Computational Statistics and Data Analysis 54 (2010) 3324–3335

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

New approaches to compute Bayes factor in finite mixture models

M.J. Rufo *, J. Martín, C.J. Pérez

Departamento de Matemáticas, Universidad de Extremadura, Avda. de la Universidad s/n, 10071 Cáceres, Spain

ARTICLE INFO

Article history:
Received 14 November 2008
Received in revised form 3 May 2010
Accepted 4 May 2010
Available online 10 May 2010

Keywords:
Bayes factor
Cluster sampling
Conjugate prior distribution
Finite mixture model
Marginal distribution
Monte Carlo methods

ABSTRACT

Two new approaches to estimate Bayes factors in a finite mixture model context are proposed. Specifically, two algorithms to estimate them and their errors are derived by decomposing the resulting marginal densities. Then, through Bayes factor comparisons, the appropriate number of components for the mixture model is obtained. The approaches are based on simple theory (Monte Carlo methods and cluster sampling), what makes them appealing tools in this context. The performance of both algorithms is studied for different situations and the procedures are illustrated with some previously published data sets.

1. Introduction

The analysis of finite mixture models has received an increasing attention during the past several years. Mixture models are excellent tools that can be used to describe complex systems. An evidence of this, is their wide range of applications in many fields of knowledge, such as, for example, astronomy, economy, biology, medicine and engineering. See Böhning and Seidel (2003) and Böhning et al. (2007) for summaries on developments and advances in mixture models, and Lee et al. (2009) for the state-of-the-art Bayesian techniques for mixture estimations.

Mixture models can be interpreted in two different ways (see Titterington et al. (1985)). The first one, the direct interpretation, suggests that the underlying population comprises several subpopulations, that each observation belongs to one of these subpopulations, and that there is no knowledge of identity of this subpopulation. In other words, the mixture model is intended to be a direct representation of the underlying phenomenon. The second one, the indirect interpretation, suggests that the mixture model is simply used as an approximation of an unknown distribution in which the subpopulations have no physical interpretation. In a clustering problem, the goal is to identify groups of observations that are generated by the same mixture component (see, McLachlan and Basford (1998)).

One of the major interests in the Bayesian analysis of finite mixture models with an unknown number of components focuses on making inferences on the model parameters. This task can be carried out by using many different approaches. One of them is based on the reversible jump MCMC method, proposed by Green (1995) and applied to finite mixture models of normal distributions by Richardson and Green (1997), and more recently by Papastamoulis and Iliopoulos (2009). A general description of reversible jump MCMC for mixtures is provided by Marin and Robert (2007). Stephens (1997, 2000) presented an alternative proposal based on the construction of a birth-and-death process. More recently, Cappé et al. (2003) investigated the similarity between both approaches. These two approaches do not try to calculate the number of components in the mixture model, but directly to make inferences on the parameters of interest by considering the number

* Corresponding address: Departamento de Matemáticas, Escuela Politécnica, Universidad de Extremadura, Avda. de la Universidad s/n, 10071 Cáceres, Spain. Fax: +34 927257203.

E-mail addresses: mrufo@unex.es (M.J. Rufo), jrmartin@unex.es (J. Martín), carper@unex.es (C.J. Pérez).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.05.002

of components as a random variable. Different approaches can be derived when the interest is focused on the choice of the number of components besides the subsequent inferences. This can be considered as a particular case of the more general model choice problem. One of these approaches is based on the Kullback–Leibler divergence (see Mengersen and Robert (1996) and Sahu and Chen (2003)). Another one is based on the Bayes factor (see, for example, Gelfand and Dey (1994), Ishawaran et al. (2001), Song and Lee (2002), Berkhof et al. (2003) and Congdon (2006)). Some other approach describes how to estimate the marginal likelihoods (see, among others, Marin and Robert (2008), Frühwirth-Schnatter (2004), Chib and Jeliazkov (2005) and references therein). A hybrid approach is based on an extension of the bridge sampling approach of Men and Wong (1996). It uses the output of the reversible jump algorithm to estimate the Bayes factors (see Bartolucci et al. (2006) for a detailed description).

