end{aligned}
\tag{31}
$$

The result obtained from Eq. (31) is the final joint probability distribution function required in this paper.

Here, the calculation of the observed likelihood for each Bernoulli sample is done directly. Thus, the likelihood calculation of each Bernoulli sample is carried out directly. This process can be described as follows:

The calculation result can be analytically generated due to the integration order of beta variable π_k and intermediate variable w is exchanged in the function to be integrated, and then Taylor expansion is carried out on the power of $(1-\pi_k)$ in the exchange result.

The flow chart of the calculation of the final observation variable Z in the Beta Bernoulli process is almost the same as that in Fig. 4. Since the probability $p(\pi_I | count, b)$ in the process described in Figure 4 can be directly used here, due to that the introduction of conditional probability distribution of the observation variable $\vec{z}$, and distribution function $p(\overrightarrow{z_I} | count)$ can be analytically calculated by marginalizing variable π_I, i.e., $\int p(\pi_I | count, b) \cdot p(\overrightarrow{Z_I} | \pi_I) d\pi_I$, then the final result $p(Z)$ can be obtained. Where, $\overrightarrow{Z_I}$ represents the I_{th} row of matrix Z and $count$ represents the number of rounds of the I_{th} variable π_I occurrence in the Beta Process π.

5 Beta process factor analysis and the logarithmic likelihood function of the joint probability distribution for beta process

The most commonly used method in machine learning is variational inference, which is often called EM algorithm in parameter estimation. One of its core steps is to calculate the joint probability distribution function of the observed variable and the hidden variable. At the same time, the convexity of the final objective function is guaranteed by taking logarithm of the joint probability distribution function. Therefore, it is one of the most important tasks in machine learning to find the logarithmic likelihood of the joint probability distribution function. For the same reason, the logarithmic likelihood of the joint distribution of Beta Process with finite observations is also calculated below.

The key use of the Beta Process is for Beta Process Factor Analysis [John and Lawrence (2009); John and Lawrence (2011); Ishwaran and James (2001)]. Among this, the Bernoulli Process, which takes the Beta Process as a parameter, will be used for factor selection in the set of factors. Therefore, the Factor Analysis of finite observation Beta Process will be discussed in the following part.

5.1 Beta process factor analysis

Beta Process Factor Analysis is mainly described as: Define the matrix $\Phi = \left[\overrightarrow{\varphi_1}, \cdots, \overrightarrow{\varphi_G}\right]$, and define a set of vector $X = \left[\overrightarrow{x_1}, \cdots, \overrightarrow{x_k}\right]$. Here $\left\{\overrightarrow{\varphi_i}\right\}$ is the basis vector for the space of X. So $\overrightarrow{x_j}$ has the same dimension as $\overrightarrow{\varphi_i}$. Here we can set $\overrightarrow{\varphi_i} \in R^P$, that is, the X space is the P dimensional space, $\overrightarrow{x_i} \in R^P$. At the same time, we can define a matrix $W = \left[\overrightarrow{w_1}, \cdots, \overrightarrow{w_k}\right]$ so that $X = \Phi(Z \circ W) + E$ holds. Among them, define a matrix

$E=\left[\overrightarrow{\varepsilon_1},\cdots,\overrightarrow{\varepsilon_k}\right]$, satisfy $\overrightarrow{\varepsilon_i}\sim N\left(0,\sigma_\varepsilon^{-1}I_P\right)$. According to the BPFA definition, we can introduce Beta Bernoulli Process matrix Z, so that:

$$\Phi\left(Z\circ W\right)+E=\sum_{s=1}^{G}\overrightarrow{\varphi_s}\cdot\left(z_{s1}w_{s1},\cdots,z_{sk}w_{sk}\right)+\left(\overrightarrow{\varepsilon_1},\cdots,\overrightarrow{\varepsilon_k}\right)$$

$$=\sum_{s=1}^{G}\left(z_{s1}w_{s1}\overrightarrow{\varphi_s},\cdots,z_{sk}w_{sk}\overrightarrow{\varphi_s}\right)+\left(\overrightarrow{\varepsilon_1},\cdots,\overrightarrow{\varepsilon_k}\right)=\left(\sum_{s=1}^{G}z_{s1}w_{s1}\overrightarrow{\varphi_s},\cdots,\sum_{s=1}^{G}z_{sk}w_{sk}\overrightarrow{\varphi_s}\right)+\left(\overrightarrow{\varepsilon_1},\cdots,\overrightarrow{\varepsilon_k}\right)$$

$$=\left(\sum_{s=1}^{G}z_{s1}w_{s1}\overrightarrow{\varphi_s}+\overrightarrow{\varepsilon_1},\cdots,\sum_{s=1}^{G}z_{sk}w_{sk}\overrightarrow{\varphi_s}+\overrightarrow{\varepsilon_k}\right)=X=\left(\overrightarrow{x_1},\cdots,\overrightarrow{x_k}\right)$$

Through the representation of vector equality, we can get: $\overrightarrow{x_j}=\sum_{s=1}^{G}z_{sj}w_{sj}\overrightarrow{\varphi_s}+\overrightarrow{\varepsilon_j}$, $1\le j\le k$. It can be seen here that the random variable z_{sj} indicates whether the component of observation $\overrightarrow{x_j}$ contains vector $\overrightarrow{\varphi_s}$, the random variable w_{sj} represents the weight of the vector $\overrightarrow{\varphi_s}$ that makes up the observation $\overrightarrow{x_j}$.

