Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5429–5443
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A stochastic EM algorithm for a semiparametric mixture model

Laurent Bordes[a], Didier Chauveau[b,*], Pierre Vandekerkhove[c]

[a] *Université de Technologie de Compiègne, France*
[b] *MAPMO, Fédération Denis Poisson, Université d'Orléans & CNRS UMR 6628, BP 6759, 45067 Orléans cedex 2, France*
[c] *Université de Marne-la-Vallée & CNRS UMR 8050, France*

Available online 31 August 2006

Abstract

Recently, there has been a considerable interest in finite mixture models with semi-/non-parametric component distributions. Identifiability of such model parameters is generally not obvious, and when it occurs, inference methods are rather specific to the mixture model under consideration. Hence, a generalization of the EM algorithm to semiparametric mixture models is proposed. The approach is methodological and can be applied to a wide class of semiparametric mixture models. The behavior of the proposed EM type estimators is studied numerically not only through several Monte-Carlo experiments but also through comparison with alternative methods existing in the literature. In addition to these numerical experiments, applications to real data are provided, showing that the estimation method behaves well, that it is fast and easy to be implemented.

Keywords: EM algorithm; Finite mixture model; Semiparametric model; Stochastic EM

1. Introduction

Probability density functions (pdf) of m-component mixture models are defined in a general setup by

$$g(x) = \sum_{j=1}^{m} \lambda_j \, f_j(x), \quad \sum_{j=1}^{m} \lambda_j = 1, \quad x \in \mathbb{R}^p,$$

where the unknown mixture proportions $\lambda_j \geqslant 0$ and the unknown pdf's f_j have to be estimated. It is commonly assumed that the f_j's belong to a parametric family $\mathscr{F} = \{f(\cdot|\xi), \xi \in \mathbb{R}^d\}$ indexed by an Euclidean parameter ξ, so that the pdf g becomes

$$g_\theta(x) = g(x|\theta) = \sum_{j=1}^{m} \lambda_j \, f(x|\xi_j), \tag{1}$$

where $\theta = (\lambda_j, \xi_j)_{j=1,\ldots,m}$ is the Euclidean model parameter. When the number of components m is fixed the parametric mixture model of Eq. (1) has been well-studied; e.g., Titterington et al. (1985), Lindsay (1995) and McLachlan and Peel (2000) are general references to the broad literature on this topic.

* Corresponding author. Tel.: +33 1 238417009; fax: +33 1 238417205.
E-mail address: didier.chauveau@univ-orleans.fr (D. Chauveau).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.08.015

Nonparametric approaches for mixtures are motivated by the fact that the choice of a parametric family $\mathscr{F}$ may be difficult. Indeed, histograms built from actual data may reveal a mixture structure (e.g., bumps) without giving evidence of an appropriate parametric family (particularly for heavy-tailed distributions, see, e.g., Hunter et al., 2006). In that case, the semiparametric approach may even be viewed as an exploratory data analysis tool, in an attempt to validate a parametric assumption, prior to the application of a possibly more efficient parametric estimation procedure.

However, model (1) can be made more flexible assuming that the number of components m is unknown; in that case m has to be estimated, see e.g., Leroux (1992), Dacunha-Castelle and Gassiat (1999) and Lemdani and Pons (1999). But if the number of components is specified but little is known about subpopulations (e.g., tails), another way to make the model more flexible is to avoid parametric assumption on $\mathscr{F}$. For example, one may state the model where for $j = 1, \ldots, m$ we have $f_j \in \mathscr{F} = \{\text{continuous pdf on } \mathbb{R}^p\}$. Of course, such a model is very flexible since each component distribution can itself be a mixture distribution, and obviously, without additional assumptions on $\mathscr{F}$ the resulting model parameters are not identifiable. Nevertheless, if training data are available such models become identifiable, and then, the component distributions can be estimated nonparametrically, see for example Hall (1981) and Titterington (1983).

Note also that in the nonparametric setup without training data, specific methods to estimate mixture weights have been developed by Hettmansperger and Thomas (2000) and Cruz-Medina and Hettmansperger (2004).

Recently, Hall and Zhou (2003) looked at p-variate data drawn from a mixture of two distributions, each having independent nonparametric components, and proved that under mild regularity assumptions their model is identifiable for $p \geqslant 3$. The non-identifiability for $p \leqslant 2$ requires to restrain the class of pdf $\mathscr{F}$. For example, for $p = 1$, restraining $\mathscr{F}$ to the location-shifted symmetric pdf, we obtain the following semiparametric mixture model:

$$g_\varphi(x) = g(x|\varphi) = \sum_{j=1}^{m} \lambda_j \, f(x - \mu_j), \quad x \in \mathbb{R}, \tag{2}$$

where the λ_j's, the μ_j's and $f \in \mathscr{G} = \{\text{even pdf on } \mathbb{R}\}$ are unknown. Hence the model parameter is

$$\varphi = (\theta, f) = ((\lambda_j, \mu_j)_{j=1,\ldots,m}, f) \in \Phi = \Theta \times \mathscr{G},$$

where

$$\Theta = \left\{ (\lambda_j, \mu_j)_{j=1,\ldots,m} \in \{(0,1) \times \mathbb{R}\}^m; \sum_{j=1}^{m} \lambda_j = 1, \mu_i \neq \mu_j, 1 \leqslant i < j \leqslant m \right\}.$$

See Bordes et al. (2005) and Hunter et al. (2006) for identifiability results. In Bordes et al. (2005), for $m = 2$, the authors propose an estimator of (θ, f) for $m = 2$. Because $g = A_\theta f$, where A_θ is an invertible operator from $L^1(\mathbb{R})$ to $L^1(\mathbb{R})$, and f is an even pdf, they propose a contrast function for θ that depends only on g. Given a sample of independent g-distributed random variables they estimate g. Then, replacing g by its estimator in the contrast function, they propose a minimum contrast estimator for θ, and then, inverting A_θ and replacing θ by its estimator they obtain an estimator of the pdf f (which generally is not a pdf because the estimator of g has no reason to be in the range of the operator A_θ). This method has several limitations. For example, for $m = 3$, even if the model is identifiable (see Hunter et al., 2006) the operator A_θ may not be invertible and then the estimation method fails. On the other hand, the method cannot be naturally generalized to p-variate data. Furthermore, the numerical computation involved by the method is time consuming which can be a drawback for a large sample size. In Hunter et al. (2006) an alternative method of estimation is proposed but it seems that it suffers from similar weakness.

In parametric setup one main problem is the computation of maximum likelihood (ML) estimates; parameter estimates cannot in general be obtained in closed form from mixture structures. Conventional algorithms, such as the Newton–Raphson, have long been known to lead to difficulties; see Lindsay (1995, p. 65). The computational issue has largely been resolved, however, with the development of the EM algorithm after its formalization by Dempster et al. (1977). See McLachlan and Krishnan (1997) for a detailed account of the EM algorithm. Moreover, in the parametric setup, the ML method can be applied easily, which is no longer true in the semiparametric setup. This is another