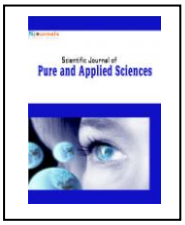
Sjournals
Scientific Journals
Scientific Journal of Pure and Applied Sciences (2014) 3(8) 809-815
ISSN 2322-2956
doi: 10.14196/sjpas.v3i8.1516

Contents lists available at Sjournals
Scientific Journal of
Pure and Applied Sciences
Journal homepage: www.Sjournals.com



Original article

The fundamental problem of gibbs sampler in mixture models

G.H. Gholami[a,*], A. Etemadi[b], H. Rasi[c]
Department of Mathematics, Faculty of science, Urmia University, Urmia, IRAN.
Department of Mathematics, Urmia Branch, Islamic Azad University, Urmia, IRAN.
Department of Statistics, Faculty of Mathematics, Tabriz University, Tabriz, IRAN.

*Corresponding author; Department of Mathematics, Faculty of science, Urmia University, Urmia, IRAN.

ARTICLE INFO

Article history,
Received 11 July 2014
Accepted 22 August 2014
Available online 29 August 2014

Keywords,
Gibbs sampler
Mixture models
Latent variable
Posterior distribution
Prior distribution

ABSTRACT

The mixture models were firstly studied by Pearson in 1894. These models are strong tools, through which the complicated systems can be analyzed in a wide range of disciplines such as Astronomy, Economics, Mechanics, etc. although the structure of these models is apparently simple, it is very complicated to obtain maximum likelihood estimators and Bayesian ones in particular and it needs to be approximated in most cases. In this paper, we apply the Gibbs Sampling in order to approximate the Bayesian Estimator in Mixture models, present the Gibbs algorithms for the family of exponential distributions and finally, we would show the disadvantage of this algorithm through an example.

1. Introduction

Assume that a population includes a number of sub-populations. Through considering a common distribution for all of these subpopulations, they would be indexed due to one or more parameters such as average, Variance, etc. if we select a sample from this general population, it can be chosen from each subpopulation. Assume that we have k subpopulations which are represented as A1, j = 1... K. also assume that B is the represents the event of choosing a sample from this general population. According to the Law of total probability:

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(A_j)\, P(B|A_j) \quad (1)$$

In the representation above, P (A_j) is the possibility of choosing each of the subpopulations and P$(B|A_j)$the possibility of the considered sample being from j population. In fact, we firstly choose one of the sub-population with the probability of P (A_j), and then, we take a sub-sample from this subpopulation. If B is the event of X = x random variant and P (A_j) = P (J = j) where J is the index of J subpopulation:

$$P(X = x) = \sum_{j=1}^{k} P(J = j).(P(X = x|J = j)$$

The rewritten form of the relation above is:

$$fx(x) = \sum_{j=1}^{k} p_j f_j(x) \qquad (2)$$

This model is called Finite Mixture Model in which p_j is the probability of the sairiple being from j subpopulation and fj(x) represents the density function for this subpopulation. If the number of the components of (k) model is infinite, then it is called au Infinite Mixture Model. Such models were firstly studied by Pearson in 1894. They have performed strong tools, through which the complicated systems can be analyzed in a wide range of disciplines such as Astronomy, Economics, Mechanics, etc.

In this paper, we assume that each of the components (fj (.)) is characterized by a vector of θ_j, the corresponding parameter:

$$fx(x) = \sum_{j=1}^{k} p_j f_j(x|\theta_j) \qquad (3)$$

This representation is called the density function of the parametric mixture model. The parameter space is

$$\Theta = R^k \times [0,1]^k$$

Although this definition of mixture model is completely simple and rather initial, its estimations of maximum likelihood (if there is any) or Bayesian estimators are not easy. Assume that, x = (x1 …. xn)

is the iid observation we have from the model three with the parameters:

p= (p1,…,pk) , $\theta = (\theta_1, \dots, \theta_k)$

The complete (accurate) calculation of posterior distribution and representation of average posterior in particular in a clear forum requires the expansion of the likelihood function:

$$L(\theta, p|x) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f_j(x_i|\theta_j) \qquad (4)$$

Proportional to the total rate of kn . Except for several special conditions, the calculation of this likelihood function is not possible for a large number of observations. This substantial computational problem causes us to apply approximate algorithm such as MCMC for the model approximation.