The Beta Process Factor Analysis is dealing with: $\left(\pi_1,\cdots,\pi_k\right)\xrightarrow{generate}\left(\overrightarrow{x_1},\cdots,\overrightarrow{x_k}\right)$. That is, another matrix X is generated from one matrix Φ through vector transformation. Here, element π_f in vector $\vec{\pi}$ corresponds to a vector $\overrightarrow{x_f}$ in matrix X. Specifically, G Bernoulli 0/1 samples $\left\{z_{vj}\right\}_{v=1}^{G}$ are generated from a Bernoulli Process with a parameter of π_j, then G basis vectors $\left\{\overrightarrow{\varphi_l}\right\}_{l=1}^{G}$ need to be extracted from the corresponding space Φ to constitute $\overrightarrow{x_j}$.

Generally, in the definition of BPFA model, the prior distributions are $\overrightarrow{\varphi_s}\sim N\left(0,\sigma_\varphi^{-1}I_P\right)$, $\overrightarrow{\varepsilon_f}\sim N\left(0,\sigma_\varepsilon^{-1}I_P\right)$ and $\overrightarrow{w_f}\sim N\left(0,\sigma_w^{-1}I_G\right)$.

It should be noted that: numerical value G here can be arbitrarily large or even tend to infinity. For any countable dimensional space, the description of the space can be realized only by taking out countable basis vectors $\left\{\overrightarrow{\varphi_i}\right\}$. Therefore, as long as the number of columns of the matrix Φ is adjusted correspondingly for any countable observation, namely adding irrelevant columns to the matrix, the space of any dimension can be constituted.

The joint probability distribution function can be directly expressed as $p\left(X,Z,\Phi,\mathrm{W}\right)=p\left(Z\right)p\left(\Phi\right)p\left(\mathrm{W}\right)p\left(X\middle|Z,\Phi,\mathrm{W}\right)$.

According to the definition of distribution function, it can be directly obtained:

$$\log P(\Phi) = -\frac{1}{2\sigma_{\varphi}^{2P}} \sum_{i=1}^{G} \sum_{v=1}^{P} \varphi_{vi}^{2} - PG \log \sigma_{\varphi} - PG \log \sqrt{2\pi}$$

$$\log P(W) = -\frac{1}{2\sigma_{w}^{2G}} \sum_{j=1}^{k} \sum_{v=1}^{G} w_{vj}^{2} - kG \log \sigma_{w} - kG \log \sqrt{2\pi}$$

$$\log P(X|Z,\Phi,\mathrm{W}) = -\frac{1}{2\sigma_{\varepsilon}^{2P}} \sum_{v=1}^{P} \sum_{l=1}^{k} \left(x_{vl} - \sum_{s=1}^{G} z_{sl} w_{sl} \varphi_{vs} \right)^{2} - kP \log \sigma_{\varepsilon} - kP \log \sqrt{2\pi}$$

In this case, the probability distribution function of $p(Z)$ can be directly described by Eq. (31). It is usually straightforward to set $g_i = G$ to $\forall i$. Only X is observable here, and the rest are unobservable variables. Theoretically, when the joint probability distribution function is constructed, the work of inference and machine learning can be performed.

5.2 Logarithmic likelihood of the joint probability distribution of beta process

Through Eq. (10), the joint probability distribution function can be directly described. Then, by introducing the intermediate variable $\left\{\vec{C}, \vec{T}, d_k, R\right\}$ and performing the operation, the logarithmic likelihood of the implicit variable can be obtained:

Here, when $d_k = 2$, the distribution function of the observation sequence can be directly obtained from Eq. (8) as follows:

$$\left\{ \frac{\prod_{q=1}^{k} p(\pi_q|1,b)}{\prod_{r=C_1+1}^{k} p(\pi_r|1,b)} \right\} \left\{ \prod_{f=C_1+1}^{k} p(\pi_f|2,b) \right\} = \prod_{q=1}^{k} p(\pi_q|1,b) \times \left[\prod_{t=C_1+1}^{k} \frac{p(\pi_t|2,b)}{p(\pi_t|1,b)} \right]$$

When $d_k \geq 3$, we can calculate the distribution function of the observation sequence from Eq. (9):

$$\left[\frac{\prod_{q=1}^{k} p(\pi_q|1,b)}{\prod_{r=C_1+1}^{k} p(\pi_r|1,b)} \right] \cdot \prod_{t=2}^{d_k-1} \left[\frac{\prod_{s=\sum_{j=1}^{t-1} C_j+1}^{k} p(\pi_l|t,b)}{\prod_{s=\sum_{i=1}^{t} C_i+1}^{k} p(\pi_s|t,b)} \right] \cdot \left[\prod_{f=\sum_{m=1}^{d_k-1} C_m+1}^{k} p(\pi_f|d_k,b) \right]$$

$$= \prod_{q=1}^{k} p(\pi_q|1,b) \times \coprod_{s=1}^{d_k-1} \left\{ \prod_{t=\sum_{j=1}^{s} C_j+1}^{k} \frac{p(\pi_t|s+1,b)}{p(\pi_t|s,b)} \right\}$$

