Computational Statistics and Data Analysis 103 (2016) 1–16

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Nonparametric mixture models with conditionally independent multivariate component densities

Didier Chauveau *, Vy Thuy Lynh Hoang

Univ. Orléans, CNRS, MAPMO, UMR 7349, Orléans, France

ARTICLE INFO

Article history:
Received 18 January 2015
Received in revised form 26 April 2016
Accepted 28 April 2016
Available online 6 May 2016

Keywords:
EM algorithm
Multivariate kernel density estimation
Multivariate mixture
Nonparametric mixture

ABSTRACT

Models and algorithms for nonparametric estimation of finite multivariate mixtures have been recently proposed, where it is usually assumed that coordinates are independent conditional on the subpopulation from which each observation is drawn. Hence in these models the dependence structure comes only from the mixture. This assumption is relaxed, allowing for independent multivariate *blocks* of coordinates, conditional on the subpopulation from which each observation is drawn. Otherwise the density functions of these blocks are completely multivariate and nonparametric. An EM-like algorithm for this model is proposed, and some strategies for selecting the bandwidth matrix involved in the nonparametric estimation step of it are derived. The performance of this algorithm is evaluated through several numerical simulations. A real dataset of reasonably large dimension is experimented on this new model and algorithm to illustrate its potential from the model based, unsupervised clustering perspective.

1. Introduction

Populations of individuals may often be divided into subgroups. Examining a sample of measurements to discern and describe subgroups of individuals, even when there is no observable variable that readily indicates into which subgroup an individual properly belongs, is sometimes referred to as "unsupervised clustering" in the literature, and in fact mixture models may be generally thought of as comprising the subset of clustering methods known as model-based clustering.

Finite mixture models may also be used in situations beyond those for which clustering of individuals is of interest. For one thing, finite mixture models give descriptions of entire subgroups (called *components*), rather than assignments of individuals to those subgroups. Indeed, even the subgroups may not necessarily be of interest; sometimes finite mixture models merely provide a means for adequately describing a particular distribution, such as the distribution of residuals in a linear regression model where outliers are present. Much of the theory of these models involves the assumption that the subgroups are distributed according to a particular parametric shape and quite often this parametric family is univariate or multivariate normal.

The most general model for nonparametric multivariate mixtures is as follows: suppose the vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are a simple random sample from a finite mixture of $m > 1$ arbitrary distributions. The density of each $\boldsymbol{X}_i$ may be written

$$g_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \sum_{j=1}^{m} \lambda_j f_j(\boldsymbol{x}_i), \tag{1}$$

* Correspondence to: Univ. Orléans, CNRS, MAPMO, UMR 7349, BP 6759, 45067 Orléans cedex 2, France. Tel.: +33 2 38 49 46 81; fax: +33 2 38 41 72 05.
E-mail addresses: didier.chauveau@univ-orleans.fr (D. Chauveau), vy-thuy-lynh.hoang@etu.univ-orleans.fr (V.T.L. Hoang).

http://dx.doi.org/10.1016/j.csda.2016.04.013
0167-9473/

where $\boldsymbol{x}_i \in \mathbb{R}^r$, and $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{f}) = (\lambda_1, \ldots, \lambda_m, f_1, \ldots, f_m)$ denotes the parameters of the statistical model. In this model λ_j denotes the proportion (weight) of component j in the population; the λ_j's are thus positive and $\sum_{j=1}^m \lambda_j = 1$. The f_j's are the component densities, drawn from some family of multivariate density functions $\mathcal{F}$ absolutely continuous with respect to Lebesgue measure. Note that the univariate ($r = 1$) case will only be briefly considered, since this paper focuses on multivariate extensions.

Model (1) is not identifiable if no restrictions are placed on $\mathcal{F}$, where "identifiable" means that $g_{\boldsymbol{\theta}}$ has a *unique* representation of the form (1) and also that we do not consider that "label-switching" – i.e., reordering the m pairs $(\lambda_1, f_1), \ldots, (\lambda_m, f_m)$ – produces a distinct representation. The most common restriction in the mixture literature is to assume that the family $\mathcal{F}$ is *parametric*, i.e. that any $f \in \mathcal{F}$ is completely specified by a finite-dimensional parameter. The most used and studied parametric mixture model is the Gaussian mixture, where f_j is the density of a (univariate or multidimensional) Gaussian distribution with mean μ_j and variance (matrix) Σ_j. Section 1.2 presents various ways of relaxing this parametric assumption while preserving an identifiability property. In the recent literature, finite mixtures of non-normal distributions have been considered as alternatives to the traditional Gaussian mixture, see, e.g., Lee and McLachlan (2013) which provides a comprehensive overview. These non-normal mixtures are mostly proposed to model heavy-tailed or skewed normal distributions, but are not appropriate for, e.g., non-elliptical clusters (see the model we consider in Section 4.4 for an example). Another point is that each non-normal but parametric mixture requires a specific multivariate EM algorithm, whereas our approach follows a different track: it is fully general since it allows for modeling of any cluster shape, and only requires for clustering the algorithm we propose.

