Computational Statistics and Data Analysis 71 (2014) 116–127

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Robust mixture regression using the t-distribution

Weixin Yao *, Yan Wei, Chun Yu

Kansas State University, USA

ARTICLE INFO

Article history:
Received 26 June 2012
Received in revised form 11 July 2013
Accepted 12 July 2013
Available online 1 August 2013

Keywords:
EM algorithm
Mixture regression models
Outliers
Robust regression
t-distribution

ABSTRACT

The traditional estimation of mixture regression models is based on the normal assumption of component errors and thus is sensitive to outliers or heavy-tailed errors. A robust mixture regression model based on the t-distribution by extending the mixture of t-distributions to the regression setting is proposed. However, this proposed new mixture regression model is still not robust to high leverage outliers. In order to overcome this, a modified version of the proposed method, which fits the mixture regression based on the t-distribution to the data after adaptively trimming high leverage points, is also proposed. Furthermore, it is proposed to adaptively choose the degrees of freedom for the t-distribution using profile likelihood. The proposed robust mixture regression estimate has high efficiency due to the adaptive choice of degrees of freedom.

1. Introduction

Mixture regression models are well known as switching regression models in the econometrics literature, which were introduced by Goldfeld and Quandt (1973). These models have been widely used to investigate the relationship between variables coming from several unknown latent homogeneous groups and applied in many fields, such as business, marketing, and social sciences (Jiang and Tanner, 1999; Böhning, 1999; Wedel and Kamakura, 2000; McLachlan and Peel, 2000; Skrondal and Rabe-Hesketh, 2004; Frühwirth-Schnatter, 2006).

Let Z be a latent class variable such that given $Z = j$, the response y depends on the p-dimensional predictor $\mathbf{x}$ in a linear way

$$y = \mathbf{x}^T\boldsymbol{\beta}_j + \epsilon_j, \quad j = 1, 2, \ldots, m, \tag{1.1}$$

where m is the number of homogeneous groups (also called components in mixture models) in the population and $\epsilon_j \sim N(0, \sigma_j^2)$ is independent of $\mathbf{x}$. Suppose $P(Z = j) = \pi_j, j = 1, 2, \ldots, m$, and Z is independent of $\mathbf{x}$, then the conditional density of Y given $\mathbf{x}$, without observing Z, is

$$f(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j\phi(y; \mathbf{x}^T\boldsymbol{\beta}_j, \sigma_j^2), \tag{1.2}$$

where $\phi(\cdot; \mu, \sigma^2)$ is the density function of $N(\mu, \sigma^2)$ and $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \sigma_1, \ldots, \pi_m, \boldsymbol{\beta}_m, \sigma_m)^T$. The model (1.2) is the so called *mixture of regression models*. Hennig (2000) proved identifiability of model (1.2) under some general conditions for the covariates. In general, the model (1.2) is identifiable if the number of components, m, is smaller than the number of distinct $(p-1)$-dimensional hyperplanes that one needs to cover the covariates of each cluster. The above conditions are usually satisfied if the domain of $\mathbf{x}$ contains an open set in $\mathbb{R}^p$.

* Correspondence to: Department of Statistics, Kansas State University, Manhattan, KS 66506, USA. Tel.: +1 785 532 0522; fax: +1 785 532 7736.
E-mail address: wxyao@ksu.edu (W. Yao).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.07.019

The unknown parameter $\boldsymbol{\theta}$ in (1.2), given observations $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, is traditionally estimated by the maximum likelihood estimate (MLE):

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \left[\sum_{j=1}^{m} \pi_j \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_j, \sigma_j^2) \right]. \quad (1.3)$$

Note that the maximizer of (1.3) does not have an explicit solution and is usually estimated by the EM algorithm (Dempster et al., 1977).