If the statistical inference merely aims to estimate the parameters, the case is called the Parameter Estimation Problem. The number of the components of the model might be unknown, if so the parametric spatial dimension would he indeterminate as well. In such a case, the statistical inference on the model is much harder compared to the previous condition. The estimation of the number of these components is called Model Learning process. Some times the number of components can he specified previously and the inference mainly aims to allocate each of the samples to each subpopulation. This kind of inference is called Clustering.

2. The structure of the latent variable

Considering the law of total probability, all of the models can be shown as:

$$f(x|\theta) = \int g(x, z|\theta)dz$$

This method is known as marginalization process. In this representation, x is the observation variable and z is called the latent variable. Note that this is a continuous representation of the law of total probability. Applying the conjugation (completion) of the variant X, and the latent variable, z , we extend the missing data, zi for each xi observation:

P (zj=j) =pj xi|zi=z~f (x|θ_j) j=1,..., k i=1,...,m,

zi are also called the label variables for their value specifies that the related xj has. Come from which subpopulation. The (x, z) pair is called a complete data, because having this pair enables us to tell that each variant belongs to which component of the mixed distribution.

As it was mentioned above, the likelihood function includes kn terms. Theoretically, this means that the zi latent variable is considered in calculation of likelihood function along with all of its values. Since the likelihood function of a given value (θ, p) can be calculated from the nk stage of the operation, the computational problems derived by extended version of the 4 equation, prevent the performance of an accurate classic and Bayesian extension.

According to the prior 'π(θ,p) for the parametric vector of (θ,p), the posterior distribution would be given:

$$\pi(\theta, p|x) \propto \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f_j(x_i|\theta_j)\pi(\theta, p)$$

The right side of the equation above has an operational stage of nk. Like the likelihood function, obtaining an intuitive posterior distribution is not possible without the extension of the right side of the equation.

The z = (z1,. . . , zn) is considered as the latent variable vector. zi i = 1, . . . , n can have each of the k various values (i.e. each observation can belong to any of the subpopulations). This vector is also known as allocation vector. The set of all kn allocations of this vector is represented by x. The following partition has been considered for the vector (n1,. . . , nk) in which nj is the share of the j subpopulation from the sample:

$$\chi_j = \left\{ z: \sum_{i=1}^{n} II_{(z_i=1)} = n_1, \ldots, \sum_{i=1}^{n} II_{(z_i=k)} = n_k \right\}$$

The j index in the mentioned partition indexes the different states of the vector (n1,. . . , nk). The total number of these states is equal to the number of the analytical non-negative right answers, n1, proportional to k parts which require the condition of n1 + ... + nk = n. this number is equal to:

$$r = \binom{n + k - 1}{n}$$

Therefore, j = 1,…,r would change. Association of j with an ideal combination of the vector (n1 . . . , nk) call be performed fully arbitrary. What is important is that the value of j can only be corresponding to one vector.

One methods to correspond the value of j and the vector (n1. . . , nk) is to consider time alphabetic order on this vector. This is also known as an arrangement according to the dictionary.

For instance, time vector (a2, b2, c2) is considered bigger than i (a1, b1, c1), whenever on of time following three conditions is required:

1 a1 <a2

2 if a1 = a2 then b1 < b2

3 if a1 = a2, b1 = b2 then c1 < c2

Thus, all of the different partitions of z can he represented as:

$$\chi = \bigcup_{j=1}^{r} \chi_j$$

Note that the one-to-one correspondence between the different values of j and the vector (n1. . . , nk) would not be transferred to the vector z, so there can be more than one z vector for each j value.

Applying these partitions, the posterior distribution can be rewritten as:

$$\pi(\theta, p|x) = \sum_{i=1}^{r} \sum_{z \in \chi_j} W(z)(\theta, p|x, z) \qquad (5)$$

Where W(z) represents the posterior possibility of the z vector.