In this paper, a Bayesian mixture model selection by using Bayes factor-based approaches is addressed. In general, the Bayes factor cannot be analytically obtained and, therefore, it has to be estimated. Many estimation techniques use MCMC methods with the corresponding chain convergence assessment for the posterior distribution. Besides, assessing the accuracy of the marginal distributions is not trivial with MCMC methods because of the dependence between successive samples. In order to avoid these problems, two sampling-based approaches to estimate the marginal distributions in the Bayes factor are proposed.

In this context, improper prior distributions cannot be used, since they may lead to an improper posterior distribution and, consequently, the Bayes factor is not defined. In order to solve this problem, Song and Lee (2002) and Berkhof et al. (2003), among others, considered examples with conjugate prior distributions. Here independent conjugate prior distributions are used for the parameter components and an expression involving the marginal distribution for each component is derived. This fact allows to estimate the Bayes factor by using Monte Carlo methods. In the first approach, latent allocation variables are generated in two steps. The corresponding Monte Carlo estimate provides precise estimations when the number of observations is moderate. When the number of observations is large, a second approach based on a two-stage cluster sampling is presented.

The outline of the paper is as follows. Section 2 presents the basic concepts, including a suitable decomposition of the marginal distribution. In Section 3, the two approaches are developed by presenting the Bayes factor estimate and its error. Section 4 shows some illustrative examples. The conclusions are presented in Section 5. Finally, an Appendix presents some theoretical results.

2. Background

This section presents some basic concepts that will be used throughout the paper.

2.1. Finite mixture models

Data $x_1, \ldots, x_n$ are assumed to be independent observations from a mixture density with k components given by:

$$p\left(x|\boldsymbol{\omega}, \boldsymbol{\theta}\right) = \sum_{j=1}^{k} \omega_j p\left(x|\theta_j\right), \tag{1}$$

where $\omega_j, j = 1, \ldots, k$, are the mixture weights (which are restricted to be non-negative and sum to unity) and $p(\cdot|\theta)$ is a parametric family where θ_j is the parameter for component j.

This model is augmented by introducing latent variables. For each observation $x_i, i = 1, \ldots, n$, a latent variable Z_i is introduced such that $Z_i = j$ indicates that the observation x_i belongs to the j-th component of the mixture. Latent random variables are independent and identically distributed with mass function given by:

$$p\left(Z_i = j|\boldsymbol{\omega}, \boldsymbol{\theta}\right) = \omega_j,$$

and verifying $p(x_i|Z_i = j, \boldsymbol{\theta}, \boldsymbol{\omega}) = p(x_i|\theta_j)$.

The introduction of these latent random variables makes the generating process easier in the Bayesian framework. Integrating out with respect to $Z_1, \ldots, Z_n$, the expression (1) is obtained, i.e.:

$$p\left(x_i|\boldsymbol{\omega}, \boldsymbol{\theta}\right) = \sum_{j=1}^{k} p\left(Z_i = j|\boldsymbol{\omega}, \boldsymbol{\theta}\right) p\left(x_i|Z_i = j, \boldsymbol{\omega}, \boldsymbol{\theta}\right) = \sum_{j=1}^{k} \omega_j p\left(x_i|\theta_j\right).$$

For a more detailed review on mixture models, see, among others, McLachlan and Peel (2000) and Frühwirth-Schnatter (2006).

2.2. Bayes factor

Let M_h and M_i be two competing models with h and i components, respectively. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be the data vector. Therefore, M_h is composed of a mixture model with h component, $p(x|\boldsymbol{\omega}, \boldsymbol{\theta}) = \sum_{j=1}^{h} \omega_j p(x|\theta_j)$, and the prior distribution for the parameters $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_h)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_h)$. An analogous interpretation is made for the model M_i.