Computational Statistics and Data Analysis 93 (2016) 162–176

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Mixtures of quantile regressions

Qiang Wu [a,*], Weixin Yao [b]

[a] Department of Biostatistics, East Carolina University, Greenville, NC 27834, United States

[b] Department of Statistics, Kansas State University, Manhattan, KS 66506, United States

ARTICLE INFO

Article history:
Received 11 June 2013
Received in revised form 9 April 2014
Accepted 17 April 2014
Available online 5 May 2014

Keywords:
EM algorithm
Kernel density estimation
Mixture model
Multiple imputation
Quantile regression
Semi-parametric model

ABSTRACT

A semi-parametric mixture of quantile regressions model is proposed to allow regressions of the conditional quantiles, such as the median, on the covariates without any parametric assumption on the error densities. The median as a measure of center is known to be more robust to skewness and outliers than the mean. Modeling the quantiles instead of the mean not only improves the robustness of the model but also reveals a fuller picture of the data by fitting varying quantile functions. The proposed semi-parametric mixture of quantile regressions model is proven to be identifiable under certain weak conditions. A kernel density based EM-type algorithm is developed to estimate the model parameters, while a stochastic version of the EM-type algorithm is constructed for the variance estimation. A couple of simulation studies and several real data applications are conducted to show the effectiveness of the proposed model.

1. Introduction

Mixtures of regressions, or clusterwise regressions, have been a longstanding topic in the research of model-based clustering. When the population is heterogeneous and consists of several homogeneous groups, several regression models are simultaneously built to explain the relationships between the response variable and the covariates. The subjects are clustered based on the estimated classification probabilities. Some early results trace back to DeSarbo and Corn (1988), Jones and McLachlan (1992), and Arminger et al. (1999). In a classical mixture of regressions model, the conditional distribution of the response variable Y given the covariates $\mathbf{x}$ can be written as

$$f(y|\mathbf{x},\boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j \phi(y; \mathbf{x}^T\boldsymbol{\beta}_j, \sigma_j^2), \tag{1.1}$$

where $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \sigma_1, \ldots, \pi_m, \boldsymbol{\beta}_m, \sigma_m)$ and $\phi(\cdot; \mu, \sigma^2)$ is the normal probability density function (pdf) with mean μ and variance $\sigma^2 > 0$. In model (1.1), the unknown parameters include $\pi_j > 0$, $\boldsymbol{\beta}_j = (\beta_{0j}, \ldots, \beta_{pj})^T$, and $\sigma_j^2 > 0$ for $j = 1, \ldots, m$ where the mixing probabilities satisfy $\sum_j \pi_j = 1$. The covariates $\mathbf{x} = (1, \tilde{\mathbf{x}}^T)^T$ usually contain a leading one for fitting intercepts. The parameters can be estimated by the maximum likelihood estimator (MLE) using an EM algorithm. A number of applications of model (1.1) can be found at Wu and Sampson (2009), Skrondal and Rabe-Hesketh (2004), and Wedel and Kamakura (2000).

Much effort has been made recently to improve the robustness of model (1.1). For example, Garcia-Escudero et al. (2010) illustrate a robust clusterwise linear regressions method which trims off a fixed proportion of outlying observations and fits

[*] Corresponding author. Tel.: +1 252 744 6047; fax: +1 252 744 6044.
E-mail addresses: wuq@ecu.edu (Q. Wu), wxyao@ksu.edu (W. Yao).

http://dx.doi.org/10.1016/j.csda.2014.04.014

the rest of the data via a mixture of linear regressions model. This method has improved robustness to noisy data. Following Ingrassia et al. (2012), Ingrassia et al. (2014) develop a family of twelve mixture models each inheriting from a linear t-cluster weighted model. Such models allow the group assignments to depend on the covariates and the component distributions to feature heavier than normal tails. Wei (2012) and Yao et al. (2014) review some robust mixture regression models and propose a new one using the t-distributions as its components. While being robust to heavy tails of the component distributions, this method also trims the data based on a modified Mahalanobis distance to deal with possible high leverage points. Similarly, Song et al. (2014) introduce a robust mixture model fitting by the Laplace distribution.

Most relevantly to the research in this paper, Hunter and Young (2012) consider a semi-parametric mixture of regressions model

$$f(y|\mathbf{x},\boldsymbol{\theta},\mathbf{G}) = \sum_{j=1}^{m} \pi_j g(y - \mathbf{x}^T\boldsymbol{\beta}_j), \tag{1.2}$$

trying to relax the normality assumption to the greatest extent, where $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \ldots, \pi_m, \boldsymbol{\beta}_m)$ and g is an unknown symmetric pdf with mean equal to zero (its median is also zero since g is symmetric). Hunter and Young (2012) prove that model (1.2) is identifiable for the parameters $\boldsymbol{\theta}$ and the error pdf g up to a permutation on $\boldsymbol{\theta}$ if $\tilde{\boldsymbol{\beta}}_j = (\beta_{1j}, \ldots, \beta_{pj})$ for $j = 1, \ldots, m$ are distinct vectors in $\mathbb{R}^p$ and the domain of $\tilde{\mathbf{x}}$ contains an open set in $\mathbb{R}^p$. A location-shifted model is a special example of model (1.2). In their method, the parameters and the error pdf are estimated by a kernel density based EM-type algorithm.