(x,z) is the complete data, and $\pi(\theta, p|x, z)$ is its posterior distribution. Considering the full likelihood function, this full posterior distribution is obtained:

$$L(\theta, p|x, z) = \prod_{i=1}^{n} f(x_i, z_i|\theta) = \prod_{i=1}^{n} Pz_i f(x_i|\theta z_i)$$

Considering the expected value of equation 5, the posterior estimator which indicates the Bayes estimator of (θ,p) is:

$$E(\theta, p|x) = \int (\theta, p) \sum_{i=1}^{r} \sum_{z \in \chi_j} W(z)\pi(\theta, p|x) d_{(\theta,p)}$$

$$= \sum_{i=1}^{r} \sum_{z \in \chi_j} (\theta, p)\pi(\theta, p|x) d_{(\theta,p)}$$

$$= \sum_{i=1}^{r} \sum_{z \in \chi_j} W(z) E^n[\theta, p|x, z]$$

Inferentially, it is very important to breakdown the equation 5. This indicates that the posterior distribution would firstly consider all of the possible z allocations of the date, then, it attributes the posterior possibility of W(z) to each of then amid finally, the posterior distribution of 'π(θ, p|x, z) would be grounded based on this allocation.

3. The approximate inference for mixture models

In this part, the approximate inference for the mixture models and its problems would be investigated. In order to perform an approximate inference the Gibbs algorithm would he applied.

3.1. The gibbs algorithm for mixture models

Gibbs sampling is one of the more usable methods in Bayesian inference of the mixture models. For this type of sampling take advantages of the structure of latent data. Gibbs sampling for the mixture models is based on a continuous simulation of z, p, θ distributions providing for one another and the data. I.e. we first obtain the total conditional distributions of parameters and the latent variable, z, and then, we perform the continuous sampling from them. If prior p and θ are independemit, the full conditional distributions of p and θ are:

$$\pi(p|x, z) = \pi(p|z)$$
$$\pi(\theta|x, p, z) = \pi(p|z)$$

This algorithm can be coded in algorithm 1.

The investigation of the convergence of this algorithm is not easy in spite of its apparent simplicity. In fact, when according to Ergodic theory, the chain is geometrically uniform, aim increase iii its dimension resulted by the stage of adding z variable, can cause substantial problems for the chain convergence. One of time natural characteristics of Gibbs algorithm due to time mixture models is that the chain might be place in a situation that its produced values convene around a local maximum. In order to get out such situations a large number of repetitions are needed. This situation is called a trap.

If the conjugate prior is applied for Pjs , it would be very easy to sample them. On the contrary, simulation of θ_js extremnely depends on sampling distribution of f (.$|\theta_j$) amid the prior π(.).

The marginal distribution of zis is a polynomial one which belongs to the exponential family of distributions. Therefore, we can apply conjugate prior which means time Dirichlet distribution.

While sampling, the Gibbs algorithm for the mixture models distribution belongs to the exponential

Algorithm 1 mixture gibbs sampler

(0) Initialization: Choose p(0) and $\theta^{(0)}$arbitrarily.

(1) Iteration t :for t = 1,2....

(1.1) generate $z_i^{(t)}$ such that:

$P(z_i^{(t)}=j|p_j^{(t-1)}, p_j^{(t-1)}, x_i) \propto p_j^{(t-1)} f(x_i|\theta_j^{(t-1)})$

(1.2) Generate p(t) according to $\pi(p|z^{(t)})$

(1.3) Generate $\theta^{(t)}$ according to $\pi(\theta|z^{(t)}, x)$

family and the conjugate prior would be in form of the algorithm 2:

Algorithm 2 mean mixture gibbs sampler.

(0) Initialization: Choose p(0) and $\theta^{(0)}$ arbitrarily.