From Eqs. (8) and (9), it can be inferred that the above two terms can be uniformly expressed, that is, the conditional distribution of the Beta Process observation sequence can be uniformly expressed as:

$$p\left(\pi_1,\cdots,\pi_k\left|C_1,\cdots \mathrm{C}_{d_k-1},d_k\right.\right)=\left\{\prod_{j=1}^{k}p\left(\pi_j\left|1,b\right.\right)\right\}\times\left[\prod_{s=1}^{d_k-1}\left[\prod_{t=\sum_{j=1}^{s}C_j+1}^{k}\frac{p\left(\pi_t\left|s+1,b\right.\right)}{p\left(\pi_t\left|s,b\right.\right)}\right]\right]^{\mathrm{I}(d_k\geq 2)} \tag{32}$$

At the same time, the joint probability distribution of random variable $\left\{\left\{\pi_j\right\}_{j=1}^{k},\left\{C_i\right\}_{i=1}^{d_k-1},d_k\right\}$ can be deduced from Eqs. (5) and (32):

$$\begin{aligned}&p\left(\pi_1,\cdots,\pi_k,C_1,\cdots,C_{d_k-1},d_k\left|a,b\right.\right)=\left[\prod_{j=1}^{k}p\left(\pi_j\left|1,b\right.\right)\right]\left\{\left[1-\sum_{C_1=0}^{k-1}Poi\left(C_1\right)\right]\right\}^{\mathrm{I}(d_k=1)}\cdot\\&\left\{\prod_{s=1}^{d_k-1}\left[Poi\left(C_s\right)\prod_{t=\sum_{j=1}^{s}C_j+1}^{k}\frac{p\left(\pi_t\left|s+1,b\right.\right)}{p\left(\pi_t\left|s,b\right.\right)}\right].\left[1-\sum_{C_{d_k}=0}^{k-\sum_{l=1}^{d_k-1}C_l-1}Poi\left(C_{d_k}\right)\right]\right\}\end{aligned} \tag{33}$$

For Poisson Process, introducing random variable R can be expressed by equivalent substitution: $1-\sum_{C=0}^{N-1}Poi\left(C\right)=\sum_{R=0}^{\infty}Poi\left(N+R\right)$. Then a joint probability distribution function for $\left\{\left\{\pi_j\right\}_{j=1}^{k},\left\{C_i\right\}_{i=1}^{d_k-1},d_k,R\right\}$ can be obtained:

$$\begin{aligned}&p\left(\pi_1,\cdots,\pi_k,C_1,\cdots,C_{d_k-1},d_k,\mathrm{R}\left|a,b\right.\right)=\left[\prod_{j=1}^{k}p\left(\pi_j\left|1,b\right.\right)\right]\left\{\left[Poi\left(k+R\right)\right]\right\}^{\mathrm{I}(d_k=1)}\cdot\\&\left\{\prod_{s=1}^{d_k-1}\left[Poi\left(C_s\right)\prod_{t=\sum_{j=1}^{s}C_j+1}^{k}\frac{p\left(\pi_t\left|s+1,b\right.\right)}{p\left(\pi_t\left|s,b\right.\right)}\right].\left[Poi\left(k-\sum_{l=1}^{d_k-1}C_l+R\right)\right]\right\}^{\mathrm{I}(d_k\geq 2)}\end{aligned} \tag{34}$$

Similarly, through Eq. (4), and variable substitution $T=\ln w$, the joint distribution function of random variables $\left\{\pi,T\right\}$ can be obtained:

$$p\left(\pi_k,T_k\left|d_k,b\right.\right)=\frac{b^{d_k}}{\Gamma\left(d_k-1\right)}T_k^{d_k-2}\left(e^{-T_k}-\pi_k\right)^{b-1} \tag{35}$$

Here, the range of values for random variables can be limited to $0 \le T_k \le -\ln \pi_k$, and $d_k \ge 2$ is required at the same time.

In addition, random variable sequence $\left\{T_j\right\}_{j=1}^{k}$ is introduced in Eq. (34). The joint probability distribution function of $\left\{\left\{\pi_j\right\}_{j=1}^{k}, \left\{T_j\right\}_{j=1}^{k}, \left\{C_i\right\}_{i=1}^{d_k-1}, d_k, R\right\}$ can be obtained:

$$p\left(\pi_1,\cdots,\pi_k,T_1,\cdots,T_k,C_1,\cdots,C_{d_k-1},d_k,R \mid a,b\right)=\left[\prod_{j=1}^{k} p\left(\pi_j,T_j \mid 1,b\right)\right]\left\{\left[Poi\left(k+R\right)\right]\right\}^{\mathrm{I}(d_k=1)}$$
$$\cdot\left\{\prod_{s=1}^{d_k-1}\left[Poi\left(C_s\right)\prod_{t=\sum_{j=1}^{s}C_j+1}^{k}\frac{p\left(\pi_t,T_t \mid s+1,b\right)}{p\left(\pi_t,T_t \mid s,b\right)}\right].\left[Poi\left(k-\sum_{l=1}^{d_k-1}C_l+R\right)\right]\right\}^{\mathrm{I}(d_k\ge 2)} \tag{36}$$

Eq. (36) is the joint probability distribution function form of all random variables that are really needed. It will be calculated below to simplify its representation.

