Computational Statistics and Data Analysis 54 (2010) 2253–2266

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Mixtures of regressions with predictor-dependent mixing proportions

D.S. Young *, D.R. Hunter

The Pennsylvania State University, Department of Statistics, University Park, PA, 16802, United States

ARTICLE INFO

Article history:
Received 29 January 2009
Received in revised form 17 July 2009
Accepted 5 April 2010
Available online 22 April 2010

Keywords:
EM algorithms
Hierarchical mixture of experts
Mixture models
Mixtures of regressions

ABSTRACT

We extend the standard mixture of linear regressions model by allowing the mixing proportions to be modeled nonparametrically as a function of the predictors. This framework allows for more flexibility in the modeling of the mixing proportions than the fully parametric mixture of experts model, which we also discuss. We present an EM-like algorithm for estimation of the new model. We also provide simulations demonstrating that our nonparametric approach can provide a better fit than the parametric approach in some instances and can serve to validate and thus reinforce the parametric approach in others. We also analyze and interpret two real data sets using the new method.

1. Introduction

In a typical multivariate finite mixture model, the k-dimensional vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are a simple random sample from an m-component mixture distribution such that $\mathbf{Y}_i$ has density

$$f(\mathbf{y}_i; \boldsymbol{\psi}) = \sum_{j=1}^{m} \lambda_j g(\mathbf{y}_i; \boldsymbol{\theta}_j), \tag{1}$$

where $m > 1$ is fixed (and assumed known for now) and the λ_j, called the weights (or *mixing proportions*) for the components, are positive and sum to unity. The density g is assumed to come from a parametric family with parameter $\boldsymbol{\theta}_j \in \Theta_j \subseteq \mathbb{R}^q$, where Θ_j is open in $\mathbb{R}^q$. The mixture density f is parameterized by $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ represents the parameter space for all unknown parameters in the mixture model, i.e., $(\lambda_1, \ldots, \lambda_m, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$.

Suppose now a vector of predictors, say $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})^{\mathrm{T}}$ for $p < n$, is also observed with each response $\mathbf{Y}_i$. The goal is to describe the conditional distribution of $\mathbf{Y}_i|\mathbf{X}_i$ using a mixture of linear regressions with assumed Gaussian errors. In this article, we will restrict attention to a univariate response variable Y; this is common in the mixture-of-regressions literature, and the univariate case is sufficient to elucidate the ideas we present. Thus, Eq. (1) becomes

$$f(y_i; \mathbf{x}_i, \boldsymbol{\psi}) = \sum_{j=1}^{m} \lambda_j \phi(y_i; \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j, \sigma_j^2), \tag{2}$$

where $\phi(\cdot; \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j, \sigma_j^2)$ is the univariate normal probability density function with mean $\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j$ and variance σ_j^2 for some $(\boldsymbol{\beta}_j, \sigma_j^2) \in \mathbb{R}^p \times \mathbb{R}_*^+$.

The uses of mixtures of regressions fall into two primary categories. The first involves estimating a set of regression coefficients for all observations coming from a possibly unknown number of heterogeneous classes. This scenario arises

* Corresponding author.
E-mail addresses: dsy109@stat.psu.edu (D.S. Young), dhunter@stat.psu.edu (D.R. Hunter).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.04.002