(1) Iteration t : for t = 1, 2,…

(1.1) Generate $z_i^{(t)}$from:

$P(z_i^{(t)})=j|p_j^{(t-1)}, p_j^{(t-1)}, x_i) \propto p_j^{(t-1)} f(x_i|\theta_j^{(t-1)})$

(1.2) compute

$$n_i^{(t)} = \sum_{i=1}^{n} II_{z_i^{(t)}} = j \qquad s_j^{(t)} = \sum_{i=1}^{n} II_{z_i^{(t)}} t(x_i)$$

(1.3) Generate p(t) from:

$$D(\gamma_1 + n_1, \dots, \gamma_k + n_k)$$

(1.4) Generate for j = 1,. . . , k according to:

$$\pi\left(\theta_j \middle| z^{(t)}, x\right) \propto \exp\left(R\left(\theta_J\right).\left(\alpha + s_j^{(t)}\right) - \psi(\theta_J)(n_j + \beta)\right)$$

Gibbs algorithm would be operated in the following example, assuming that the value of p parameter is determinate. The example shows that Gibbs algorithm is not adequately capable of traversing along the parametric space and it might even he trapped in very large number of repetitions in a parametric area.

Example 1 assume that (x1,. . . , xn,) is a random sample of the following mixture distribution:

$$0.2N(\mu_1, 1) + 0.8N(\mu_2, 1)$$

Considering the normal independent prior distributions, the posterior distributions for μ_1 and, μ_2 are calculated:

$$N\left(\frac{\lambda\delta + 1x_1(z)}{\lambda + 1}, \frac{1}{\lambda + 1}\right)$$

$$N\left(\frac{\lambda\delta + (n-1)x_2(z)}{\lambda + n - 1}, \frac{1}{\lambda + n - 1}\right)$$

Thus 'we can sample the posterior distributions applying algorithm 3.

Algorithm 3 Mixture Gibbs Sampler for example 3

(0) Initialization: choose $\mu_1^{(0)}$ and $\mu_2^{(0)}$

(1) Step t: fort=1,2....

(1.1) Generate $z_i^{(t)}$ from:

$$P(z_i = 1) \propto p \exp\left\{-\frac{1}{2}\left(x_i - \mu_1^{(t-1)}\right)^2\right\}$$

$$P(z_i = 2) \propto (1-p)\exp\left\{-\frac{1}{2}\left(x_i - \mu_2^{(t-1)}\right)^2\right\}$$

(1.2) compute:

$$1 = \sum_{i=1}^{n} II_{z_i^{(t)}} = 1 \qquad \bar{x}_j((z)) = \sum_{i=1}^{n} II_{z_i^{(t)}=j}\,(x_i)$$

(1.3) Generate $\mu_1^{(t)}$ from

$$N\left(\frac{\lambda\delta + 1x_1(z)}{\lambda + 1}, \frac{1}{\lambda + 1}\right)$$

(1.4) Generate (t) from

$$N\left(\frac{\lambda\delta + (n-1)x_2(z)}{\lambda + n - 1}, \frac{1}{\lambda + n - 1}\right)$$

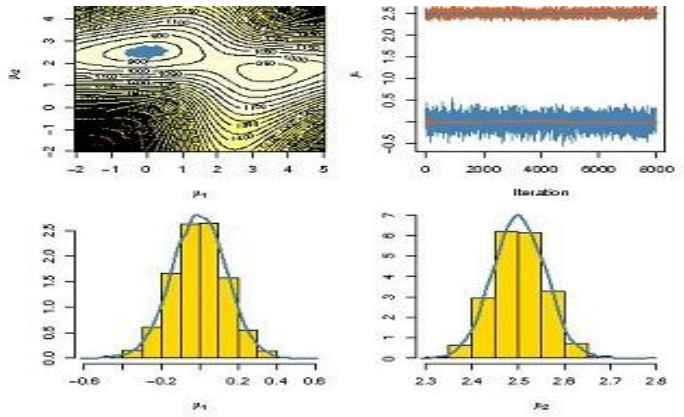


Fig. 1. Initialized at random.