5.2.1 Quotient calculation of $\left\{\pi,T\right\}$ joint probability

According to Eq. (35), the division of the two terms can be directly calculated. When $s \ge 2$, it can be deduced:

$$\frac{p\left(\pi_t,T_t \mid s+1,b\right)}{p\left(\pi_t,T_t \mid s,b\right)}=\frac{\frac{b^{s+1}}{\Gamma\left(s\right)}T_t^{s-1}\left(e^{-T_t}-\pi_t\right)^{b-1}}{\frac{b^{s}}{\Gamma\left(s-1\right)}T_t^{s-2}\left(e^{-T_t}-\pi_t\right)^{b-1}}=\frac{b}{s-1}T_t \tag{37}$$

When $s=1$, it can be calculated that:

$$\frac{p\left(\pi_t,T_t \mid 2,b\right)}{p\left(\pi_t,T_t \mid 1,b\right)}=\frac{b^2\left(e^{-T_t}-\pi_t\right)^{b-1}}{b\left(1-\pi_t\right)^{b-1}}=b\left(\frac{e^{-T_t}-\pi_t}{1-\pi_t}\right)^{b-1} \tag{38}$$

By substituting Eqs. (37) and (38) into Eq. (36), the joint probability distribution function of all required random variables can be obtained:

$$p\left(\pi_1,\cdots,\pi_k,T_1,\cdots,T_k,C_1,\cdots,C_{d_k-1},d_k,R\mid a,b\right)=\left[\prod_{j=1}^{k}b\left(1-\pi_j\right)^{b-1}\right]\left\{\left[Poi\left(k+R\right)\right]\right\}^{\mathrm{I}(d_k=1)}$$
$$\cdot\left\{\left[Poi\left(C_1\right)Poi\left(k-\sum_{l=1}^{d_k-1}C_l+R\right)\prod_{t=C_1+1}^{k}\left[b\left(\frac{e^{-T_t}-\pi_t}{1-\pi_t}\right)^{b-1}\right]\right]\right\}^{\mathrm{I}(d_k>1)}$$
$$\cdot\left\{\prod_{s=2}^{d_k-1}\left[Poi\left(C_s\right)\prod_{t=\sum_{j=1}^{s}C_j+1}^{k}\left[\left(\frac{b}{s-1}T_t\right)\right]\right]\right\}^{\mathrm{I}(d_k>2)} \tag{39}$$

5.2.2 Logarithmic likelihood of joint probability distribution function

Taking the logarithm of Eq. (39) here, we can deduce:

$$\begin{aligned}
&\log p\left(\vec{\pi},\vec{T},\vec{C},d_k,R\right)\\
&=\sum_{j=1}^{k}\left\{\log b+\left(b-1\right)\log\left(1-\pi_j\right)\right\}+\mathrm{I}\left(d_k=1\right)\log Poi\left(k+R\right)\\
&+\mathrm{I}\left(d_k>1\right)\left\{\sum_{t=C_1+1}^{k}\left[\log b+\left(b-1\right)\log\left(e^{-T_t}-\pi_t\right)-\left(b-1\right)\log\left(1-\pi_t\right)\right]\right.\\
&\left.+\log Poi\left(C_1\right)+\log Poi\left(k-\sum_{l=1}^{d_k-1}C_l+R\right)\right\}\\
&+\mathrm{I}\left(d_k>2\right)\left\{\sum_{s=2}^{d_k-1}\left[\log Poi\left(C_s\right)+\sum_{t=\sum_{j=1}^{s}C_j+1}^{k}\log b+\sum_{t=\sum_{j=1}^{s}C_j+1}^{k}\log T_t-\sum_{t=\sum_{j=1}^{s}C_j+1}^{k}\log\left(s-1\right)\right]\right\}
\end{aligned} \tag{40}$$

The structure of joint log-likelihood is analyzed below.

