Computational Statistics and Data Analysis 125 (2018) 44–56

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Unsupervised learning of mixture regression models for longitudinal data

Peirong Xu [a], Heng Peng [b], Tao Huang [c,*]

[a] College of Mathematics and Sciences, Shanghai Normal University, Shanghai, China
[b] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
[c] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China

ARTICLE INFO

Article history:
Received 6 June 2017
Received in revised form 3 November 2017
Accepted 18 March 2018
Available online 30 March 2018

Keywords:
Unsupervised learning
Model selection
Longitudinal data analysis
Quasi-likelihood
EM algorithm

ABSTRACT

This paper is concerned with learning of mixture regression models for individuals that are measured repeatedly. The adjective "unsupervised" implies that the number of mixing components is unknown and has to be determined, ideally by data driven tools. For this purpose, a novel penalized method is proposed to simultaneously select the number of mixing components and to estimate the mixture proportions and unknown parameters in the models. The proposed method is capable of handling both continuous and discrete responses by only requiring the first two moment conditions of the model distribution. It is shown to be consistent in both selecting the number of components and estimating the mixture proportions and unknown regression parameters. Further, a modified EM algorithm is developed to seamlessly integrate model selection and estimation. Simulation studies are conducted to evaluate the finite sample performance of the proposed procedure. And it is further illustrated via an analysis of a primary biliary cirrhosis data set.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

In many medical studies, the marker of disease progression and a variety of characteristics are routinely measured during the patients' follow-up visit to decide on future treatment actions. Consider a motivating Mayo Clinic trial with primary biliary cirrhosis (PBC), wherein a number of serological, clinical and histological parameters were recorded for each of 312 patients from 1974 to 1984. This longitudinal study had a median follow-up time of 6.3 years as some patients missed their appointments due to worsening medical condition of some labs. It is known that PBC is a fatal chronic cholesteric liver disease, which is characterized histopathologically by portal inflammation and immune-mediated destruction of the intrahepatic bile ducts (Pontecorvo et al., 1992). It can be divided into four histologic stages, but with nonuniformly affected liver. The diagnosis of PBC is important for the medical treatment with Ursodiol has been shown to halt disease progression and improve survival without need for liver transplantation (Talwalkar and Lindor, 2003). Therefore, one goal of the study was the investigation of the serum bilirubin level, an important marker of PBC progression, in relation to the time and to potential clinical and histological covariates. Another issue that should be accounted for is the unobservable heterogeneity between subjects that may not be explained by the covariates. The changes in inflammation and bile ducts occur at different rates and with varying degrees of severity in different patients, so the heterogeneous patients could potentially belong to different latent groups. To address these problems, there is a demand for mixture regression modeling for subjects on the basis of longitudinal measurements.

* Corresponding author.
E-mail address: huang.tao@mail.shufe.edu.cn (T. Huang).

https://doi.org/10.1016/j.csda.2018.03.012
0167-9473/© 2018 Elsevier B.V. All rights reserved.

There are various research works on mixture regression models for longitudinal outcome data, particularly in the context of model-based probabilistic clustering (Fraley and Raftery, 2002). For example, De la Cruz-Mesía et al. (2008) proposed a mixture of non-linear hierarchical models with Gaussian subdistributions; McNicholas and Murphy (2010) extended the Gaussian mixture models with Cholesky-decomposed group covariance structure; Komárek and Komárková (2013) introduced a generalized linear mixed model for components' densities under the Gaussian mixture framework; Heinzl and Tutz (2013) considered linear mixed models with approximate Dirichlet process mixtures. Other relevant work includes Celeux et al. (2005), Booth et al. (2008), Pickles and Croudace (2010), Maruotti (2011), Erosheva et al. (2014) and some of the references therein. Compared with heuristic methods such as the k-means method (Genolini and Falissard, 2010), issues like the selection of the number of clusters (or components) can be addressed in a principled way. However, most of them assume a parametric mixture distribution, which may be too restrictive and invalid in practice when the true data-generating mechanism indicates otherwise.

