Applied Mathematics and Computation 244 (2014) 959–975

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Mixture models with an unknown number of components via a new posterior split–merge MCMC algorithm

Erlandson F. Saraiva [a], Francisco Louzada [b,*], Luis Milan [c]

[a] Instituto de Matemática, Universidade Federal de Mato Grosso do Sul, Brazil
[b] Departamento de Matemática Aplicada e Estatística, Universidade de São Paulo, Brazil
[c] Departamento de Estatística, Universidade Federal de São Carlos, Brazil

ARTICLE INFO

Keywords:
Mixture modeling
Bayesian inference
Metropolis–Hastings
Split–merge update

ABSTRACT

In this paper we introduce a Bayesian analysis for mixture models with an unknown number of components via a new posterior split–merge MCMC algorithm. Our strategy for splitting is based on data in which allocation probabilities are calculated based on posterior distribution from the previously allocated observations. This procedure is easy to be implemented and determines a quick split proposal. The acceptance probability for split–merge movements are calculated according to metropolised Carlin and Chib's procedure. The performance of the proposed algorithm is verified using artificial datasets as well as two real datasets. The first real data set is the benchmark galaxy data, while the second is the publicly available data set on *Escherichia coli* bacterium.

1. Introduction

Mixture models are widely used for modeling data where each observation is assumed to have arisen from one of the k subpopulations, where k can be known or unknown, and each subpopulation may be suitably modeled by a density from some parametric family. The density of each subpopulation is called the component of the mixture, which is weighted by the relative frequency of the subpopulation in the overall population. Furthermore, mixture models provide a convenient and flexible family of distributions for fitting data which are not well modeled by any standard parametric family, and can be seen as a parametric alternative to non-parametric methods of density estimation [25]. This kind of modeling finds a wide range of applications, ranging from client classification in market studies to grouping genes and proteins in bioinformatics. This paper is concerned with the estimation of the parameters of a mixture model framework, particularly when the number of components k is unknown.

For the case of unknown k, Carlin and Chib [3] and Chib [6] propose to estimate the marginal likelihood of a mixture model with k components and use the Bayes factor to test a model with k components *versus* a model with $k+1$ components. Mengersen and Robert [20] also use a testing approach considering the Kullback–Leibler divergence as a measure of distance between mixture models with k and $k+1$ components.

[*] Corresponding author. Addresses: Departamento de Matemática Aplicada e Estatística, Universidade de São Carlos, Brazil; Caixa Postal 676, CEP 13565-905, São Carlos, São Paulo, Brazil.

E-mail address: louzada@icmc.usp.br (F. Louzada).

http://dx.doi.org/10.1016/j.amc.2014.07.032
0096-3003/© 2014 Elsevier Inc. All rights reserved.

The advance of Markov chain Monte Carlo (MCMC) methods has revolutionized the Bayesian approach of finite mixture models. Some key papers on Bayesian analysis of finite mixtures using MCMC methods are due to Diebolt and Robert [10,11], Carlin and Chib [3], Escobar and West [13], Mengersen and Robert [20] and Roeder and Wasserman [23].

In the last years, sophisticated methods were proposed to jointly estimate the number of components and the component parameters of mixture models. Among they, Escobar and West [13] propose a Bayesian semi-parametric approach with the Dirichlet process prior; Richardson and Green [22] propose a Bayesian approach by considering the reversible jump algorithm with birth–death and split–merge moves. Dellapotas and Papgeorgiou [9] extend the reversible jump algorithm for the multivariate case constructing split–merge movements on the space of eigenvectors and eigenvalues of the current covariance matrix. The authors focus on predictive inference once it is invariant to the label switching problem. Stephens [25] proposes a birth-and-death process where each component is considered a point in the parameter space and constructs an ergodic Markov chain with an appropriate stationary distribution. Jain and Neal [17] propose a split–merge MCMC procedure for the conjugated Dirichlet process mixture model using a restricted Gibbs sampling scan to determine a split proposal, where the number of scans (tuning parameter) must be previously fixed by the user. Jain and Neal [18] extend their method to a nonconjugated Dirichlet process mixture model.