First of all, the coefficient of item $\log b$ in Part $d_k>2$ is calculated, and we can get:

$$\sum_{s=2}^{d_k-1}\sum_{t=\sum_{j=1}^{s}C_j+1}^{k}1=\left(d_k-2\right)k-\sum_{s=2}^{d_k-1}\sum_{j=1}^{s}C_j \tag{41}$$

With Eq. (41), the calculation of C_j can be directly completed:

$$\sum_{s=2}^{d_k-1}\sum_{j=1}^{s} C_j = d_k \sum_{j=1}^{d_k-1} C_j - \sum_{j=1}^{d_k-1} jC_j - C_1 \tag{42}$$

Then, the cumulative calculation of the $\log(s-1)$ items in Part $d_k > 2$ can be completed:

$$\begin{aligned}&\sum_{s=2}^{d_k-1}\sum_{t=\sum_{j=1}^{s} C_j+1}^{k} \log(s-1)\\ &= k\left(\log 1+\log 2+\cdots+\log(d_k-2)\right)-\log 1\sum_{j=1}^{2} C_j-\log 2\sum_{j=1}^{3} C_j-\cdots-\log(d_k-2)\sum_{j=1}^{d_k-1} C_j\end{aligned} \tag{43}$$

For the calculation of the second half of Eq. (43), it can be described as:

$$\begin{aligned}&\log 1\sum_{j=1}^{2} C_j+\log 2\sum_{j=1}^{3} C_j+\cdots+\log(d_k-2)\sum_{j=1}^{d_k-1} C_j\\ &= \log\Gamma(d_k-1)\sum_{j=1}^{d_k-1} C_j-\sum_{s=3}^{d_k-1} C_s\log\Gamma(s-1)\end{aligned} \tag{44}$$

By substituting Eq. (44) into Eq. (43), we can obtain:

$$\begin{aligned}&\sum_{s=2}^{d_k-1}\sum_{t=\sum_{j=1}^{s} C_j+1}^{k} \log(s-1) = k\log\Gamma(d_k-1)-\left(\sum_{j=1}^{d_k-1} C_j\log\Gamma(d_k-1)-\sum_{s=3}^{d_k-1} C_s\log\Gamma(s-1)\right)\\ &= k\log\Gamma(d_k-1)-\sum_{j=1}^{d_k-1} C_j\log\Gamma(d_k-1)+\sum_{s=3}^{d_k-1} C_s\log\Gamma(s-1)\end{aligned} \tag{45}$$

By substituting Eqs. (41), (42) and (45) into Eq. (40), the following results can be obtained:

$$
\begin{aligned}
&\log p\left(\vec{\pi},\vec{T},\vec{C},d_k,R\right)\\
&=k\log b+\left(b-1\right)\sum_{j=1}^{k}\log\left(1-\pi_j\right)+\mathrm{I}\left(d_k=1\right)\log Poi\left(k+R\right)\\
&+\mathrm{I}\left(d_k>1\right)\left(k-C_1\right)\log b+\mathrm{I}\left(d_k>1\right)\left(b-1\right)\sum_{t=C_1+1}^{k}\log\left(e^{-T_t}-\pi_t\right)\\
&-\mathrm{I}\left(d_k>1\right)\left(b-1\right)\sum_{t=C_1+1}^{k}\log\left(1-\pi_t\right)+\mathrm{I}\left(d_k>1\right)\log Poi\left(C_1\right)\\
&+\mathrm{I}\left(d_k>1\right)\log Poi\left(k-\sum_{l=1}^{d_k-1}C_l+R\right)+\mathrm{I}\left(d_k>2\right)\sum_{s=2}^{d_k-1}\log Poi\left(C_s\right)\\
&+\mathrm{I}\left(d_k>2\right)\log b\left[\left(d_k-2\right)k-\left(d_k\sum_{j=1}^{d_k-1}C_j-\sum_{j=1}^{d_k-1}jC_j-C_1\right)\right]\\
&+\mathrm{I}\left(d_k>2\right)\sum_{s=2}^{d_k-1}\sum_{t=\sum_{j=1}^{s}C_j}^{k}\log T_t\\
&-\mathrm{I}\left(d_k>2\right)\left[k\log\Gamma\left(d_k-1\right)-\sum_{j=1}^{d_k-1}C_j\log\Gamma\left(d_k-1\right)\right]\\
&-\mathrm{I}\left(d_k>3\right)\sum_{s=3}^{d_k-1}C_s\log\Gamma\left(s-1\right)
\end{aligned}
\tag{46}
$$

Regarding Poisson distribution, it can be calculated as:

$$
\log Poi\left(C_i\right)=\log\frac{\left(\frac{a\gamma}{b}\right)^{C_i}}{\Gamma\left(C_i+1\right)}e^{-\frac{a\gamma}{b}}=C_i\log a+C_i\log\gamma-C_i\log b-\log\Gamma\left(C_i+1\right)-\frac{a\gamma}{b}
\tag{47}
$$

By substituting Eq. (47) into Eq. (46), we can obtain the final conclusion:

$$
\begin{aligned}
&\log P\left(\vec{\pi},\vec{C},\vec{T},d_k,R\right)\\
&=\left(\mathrm{I}\left(d_k=2\right)-\mathrm{I}\left(d_k>2\right)\right)k\log b+k\log a+R\log a+k\log\gamma+R\log\gamma-R\log b\\
&-\mathrm{I}\left(d_k=1\right)\log\Gamma\left(k+R+1\right)+\left(2\mathrm{I}\left(d_k>2\right)-\mathrm{I}\left(d_k=2\right)-1\right)\frac{a\gamma}{b}-\mathrm{I}\left(d_k=2\right)C_1\log b\\
&+\left(b-1\right)\sum_{j=1}^{k}\log\left(1-\pi_j\right)+\mathrm{I}\left(d_k>1\right)\left(1-b\right)\sum_{t=C_1+1}^{k}\log\left(1-\pi_t\right)+\mathrm{I}\left(d_k>1\right)\left(b-1\right)\sum_{t=C_1+1}^{k}\log\left(e^{-T_t}-\pi_t\right)\\
&-\mathrm{I}\left(d_k>1\right)\sum_{s=1}^{d_k-1}\log\Gamma\left(C_s+1\right)-\mathrm{I}\left(d_k>1\right)\log\Gamma\left(k-\sum_{l=1}^{d_k-1}C_l+R+1\right)-\mathrm{I}\left(d_k>2\right)d_k\frac{a\gamma}{b}\\
&+k\log b\mathrm{I}\left(d_k>2\right)d_k-d_k\log b\mathrm{I}\left(d_k>2\right)\sum_{j=1}^{d_k-1}C_j+\mathrm{I}\left(d_k>2\right)\log b\sum_{j=1}^{d_k-1}jC_j\\
&+\mathrm{I}\left(d_k>2\right)\sum_{s=2}^{d_k-1}\sum_{t=\sum_{j=1}^{s}C_j+1}^{k}\log T_t-k\mathrm{I}\left(d_k>2\right)\log\Gamma\left(d_k-1\right)+\mathrm{I}\left(d_k>2\right)\log\Gamma\left(d_k-1\right)\sum_{j=1}^{d_k-1}C_j\\
&-\mathrm{I}\left(d_k>3\right)\sum_{s=3}^{d_k-1}C_s\log\Gamma\left(s-1\right)
\end{aligned}
\tag{48}
$$

Then, by the definition of observation sequence of Beta Bernoulli Process, Formula $p\left(z_{sl}\left|\pi_l\right.\right)=\pi_l^{z_{sl}}\left(1-\pi_l\right)^{1-z_{sl}}$ and $p\left(Z\left|\vec{\pi}\right.\right)=\prod_{s=1}^{G}\prod_{l=1}^{k}p\left(z_{sl}\right)=\prod_{s=1}^{G}\prod_{l=1}^{k}\pi_l^{z_{sl}}\left(1-\pi_l\right)^{1-z_{sl}}$ can be obtained directly. Namely:

$$\log P\left(Z\left|\vec{\pi}\right.\right)=\sum_{s=1}^{G}\sum_{l=1}^{k}\left(z_{sl}\log\pi_l+\left(1-z_{sl}\right)\log\left(1-\pi_l\right)\right).$$

Because of the conditional independence between random variables, it can be obtained directly through substitution derivation: $P\left(Z\left|\vec{\pi},\vec{C},\vec{T},d_k,R\right.\right)=p\left(Z\left|\vec{\pi}\right.\right)$.

In this case, the logarithmic likelihood function of the final distribution can be obtained by adding the above results:

$$
\begin{aligned}
&\log p\left(X,Z,\Phi,\mathrm{W},\vec{\pi},\vec{C},\vec{T},d_k,R\right)=\\
&\log p\left(Z\left|\vec{\pi}\right.\right)+\log p\left(\Phi\right)+\log p\left(\mathrm{W}\right)+\log p\left(X\left|Z,\Phi,\mathrm{W}\right.\right)+\log p\left(\vec{\pi},\vec{C},\vec{T},d_k,R\right).
\end{aligned}
$$

In which $\left\{\vec{\pi},\vec{C},\vec{T},d_k,R\right\}$ are all implicit variables. In this way, parameters can be learned by variational EM algorithm.

6 Discussion

Theoretically, the joint probability distribution function must be able to handle any number of observations, and, importantly, the number of actual observations can be arbitrarily large, but not infinite. The method that we have given here is simple and effective in dealing with this problem, because the Nonparametric Bayesian stochastic Process we discussed here does not satisfy the Kolmogorov consistency theorem, so lead to the relationship between observed variables is not independent identically distributed. The distribution function form is much more complicated than the traditional machine learning situation, and the number of unobserved variables has a direct impact on the form of the distribution function. The method proposed here eliminates the information irrelevant to observations and thus gives the general form of any finite number of observations.

We have obtained several results of this idea through the Stick-Breaking structure proposed by Paisley et al. [Paisley and Zaas (2010)]: including the more general construction of finite observation and the new type of joint probability distribution function for Stick-Breaking Beta Processes, which indicates that the Beta Process is the superposition of a Poisson Process countable set and used as a priori of Bernoulli Process. Finally, a finite observation of a 0/1 matrix is completed.

In the future, we will extend the proposed method and use variational inference method to solve the problem that the accurate estimation of marginal distribution is too complex, so as to be applicable to the machine learning task of approximate parameter estimation of Beta Process and Beta Bernoulli Process. We will also explore some approximate inference models of distribution functions of Non-parametric process variables, hoping to obtain better and simplified performance by means of variational inference method. These similar methods can also be used for Gamma Processes [Anirban and Brian (2015)] and Gamma Poisson Processes [Michalis and Titsias (2007)]. This is the next step of our consideration.