A key concern for the performance of mixture modeling is the selection of the number of components. A mixture with too many components may overfit the data and result in poor interpretations. Many statistical methods have been proposed in the past few decades by using the information criteria. For example, see Leroux (1992), Roeder and Wasserman (1997), Hennig (2004), De la Cruz-Mesía et al. (2008) and many others. However, these methods are all based on the complete model search algorithm, which result in heavy computation burden. To improve the computational efficiency, data-driven procedures are much more preferred. Recently, Chen and Khalili (2008) used the SCAD penalty (Fan and Li, 2001) to penalize the difference of location parameters for mixtures of univariate location distributions; Komárek and Lesaffre (2008) suggested to penalize the reparameterized mixture weights in the generalized mixed model with Gaussian mixtures; Heinzl and Tutz (2014) constructed a group fused lasso penalty in linear-mixed models; Huang et al. (in press) proposed a penalized likelihood method in finite Gaussian mixture models. Most of them are developed for independent data or based on the full likelihood. However, the full likelihood is often difficult to specify in formulating a mixture model for longitudinal data, particularly for correlated discrete data.

Instead of specifying the form of distribution of the observations, a quasi-likelihood method (Wedderburn, 1974) gives consistent estimates of parameters in mixture regression models that only needs the relation between the mean and variance of each observation. Inspired by its nice property, in this paper, we propose a new penalized method based on quasi-likelihood for mixture regression models to deal with the above mentioned problems simultaneously. This would be the first attempt to handle both balanced and unbalanced longitudinal data that only requires the first two moment conditions of the model distribution. By penalizing the logarithm of mixture proportions, our approach can simultaneously select the number of mixing components and estimate the mixture proportions and unknown parameters in the semiparametric mixture regression model. The number of components can be consistently selected. And given the number of components, the estimators of mixture proportions and regression parameters can be root-n consistent and asymptotically normal. By taking account of the within-component dispersion, we further develop a modified EM algorithm to improve the classification accuracy. Simulation results and the application to the motivating PBC data demonstrate the feasibility and effectiveness of the proposed method.

The rest of the paper is organized as follows. In Section 2, we introduce a new penalized method for learning semiparametric mixture regression models with longitudinal data. Section 3 presents the corresponding theoretical properties and Section 4 provides a modified EM algorithm for implementation. In Section 5, we assess the finite sample performance of the proposed method via simulation studies. We apply the proposed method to the PBC data in Section 6, and conclude the paper with Section 7. All technical proofs are provided in Appendix.

2. Learning semiparametric mixture of regressions

2.1. Model specification

In a longitudinal study, suppose Y_{ij} is the response variable measured at the jth time point for the ith subject, and X_{ij} is the corresponding $p \times 1$ vector of covariates, $i = 1, \ldots, n, j = 1, \ldots, m_i$. Let $Y_i = (Y_{i1}, \ldots, Y_{im_i})^T$ and $X_i = (X_{i1}, \ldots, X_{im_i})^T$. In general, the observations for different subjects are independent, but they may be correlated within the same subject. We assume that the observations of each subject belong to one of K classes (components) and $u_i \in \{1, \ldots, K\}$ is the corresponding latent class variable. Assume that u_i has a discrete distribution $\mathrm{P}(u_i = k) = \pi_k$, where $\pi_k, k = 1, \ldots, K$, are the positive mixture proportions satisfying $\sum_{k=1}^{K} \pi_k = 1$. Given $u_i = k$ and X_{ij}, suppose the conditional mean of Y_{ij} is

$$\mu_{ijk} \equiv \mathrm{E}(Y_{ij} \mid X_{ij}, u_i = k) = g(X_{ij}^T \beta_k), \tag{2.1}$$

where g is a known link function, and β_k is a p-dimensional unknown parameter vector. The corresponding conditional variance of Y_{ij} is given by

$$\sigma_{ijk}^2 \equiv \mathrm{var}(Y_{ij} \mid X_{ij}, u_i = k) = \phi_k V(\mu_{ijk}), \tag{2.2}$$

where V is a known positive function and ϕ_k is a unknown dispersion parameter. In other words, conditioning on X_{ij}, the response variable Y_{ij} follows a mixture distribution

$$Y_{ij} \mid X_{ij} \sim \sum_{k=1}^{K} \pi_k f_k(Y_{ij} \mid X_{ij}^T \beta_k, \phi_k),$$

Download English Version:

https://daneshyari.com/en/article/6868665

Download Persian Version:

https://daneshyari.com/article/6868665

Daneshyari.com