When the error pdf is symmetric, the mixture of mean regressions model (1.2) works well in modeling the center location functions. However, there are situations where the error pdf is asymmetric in which case it seems reasonable to consider the median or other quantiles. The median as a measure of center is considered more robust to skewness and outliers than the mean. In this paper, a novel mixture of regressions model is introduced to allow regressions of the conditional quantiles, such as the median, on the covariates. In addition, it allows the component error densities to be different. Denote by g_j the error density of the jth component. Under a similar model specification as (1.2), the component pdf g_j is assumed to have its τth quantile equal to zero. As compared to the traditional mixtures of mean regressions, the mixtures of quantile regressions are more robust to non-normal component distributions and capable of revealing more detailed structure/information of the data by fitting varying conditional quantile functions. A kernel density based EM-type algorithm is developed to estimate the model parameters. In each iteration of the algorithm, the regression parameters are updated using a weighted quantile regression method, and the error pdfs are updated by a constrained kernel density estimation method. Moreover, a stochastic version of the EM-type algorithm based on multiple imputations is constructed for the variance estimation. A couple of simulation studies and several real data applications are conducted to demonstrate the effectiveness of the proposed model.

The rest of this article is organized as follows. In Section 2, we introduce the new mixture of quantile regressions model, prove its identifiability result, and detail the new kernel density based EM-type algorithm. In Section 3, we provide the stochastic EM-type algorithm for the variance estimation. In Sections 4 and 5, we present the simulation studies and the real data applications. Finally, some discussions are given in Section 6.

2. Mixtures of quantile regressions

The model setting for a mixture of τth quantile regressions is as follows. Let Z be a latent class variable with $\Pr(Z = j|\mathbf{x}) = \pi_j > 0$ for $j = 1, \ldots, m$, where $\mathbf{x} = (1, \tilde{\mathbf{x}}^T)^T$ is a $(p+1)$-dimensional vector of covariates with a leading one for fitting intercepts. Given $Z = j$, the response variable Y depends on the covariates $\mathbf{x}$ through

$$Y = \mathbf{x}^T\boldsymbol{\beta}_j(\tau) + \epsilon_j(\tau), \tag{2.1}$$

where $\boldsymbol{\beta}_j(\tau) = (\beta_{0j}(\tau), \ldots, \beta_{pj}(\tau))^T$ are the τth quantile regression coefficients for the jth component. The errors $\epsilon_j(\tau)$ are assumed to be independent of $\mathbf{x}$ and have pdfs $g_j(\cdot)$ whose τth quantiles are equal to zero. There is no additional constraint on the error pdfs as they are going to be estimated non-parametrically. We assume that the number of components $m > 1$ is known in advance. A regular choice for τ is 0.5 which corresponds to a median regression but it does not have to be. Since the model deals with only one quantile at a time, we suppress its dependency on τ in the following discussion for the notational ease.

Next, we prove that the mixture of τth quantile regressions model (2.1) is identifiable for $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \ldots, \pi_m, \boldsymbol{\beta}_m)$ and $\mathbf{G} = (g_1, \ldots, g_m)$ up to the same permutation on $\boldsymbol{\theta}$ and $\mathbf{G}$ if $\tilde{\boldsymbol{\beta}}_j = (\beta_{1j}, \ldots, \beta_{pj})$ for $j = 1, \ldots, m$ are distinct vectors in $\mathbb{R}^p$ and the domain of $\tilde{\mathbf{x}}$ contains an open set in $\mathbb{R}^p$. Necessary conditions and the identifiability of the model (2.1) are summarized in Theorem 2.1 whose proof is given in the Appendix. Of course, we have a flexibility to assume an equal error density $g_1 = \cdots = g_m$. In this case, a pooled density estimate can be found during the estimation. But the regression parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ must be distinct for the identifiability purpose.

Theorem 2.1. *Suppose that in the mixture of quantile regressions model (2.1), the domain of $\tilde{\mathbf{x}}$ contains an open set in $\mathbb{R}^p$, $0 < \pi_j < 1$, and $\tilde{\boldsymbol{\beta}}_j = (\beta_{1j}, \ldots, \beta_{pj})$ are distinct vectors in $\mathbb{R}^p$ for $j = 1, \ldots, m$. Then the parameters π_j, $\boldsymbol{\beta}_j$, and the error pdfs $g_j(\cdot)$ for $j = 1, \ldots, m$ are uniquely determined, up to a permutation, by the conditional density $f(y|\mathbf{x})$.*