Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

The doubly smoothed maximum likelihood estimation for location-shifted semiparametric mixtures

Byungtae Seo

Department of Statistics, Sungkyunkwan University, Seoul 03063, Republic of Korea

ARTICLE INFO

Article history:
Received 20 July 2015
Received in revised form 30 May 2016
Accepted 2 November 2016
Available online xxxx

Keywords:
Finite mixture
Semiparametric mixture
Doubly smoothed MLE

ABSTRACT

Finite mixture of a location family of distributions are known to be identifiable if the component distributions are common and symmetric. In such cases, several methods have been proposed for estimating both the symmetric component distribution and the model parameters. In this paper, we propose a new estimation method using the doubly smoothed maximum likelihood, which can effectively eliminate potential biases while maintaining a high efficiency. Some numerical examples are presented to demonstrate the performance of the proposed method.

1. Introduction

Finite mixture models have been widely applied as a model based clustering tool for classifying or clustering data in various scientific fields. In order to employ mixture models, practitioners must first determine a parametric family for the component distribution. The normal distribution is the most commonly used, but other parametric distributions can also be employed, depending on the source of the data or the area of study. For example, in survival analysis, exponential and gamma distributions are common choices, owing to the nature of the data involved.

After determining the component distribution, the next task is to specify a suitable number of components. This is related to the number of hidden homogeneous subpopulations of the whole population under study. There are some well-known criteria for choosing the number of components. The most common tools employed for this purpose are information-based criteria, such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and their relatives. Such criteria can be used to determine a suitable number of components when a certain class of component distributions is assumed.

Although the procedure described above is standard for fitting finite mixture models, a more desirable method would be to estimate the component distribution itself as well as model parameters rather than adopting an assumption. However, we should exercise a high degree of care in addressing this issue. Because mixture distributions are already flexible enough to cover almost all statistical distributions, to allow further flexibility could result in fundamental statistical problems. For example, if one leaves the component density completely unspecified in a finite mixture model, then the model is not identifiable in general.

Hall and Zhou (2003) tackled this issue in a multivariate mixture in which coordinate distributions are conditionally independent given the component membership. Under some mild conditions, they showed the identifiability of the model for more than two dimensional mixtures. Hohmann and Holzmann (2013) showed the identifiability and proposed an estimation method for bivariate mixtures under additional tail conditions on the component characteristic functions and densities. Recently, Chauveau et al. (2015) provided a comprehensive survey for such semiparametric multivariate mixtures.

E-mail address: seobt@skku.edu.

http://dx.doi.org/10.1016/j.csda.2016.11.003
0167-9473/

For the identifiability issue, the multivariate mixtures with the conditional independence can make the model identifiable. However, the identifiability of the univariate mixtures needs an additional condition. One feasible condition is that the component distribution is common and symmetric for each component. Hunter et al. (2007) and Bordes et al. (2006) studied such a model and demonstrated that this semiparametric mixture model is identifiable with a nonparametric symmetric component distribution when the number of component is two. Hunter et al. (2007) also discussed the identifiability for the mixtures with more than two components.

Despite the theoretical justification for the use of semiparametric mixtures, it is not easy to carry out the required estimation owing to the nonparametric component distribution. In order to overcome this challenge, Hunter et al. (2007) suggested a minimum distance approach. However, their method does not produce a proper estimate of the component distribution. The estimating methods proposed by Bordes et al. (2006) and Hunter et al. (2007) are not efficient in general and hard to generalize to more than two-component mixtures. Bordes et al. (2007) proposed a stochastic EM-like algorithm which can be generalized for more than two-component semiparametric mixtures, but it does not have the descent property. Recently, Chauveau et al. (2015) modified the algorithm in order to have a descent property. Although these estimation methods can produce a reasonable estimator, they are not fully likelihood-based and the choice of bandwidth remains a difficult task.