Regression analysis is one of the main research directions in the field of machine learning. At present, Gaussian Process Regression is the main regression method when stochastic process is used as the tool. Among them, Kalman Filter is the most widely used field in Gaussian Process Regression. Based on the same idea, because for the Beta Process, when the joint probability distribution function is given, the conditional probability can be calculated according to the Bayesian formula and the regression analysis can be carried out theoretically. Therefore, the regression analysis of Beta Process is also one of the directions to be considered in the next step.

The idea described above can be also used in the context of Gamma Processes similar to Beta Processes, so our results also contribute to the establishment of a general Non-parametric Bayesian inference mechanism.

A more common variant of the Beta Bernouilli Process is the Indian Buffet Process (IBP), which learns the number of features included in the model from the observed data, thus allowing the model to interpret the data more accurately. The Non-parametric Bayesian model based on IBP can automatically learn the implicit features, and can in a scalable way to determine the number of features. Therefore, in theory, better prediction performance can be achieved. In practical applications, the 0/1 output of the Beta Bernouilli Process is generally used to describe the relationship between entities. In the sample matrix of the Beta Bernoulli Process, a specific entity is described by a set of binary features, and then

the features are obtained from the observations. And try to infer the features. The sample matrix value of the Beta Bernoulli Process can be used as a basis for determining whether the entities are related. If the weight is attached to the 0/1 output of the Beta Bernouilli Process at the same time, the strength of the influence between the entities can be added while describing the correlation between the entities.

Since the distribution of the Beta Bernouilli Process is long-tailed, and the distribution functions for each round generated by the Beta Process Stick Breaking do not necessarily have the same attenuation trend as the power-rate distribution, resulting in the model being basically sufficient to describe the entity possessing any number of features. The general Beta Bernouilli Process describes the probability distribution, which can be used to describe the relationships between entities, and the relationships are not necessarily symmetric. This asymmetry relationship can be applied to some important issues such as social network connection prediction. Connection prediction is an important issue in social network modeling [Miller, Michael and Thomas (2009)]. Here, it can be assumed that the link probability from one node to another node is determined by the combined effect of pairwise feature interactions. If a weight is added to the 0/1 sample matrix of the Beta Bernouilli Process, and the positive weight corresponds to the probability of high correlation, while the negative weight corresponds to the probability of low correlation, and the zero weight indicates that there is no correlation between the two features, then the representation ability of the model will be greatly improved, and the influence relationship between nodes will have stronger performance.

The relationship between entities can be simplified. The simplified symmetric relationship is used to learn a complete symmetric weight matrix. The symmetric Beta Bernouilli Process model can also be used to describe the co-authorship relationship judgment in text mining, because the co-authorship relationship is symmetric [Teh, Jordan, Beal et al. (2006)].

Currently, IBP Process with multiple levels proposed by the academia has been applied in Deep Learning. It is used to learn the structure of Deep Belief Network, including the number of layers of neurons, the number of neurons in each layer, and the connection structure of neurons between layers [Adams, Hanna and Zoubin (2010)].

In this paper, the exact analytical form of probability distribution function of finite arbitrary dimension is directly analyzed for Beta Bernouilli Process, and its properties as the objective function of machine learning are discussed. In the next step of prediction [Miller, Michael and Thomas (2009)] and learning [Adams, Hanna and Zoubin (2010)], it can be directly used as the prior probability distribution function of the discriminant model [Miller, Michael and Thomas (2009)] and substituted into the objective function for parameter optimization.

Since the marginal probability distribution function defined here is accurate, the calculation process of parameter optimization can be carried out based on demand, or we can directly optimize the precise distribution of marginal probability, or choose sampling [Miller, Michael and Thomas (2009)] and variational inference to make approximate inference to the joint probability distribution function.

In Deep Learning, ideas similarity of Teh et al. [Teh, Jordan, Beal et al. (2006)] can also be used to conduct Deep Learning reasoning by taking the row number and column number of each row of the binary matrix output by the Beta Bernouilli Process as the layer number of multi-layer neural network and the node number of each layer. The two-parameter beta

process description adopted in this paper theoretically promotes the model in [Adams, Hanna and Zoubin (2010)], which directly adopted the Indian Buffet Process as the priori of the number of layers and the number of nodes of each layer of Deep Belief Network.

7 Conclusions

Beta Process contains a list of random variables. However, these random variables do not satisfy the stationarity or the global independent increment, so the probability distribution of these random variables has an extremely complex form. The stick breaking construction method is a means to indirectly define the Beta Process by describing the sampling process. Further analyzing and deducing the joint probability distribution of observed samples through the described sampling method is the next step necessary for machine learning.

The result presented here is an analytical method for directly calculating the probability distribution of observable variables in Beta Process. Through probability distribution calculation, on the one hand, all intermediate variables are directly marginalized, thus completely eliminating the unobservable information. On the other hand, the observation of Bernoulli Process directly generated by taking Beta Process as a parameter can have the form of analytic probability distribution function.

In the future, we will further extend the derived results and deal with the later steps of machine learning on the Beta Process.