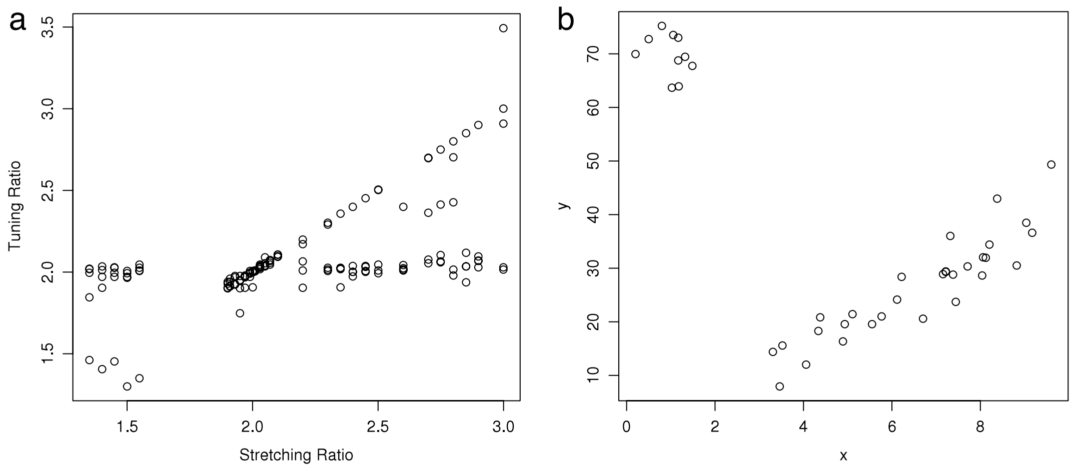

Fig. 1. Plot (a) is a scatterplot of a data set from a tone perception study reported in Cohen (unpublished). Plot (b) shows simulated data illustrating masked outliers.

when it seems inaccurate to assume that a single regression adequately explains the relationship between the variables at hand. An example of this scenario is depicted in Fig. 1(a), which shows data from an experiment on the perception of musical tones (Cohen, unpublished). This usage of mixtures of regressions has been extensively studied in the econometrics literature and was first introduced by Quandt (1972) as the *switching regimes*, or *switching regressions*, problem.

A second use for mixtures of regressions is in outlier detection or robust regression estimation. For example, one regression plane may adequately model the data, but there is an apparent class heterogeneity because of large variances attributed to some observations, which are considered outliers. Another example, as considered by Viele and Tong (2002), is when the outliers appear in clusters. Such outliers are said to be *masked*, and Viele and Tong (2002) further describe them as outliers that "cannot be detected individually by standard techniques". Pena et al. (2003) used a split and recombine (SAR) procedure to identify possible clusters in a sample, which can be extended to identifying masked outliers. Fig. 1(b) is a simulated data set with an extreme case of masked outliers, similar to those analyzed in Viele and Tong (2002) and Pena et al. (2003). The cluster of points in the upper left of the scatterplot are high leverage outliers which traditional methods fail to detect (see Justel and Pena, 1996).

Given the m-component mixture of regressions model (2) and a new observation $(y_{n+1}, \mathbf{x}_{n+1})$, one might ask which of the m regression functions in the model should be used to predict the value of the response y_{n+1}. According to the model, each regression should occur with probability equal to its corresponding λ_j, but this might not be realistic if $\mathbf{x}_{n+1}$ contains some information about the relative weights. To reflect this possibility in the notation, we may replace model (2) by

$$f(y_i; \mathbf{x}_i, \boldsymbol{\psi}) = \sum_{j=1}^{m} \lambda_j(\mathbf{x}_i)\phi(y_i; \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j, \sigma_j^2). \tag{3}$$

How should we model $\lambda_j(\mathbf{x}_i)$ in Eq. (3)? One way is to assume a parametric form, such as a logistic function, which introduces new parameters requiring estimation. This is the idea of the *hierarchical mixtures of experts* (HME) procedure (Jacobs et al., 1991), which is commonly used in neural networks. Yet just as nonparametric regression techniques are sometimes preferred over parametric regression – for instance, in situations where prediction of new observations is more important than explanation or where one wishes to verify a putative parametric form or even discover a new one without assuming one *a priori* – one might reasonably wish to determine $\lambda_j(\mathbf{x}_i)$ nonparametrically. Our method allows such a determination.

In the remainder of this article, we discuss the HME model, then introduce our alternative nonparametric method for modeling the mixing weights as functions of the predictors. We also provide an EM-like algorithm for estimation using our method. Finally, we illustrate the method using both simulated data sets and real data sets. The code for the estimation procedures is included in the package `mixtools` (Young et al., 2008) for the R statistical computing environment (R Development Core Team, 2008).

2. Hierarchical mixtures of experts

The hierarchical mixtures of experts (HME) model comes from the statistical learning, or machine learning, literature (see Jordan and Jacobs (1992), Jordan and Jacobs (1994) and Jordan and Xu (1995) for discussion). Finite mixture models, like the ones in this article, may be considered an unsupervised learning method in the sense that there are