In this paper, we propose an alternative split–merge strategy to implement a MCMC methodology to jointly estimate the number of components k and the component parameters of a mixture model. The key idea for constructing the proposed posterior split–merge (PSM) MCMC algorithm is to design data-driven split–merge proposals. For this, we develop a split strategy, in which, allocation probabilities are calculated based on posterior distribution from observations previously allocated. Given the allocation of observations the candidate-values for component parameters are generated from posterior distributions since it is available in the close form.

Due to the nature of the parameter space, which is composite by the number of components, the latent allocation variables and the component parameters, when we propose a split or merge movement, we consider pseudo-priors distributions [3,16,8] as linking in order to maintain the reversibility condition and the detailed balance equation with respect to the posterior distribution. Thus, the acceptance probability for split–merge movements is given by the standard Metropolis–Hastings acceptance probability.

Once the posterior distribution for component parameters is available in a close form then, choosing the posterior density as candidate-generating density, the likelihood ratio and the prior densities ratio (from Metropolis–Hastings acceptance probability) cancel with the corresponding term in the posterior density, eliminating the parameters from the acceptance probability. Thus, the proposed algorithm performs a standard Metropolis–Hastings update with an acceptance probability which depends only on the data associated with component(s) selected for a split or a merge.

As advantages, our approach determines a quick split proposal, new components are created based on information from clusters of observations and new candidate-values for component parameters are generated from posterior distribution. The methodology also can be applied for mixture models from any parametric family, since the component parameters can be integrated out of the model analytically.

We illustrate the performance of the methodology on some artificial data sets and two real data sets. The first real dataset is the benchmark data on velocities from distant galaxies diverging from their own, previously analyzed by Roeder and Wasserman [23], Escobar and West [13], Richardson and Green [22] and Stephens [25], available in software R. The second is the *Escherichia Coli* bacterium gene expression data, described in [1] and extracted from site www.jbc.org. We also compare results from PSM with Reversible jump (RJ) proposed by Richardson and Green [22] and with proposal of Jain and Neal [17] which hereafter is denoted by JN. The performance of algorithms are presented in terms of posterior probability for number of components k, convergence of the posterior probability for k, mixing over k and estimated autocorrelation.

The paper is organized as follows. In Section 2, we describe the mixture model and the hierarchical Bayesian approach for the unknown k case. In Section 3, we present our PSM-MCMC approach. In Section 4, we verify the performance of the methodology using artificial and real datasets. In Section 5, the paper is concluded with final remarks on the proposed methodology.

2. Mixture model and Bayesian approach

Let $\mathbf{y} = \{y_1, \ldots, y_n\}$ be independent observations from a mixture model with k components,

$$f(y_i|\mathbf{w}, \boldsymbol{\phi}_k, k) = \sum_{j=1}^{k} w_j f(y_i|\phi_j), \tag{1}$$

where, $f(y_i|\phi_j)$ is the density of a family of parametric distributions with parameters ϕ_j (scalar or vector), $\boldsymbol{\phi}_k = (\phi_1, \ldots, \phi_k)$ is the whole vector of parameters and $\mathbf{w} = (w_1, \ldots, w_k), w_j > 0$ and $\sum_{j=1}^{k} w_j = 1$, are component weights.

As is usual in mixture models consider for each observation y_i a latent variable c_i, so that, $c_i = j$ if y_i is from component j [11]. The c_i's are supposed to be independently drawn from the distribution $P(c_i = j|\mathbf{w}, k) = w_j$, for $i = 1, \ldots, n$ and $j = 1, \ldots, k$. Letting n_j be the number of observations from component j (i.e. the number of c_i's equals to j), so the joint probability for $\mathbf{c} = (c_1, \ldots, c_n)$ given $\mathbf{w}$ and k is $\pi(\mathbf{c}|\mathbf{w}, k) = \prod_{j=1}^{k} w_j^{n_j}$. Conditional on $\mathbf{c} = (c_1, \ldots, c_n), \mathbf{y} = \{y_1, \ldots, y_n\}$ are independent observations from densities $f(y_i|c_i = j, \mathbf{w}, \boldsymbol{\phi}_k, k) = f(y_i|\phi_j)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

Considering S_j as being the set of observations belonging to component $j, S_j = \{y_i; c_i = j\}$, then