The focus of this paper is to develop a likelihood-based estimation method for semiparametric univariate mixtures. In this regard, Chee and Wang (2013) proposed nonparametric mixture models for the nonparametric component density, and applied the constrained Newton method for multiple support points (CNM) to estimate the nonparametric mixing distribution. Their method provides a well-defined likelihood along with an algorithm estimating the nonparametric mixing distribution and model parameters. However, because their method still requires some extent of smoothing, a loss in efficiency is inevitable. Furthermore, although their method is less sensitive to the choice of the bandwidth than other kernel-based methods, the choice of bandwidth remains a nontrivial issue.

When a given problem requires some degree of smoothing, the use of kernel almost always produces biases, and leads to some loss in efficiency. This results from the fact that smoothing involves contamination or distortion of the initial problem, which leads to a certain deviance from the original model or data. In order to minimize this deviance, we propose the use of the doubly smoothed maximum likelihood estimator (DSMLE), as introduced by Seo and Lindsay (2013). The DSMLE was originally developed with the aim of constructing a consistent estimator when the usual maximum likelihood estimator (MLE) fails to be consistent. In addition, the DSMLE can also be useful in reducing biases or information loss when a given problem requires a certain degree of smoothing.

In this paper, we develop an estimation method for semiparametric mixture models based on a double smoothing procedure. The remainder of this paper is organized as follows. In Section 2, we introduce some recently proposed computing methods for the semiparametric mixtures. In Section 3, we present the method for applying the double smoothing procedure, along with a brief introduction to the DSMLE. Some computational issues are discussed in Section 4, and numerical examples are presented in Section 5. We then make some concluding remarks in Section 6.

2. Semiparametric mixture models

In this section, we briefly introduce two existing methods for estimating model parameters in semiparametric mixture models. In this context, let us consider the following m-component mixture density with an unspecified symmetric density g:

$$f(x;\boldsymbol{\pi},\boldsymbol{\mu},g)=\sum_{j=1}^{m}\pi_j g(x-\mu_j), \tag{2.1}$$

where $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_m)$ is a vector containing each subpopulation mean and $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_m)$ is the corresponding vector of mixing proportions, which satisfies $\sum_{j=1}^{m}\pi_j=1$ and $\pi_j>0, j=1,\ldots,m$.

To estimate $(\boldsymbol{\pi},\boldsymbol{\mu},G)$, Bordes et al. (2007) proposed a stochastic expectation–maximization (EM) algorithm. Although Bordes et al. (2007) did not provide a clear probability model for g, in some sense, their estimating algorithm is equivalent to estimating $(\boldsymbol{\pi},\boldsymbol{\mu})$ and B under

$$g_K(x;B)=\sum_{i=1}^{n}\frac{K_h(x-\tilde{x}_i)+K_h(x+\tilde{x}_i)}{2n}=\int\frac{K_h(x-\tilde{x})+K_h(x+\tilde{x})}{2}dB(\tilde{x}), \tag{2.2}$$

where K_h is a symmetric kernel density with bandwidth h, and B is the uniform distribution on $\tilde{x}_1,\ldots,\tilde{x}_n$. Using the stochastic EM algorithm, they then iteratively update $(\boldsymbol{\pi},\boldsymbol{\mu})$ and $\tilde{x}_i$'s.

Chee and Wang (2013) modeled g using a similar but more flexible model as

$$g_M(x;Q)=\int\frac{K_h(x-\eta)+K_h(x+\eta)}{2}dQ(\eta), \tag{2.3}$$

where Q is an unspecified probability distribution on $\mathbb{R}^+$. Readers should note that $g_M(x;Q)$ can be interpreted as a nonparametric mixture density with a component density $(K_h(x-\eta)+K_h(x+\eta))/2$ and an unknown mixing distribution

Please cite this article in press as: Seo, B., The doubly smoothed maximum likelihood estimation for location-shifted semiparametric mixtures. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.11.003