1.1. The EM algorithm

Mixture models are deeply connected to the EM algorithm. This algorithm, as defined in the seminal article (Dempster et al., 1977), is more properly understood to be a class of algorithms, a number of which predate even (Dempster et al., 1977) in the literature. These algorithms are designed for maximum likelihood estimation in missing data problems, of which finite mixtures are canonical examples because the unobserved labels of the individuals (as in unsupervised clustering) give an easy interpretation of missing data. A recent account of the EM algorithm principle, properties and generalizations can be found in McLachlan and Krishnan (2008), and mixture models are deeply detailed in McLachlan and Peel (2000).

In a missing data setup, the n-fold product of the probability density function (pdf) $g_{\boldsymbol{\theta}}$ of the observations corresponds to the *incomplete* data pdf, associated with the log-likelihood $\ell_{\boldsymbol{x}}(\boldsymbol{\theta}) = \sum_{i=1}^n \log g_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$. In mixture models and many other missing data situations, maximizing $\ell_{\boldsymbol{x}}(\boldsymbol{\theta})$ leads to a difficult problem. Intuitively, EM algorithms replace this unfeasible maximization by the maximization of a pseudo-likelihood that resembles the likelihood for some complete data $\boldsymbol{y}$ that is defined from the model, so that this pseudo-likelihood is easy to maximize. Assuming $\boldsymbol{y}$ comes from a complete data pdf $g^c_{\boldsymbol{\theta}}$, the EM algorithm iteratively maximizes the operator

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) := \mathbb{E}[\log g^c_{\boldsymbol{\theta}}(\boldsymbol{y})|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}],$$

the expectation being taken relatively to the conditional distribution of $(\boldsymbol{y}|\boldsymbol{x})$, for the value $\boldsymbol{\theta}^{(t)}$ of the parameter at iteration t. Given an arbitrary starting value $\boldsymbol{\theta}^{(0)}$, the EM algorithm generates a sequence $(\boldsymbol{\theta}^{(t)})_{t\geq 1}$ by iterating the following steps:

1. E-step: `compute` $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$
2. M-step: `set` $\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

In finite mixture models, the *complete data* associated with the actually observed sample $\boldsymbol{x}$ is $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{Z})$, where to each individual (multivariate) observation $\boldsymbol{x}_i$ is associated an indicator variable Z_i denoting its component of origin. It is common to define $Z_i = (Z_{i1}, \ldots, Z_{im})$ with the indicator variables

$$Z_{ij} = \mathbb{I}\{\text{observation } i \text{ comes from component } j\}, \quad \sum_{j=1}^m Z_{ij} = 1.$$

From (1), this means that $\mathbb{P}_{\boldsymbol{\theta}}(Z_{ij} = 1) = \lambda_j$, and $(\boldsymbol{X}_i|Z_{ij} = 1) \sim f_j$, $j = 1, \ldots, m$. In this case, the expectation is w.r.t. the conditional distribution of the Z_{ij}'s,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) := \mathbb{E}\Big[\sum_{i=1}^n \sum_{j=1}^m Z_{ij} \log \lambda_j f_j(\boldsymbol{x}_i)|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}\Big].$$

Conveniently, the M-step for finite mixture models always looks partly the same: No matter what form the f_j's take, the updates to the mixing proportions are given by

$$\lambda_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^n p_{ij}^{(t)}, \quad \text{for } j = 1, \ldots, m,$$

where $p_{ij}^{(t)} := \mathbb{P}_{\boldsymbol{\theta}^{(t)}}(Z_{ij} = 1|\boldsymbol{x}_i)$ is the *posterior probability* that the individual i comes from component j. The updates for the f_j's depend on the particular form of the component densities. In parametric mixtures (i.e. when the family $\mathcal{F}$ is completely specified by a finite-dimensional parameter), the updates of these parameters are often straightforward, and can be looked like weighted MLE estimates. This is the case for, e.g., Gaussian mixtures.