References

Acharya, A.; Teffer, D.; Henderson, J.; Tyler, M.; Zhou, M. et al. (2015): Gamma process poisson factorization for joint modeling of network and documents. *Joint European Conference on Machine Learning & Knowledge Discovery in Databases*, pp. 283-299.

Adams, R. P.; Wallach, H. M.; Ghahramani, Z. (2010): Learning the structure of deep sparse graphical models. *Proceedings of the13th IEEE International Conference on Artificial Intelligence and Statistics*, pp. 1-8.

Alqifari, H. N.; Coolen, F. P. (2019): Robustness of nonparametric predictive inference for future order statistics. *Journal of Statistical Theory and Practice*, vol. 13, no. 1, pp. 12.

Arfe, A.; Peluso, S.; Muliere, P. (2018): The semi-Markov beta-Stacy process: a Bayesian non-parametric prior for semi-Markov processes. arXiv:1812.00260.

Antoniak, C. E. (1974): Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, vol. 2, no. 6, pp. 1152-1174.

Doshi, F.; Miller, K.; Van Gael, J.; Teh, Y. W. (2009): Variational inference for the indian buffet process. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pp. 137-144.

Doshi-velez, F.; Mohamed, S. (2009): Large scale nonparametric Bayesian inference: data parallelisation in the Indian buffet process. *Proceedings of the 21th Annual Conference on Neural Information Processing Systems*, pp. 1294-1302.

Griffin, J. E.; Kalli, M.; Steel, M. F. (2018): Discussion of "nonparametric bayesian inference in applications": bayesian nonparametric methods in econometrics. *Statistical Methods and Applications*, vol. 27, no. 2, pp. 207-218.

Griffiths, T. L.; Ghahramani, Z. (2011): The indian buffet process: an introduction and review. *Journal of Machine Learning Research*, vol. 12, no. 4, pp. 1185-1224.

Ishwaran, H.; James, L. F. (2001): Gibbs sampling methods for stick-breaking priors. In *Journal of American Statistical Association*, vol. 96, no. 453, pp. 161-173.

Lee, H. J.; Hong, K. S. (2016): Class-specific mid-level feature learning with the Discriminative Group-wise Beta-Bernoulli process restricted Boltzmann machines. *Pattern Recognition Letters*, vol. 80, pp. 8-14.

Liu, Z.; Yu, L.; Sun, H. (2016): Image restoration via bayesian dictionary with nonlocal structured beta process. *Journal of Visual Communication and Image Representation*, vol. 52, pp. 159-169.

Miller, K. T.; Jordan, M. I.; Griffiths, T. L. (2009): Nonparametric latent feature models for link prediction. *Proceedings of the 21th Annual Conference on Neural Information Processing Systems*, pp. 1276-1284.

Nalisnick, E.; Smyth, P. (2017): Stick-breaking variational autoencoders. *International Conference on Learning Representations*. arXiv:1605.06197.

Paisley, J.; Zaas, A. (2010): A stick-breaking construction of the beta process. *Proceedings of the 27th International Conference on Machine Learning*, pp. 847-854.

Paisley, J.; Blei, D.; Jordan, M. I. (2012): Stick-breaking beta processes and the Poisson process. *Proceeding of the 15th International Conference on Artificial Intelligence and Statistics*, pp. 850-858.

Paisley, J.; Carin, L. (2009): Nonparametric factor analysis with Beta process paiors. *Proceeding of Annual International Conference on Machine Learning*, pp. 777-784.

Paisley, J.; Carin, L. (2011): variational inference for stick-breaking beta process priors. *Proceedings of the 28th International Conference on Machine Learning*, pp. 889-896.

Paisley, J.; Jordan, M. (2016): A constructive definition of the beta process. arXiv:1604.00685.

Roychowdhury, A.; Kulis, B. (2015): Gamma processes, stick-breaking, and variational inference. *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 800-808.

Seo, S.; Wallat, M.; Graepel, T.; Obermayer, K. (2000): Gaussian process regression: active data selection and test point rejection. *IEEE-INNS-ENNS International Joint Conference on Neural Networks*, pp. 241-246.

Stevens, A.; Pu, Y.; Sun, Y. *(*2017*):* Tensor-dictionary learning with deep kruskal-factor analysis. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 121-129.

Titsias, M. K. (2007): The infinite gamma-poisson feature model. *Proceedings of the 21th Annual Conference on Neural Information Processing Systems*, pp. 1513-1520.

Thibaux, R.; Jordan, M. I. (2007): Hierarchical beta processes and the Indian buffet process. *International Conference on Artificial Intelligence and Statistics*, pp. 564-571.

Teh, Y. W.; Görür, D.; Ghahramani, Z. (2007): Stick-breaking construction for the Indian Buffet Process. *Artificial Intelligence and Statistics*, pp. 556-563.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; Blei, D. M. (2006): Hierarchical dirichlet processes. *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566-1581.

Yao, Y.; Vehtari, A.; Simpson, D.; Gelman, A. (2018): Yes, but did it work? Evaluating variational inference. *International Conference on Machine Learning*, pp. 5577-5586.