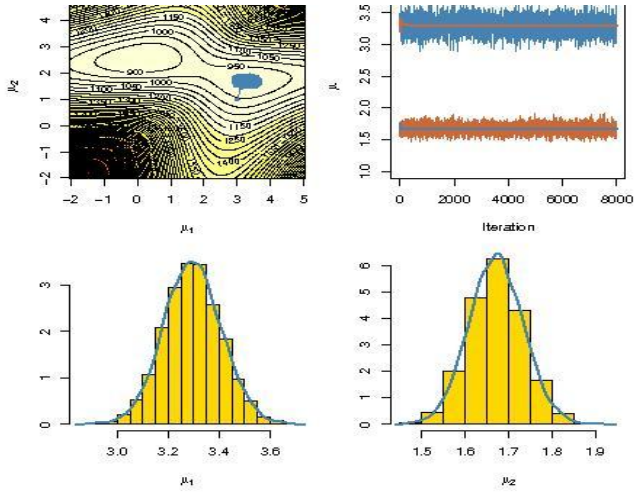


Fig. 2. Initialized close to the lower mode.

The following figures illustrate the behavior of this algorithm applying n = 500 samples when, u_1 = 0 and u_2= 2.5. Gibbs algorithm have been repeated 10000 times arid the produced models for (u_1/u_2) are plotted on the posterior logarithm balance graph. Figure 1 in which the algorithm has been converged to the real value of p, u_1 and u_2, shows the dot algorit/urru near the exponent, whereas time figure is the first point near the wrong exponent. These two figures clearly indicate that Gibbs algorithm is not able to get out of local exponents.

4. Conclusion

In this paper, the Gibbs algorithm has been applied in order to approximate Bayesian estimators in mixture models and its disadvantage due to being incapable of getting out of trap conditions was clarified. In order to solve this problem, two suggestions are presented: using other MCMC algorithms, and, applying artificial limitations.

References

1. Peter, D., Hoff, A., 2009. First Course in Bayesian Statistical Methods., Springer.
2. James, 0., 1995. Berger. Statistical Decision Theory and Bayesian Analysis Springer- Verilog., Second Edition.
3. Gholami, G.H., 2008. Change-point Problems in Regression: A Bayesian Approach., Ph. D. Thesis).
4. Christian, P., 2001. Robert., the Bayesian Choice. From Decision- Theoretic Foundations to Computational Implementation., Second Edition, Springer.
5. Christian, P., 2006. Robert. Jean-Michel Mann Bayesian Core: A Practical Approach to Computational Bayesian Statistics., Springer Texts in Statistics.
6. Jeffrey, H., 1961. Theory of Probability. Third Edition., Oxford University Press, Londomi.
7. Christian P., Robert. Kate Lee. Jean-Michel Marin. Kernie Mengersen Bayesian Inference on Mixtures of Distributions. J. Roy. Statist. Soc.
8. Christian, P., 2010. Robert. Bayesian Computational Methods., Springer.
9. A Reference for Developing a Basic Occupational Safety and Health Program for Small Businesses, April 2000. Occupational Safety & Health, State of Alaska.
10. Ganji, H., 2006. Work Psychology. savalan, tehran.

11. Hemmat joo, Y., 2004. Worker safety attitude and relationship with the events in a match factory in the city of Tabriz (Masters thesis). Faculty of Health. Tehran Univ.Med.Sci.
12. Krause, T.R., 1997. The Behavior-Based Safety process. Van Nostrand, New York.
13. Minoo, A.R., 2008. The relationship between demographic factors and attitudes to safety and safe behavior Saipa production staff. Presented at the The First International Conference on Health. Safety Env. Organizat., tehran.
14. Monazam, m., soltanzadh, a., 2007. Relationship between safety attitudes and events of a gas refinery. J. Res. Health Sci.
15. Rezaei, M., 2013. Attitudes of personnel safety and its impact on the health and safety conditions (Case Study: Khuzestan Farabi Agro Industry); Ninth Int. Conf. Manag.
16. Sheikhi, (soleiman)., 2008. Safety Attitudes operating room staff employed in hospitals of Qazv. Univ. Med. Sci. Qazvin Univ. Med. Sci., 29.
17. Williams, W., Purdy, S., Storey, L., Nakhla, M., Boon, G., 2007. Towards more effective methods for changing perceptions of noise in the workplace. Saf. Sci.