It is well known that the log-likelihood function (1.3) is unbounded and goes to infinity if one or more observations lie exactly on one component hyperplane and the corresponding component variance goes to zero. When running the EM algorithm, some initial values might converge to the boundary point with small variance and very large log-likelihood. In such situations, our objective is to find a local maximum of (1.3) in the interior of parameter space (Kiefer, 1978; Peters and Walker, 1978). However, the challenge is to find this interior local maximum. Hathaway (1985, 1986) proposed putting some constraints on the parameter space such that the component variance has some low limit. Yao (2010) proposed using the profile likelihood and a graphical method to locate the interior local maximum. Practically, the interior local maximum can usually be found by starting from some "good" initial values such as the K-means (MacQueen, 1967) and the moment method estimator (Lindsay and Basak, 1993). Chen et al. (2008) also proposed using a penalized likelihood method to avoid the unboundedness of mixture likelihood. In this article, for simplicity of computation and comparison, we assume equal variance for all components.

The MLE $\hat{\boldsymbol{\theta}}$ in (1.3) works well when the error distribution is normal. However, the normality based MLE is sensitive to outliers or heavy-tailed error distributions. There is little research about how to estimate the mixture regression parameters robustly. Markatou (2000) and Shen et al. (2004) proposed using a weight factor for each data point to robustify the estimation procedure for mixture regression models. Neykov et al. (2007) proposed robust fitting of mixtures using the trimmed likelihood estimator (TLE). Bai et al. (2012) proposed a modified EM algorithm to robustly estimate the mixture regression parameters by replacing the least squares criterion in M step with a robust criterion. Bashir and Carter (2012) extended the idea of the S-estimator to mixture of linear regression. There are also some related robust methods for linear clustering (Hennig, 2002, 2003; Mueller and Garlipp, 2005; García-Escudero et al., 2009, 2010).

In this article, we propose a new robust mixture regression model by extending the mixture of t-distributions proposed by Peel and McLachlan (2000) to the regression setting. Similar to the traditional M-estimate for linear regression (Maronna et al., 2006), the proposed estimate is expected to be sensitive to high leverage outliers. To overcome this problem, we also propose a modified version of the new method by fitting the new model to the data after adaptively trimming high leverage points. Compared to the TLE, the proportion of trimming of our new method is data adaptive instead of a fixed value. In addition, we propose to use the profile likelihood to adaptively choose the degrees of freedom for the t-distribution. The proposed estimate has high efficiency, i.e., comparable performance to the traditional MLE when the error is normal, due to the adaptive choice of degrees of freedom. Using a simulation study and real data application, we compare the new method to some existing methods, and demonstrate the effectiveness of the proposed method.

The rest of this article is organized as follows. In Section 2, we introduce our new robust mixture linear regression models based on the t-distribution. In Section 3, we propose to further improve the robustness of the proposed method against high leverage outliers by adaptively trimming high leverage points. In Section 4, we introduce how to adaptively choose the degrees of freedom for the t-distribution. In Section 5, we compare the proposed method to the traditional MLE and some other robust methods by using a simulation study and real data application. Section 6 contains a discussion of possible future work.

2. Robust mixture regression using the t-distribution

In order to more robustly estimate the mixture regression parameters in (1.2), we assume that the error density $f_j(\epsilon)$ is a t-distribution with degrees of freedom ν_j and scale parameter σ:

$$f(\epsilon; \sigma, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)\sigma^{-1}}{(\pi\nu)^{\frac{1}{2}}\Gamma\left(\frac{\nu}{2}\right)\left\{1 + \frac{\epsilon^2}{\sigma^2\nu}\right\}^{\frac{1}{2}(\nu+1)}}. \quad (2.1)$$

We first assume that ν_js are known. We will discuss about how to adaptively choose ν_js in Section 4. The unknown parameter $\boldsymbol{\theta}$ in (1.2) can be estimated by maximizing the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j f(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j; \sigma, \nu_j) \right\}. \quad (2.2)$$

Note, however, the above log-likelihood does not have an explicit maximizer. Here, we also propose to use an EM algorithm to simplify the computation. Let

$$z_{ij} = \begin{cases} 1, & \text{if the } i\text{th observation is from the } j\text{th component;} \\ 0, & \text{otherwise,} \end{cases}$$