ON MODEL-BASED CLUSTERING USING QUANTILE REGRESSION

Carlo Gaetan [1], Paolo Girardi[2] and Victor Muthama Musau [3]

[1] DAIS, Ca' Foscari of Venice (Italy) , (e-mail: gaetan@unive.it)

[2] Department of Developmental and Social Psychology, University of Padova (Italy), (e-mail: paolo.girardi@unipd.it)

[3] Department of Pure and Applied Sciences, Kirinyaga University (Kenya), (e-mail: vmusau@kyu.ac.ke)

ABSTRACT: Clustering general regression functions or curves can suffer of lack of robustness when we consider the usual Gaussian assumption. In this note we introduce a new model-based clustering method that tries to overcome this limitation.

KEYWORDS: Functional data, hierarchical Bayesian model, MCMC algorithm

1 Introduction

Unlike the classical clustering approaches such as agglomerative hierarchical clustering and K-means clustering, which are largely heuristic and not based on formal statistical models, model-based clustering takes a likelihood based approach thus permitting inference to be drawn on the clusters. These techniques are based on the finite mixture model theory (Fraley & Raftery, 2002), where each mixture component corresponds to a cluster. However, fundamental concerns remain about robustness and in particular the choice of distribution representing the within cluster density. The Gaussian mixture models are historically the most popular tool for model-based clustering. However, if the distribution of the observed variable is characterized by asymmetry and presence of outliers, a Gaussian distribution may not be an appropriate within cluster density. The direct link that exists between univariate quantile regression approach and the Asymmetric Laplace Distribution (ALD) forms our basis of introducing a clustering model based on finite mixture of ALDs to group individuals subject to heterogeneity due to regressor variables.

2 Methodology

We start by considering a vector, $\mathbf{y} = (y_1, \ldots, y_T)'$ of responses y_t and the associated design matrix $\mathbf{X} = (x_1, \ldots, x_T)'$ that collects the vectors x_t of L covariates. Further, let $Q_p(y_t|x_t)$, for $0 < p < 1$, be the pth quantile regression function of y_t given x_t which can be modelled as $Q_p(y_t|x_t) = x_t'\beta$, where β is a vector of unknown parameters to be estimated. The regression coefficient estimate is obtained by minimizing (Koenker & Bassett, 1978)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{t=1}^{T} \rho_p(y_t - x_t'\beta) \tag{1}$$

where $\rho_p(\cdot)$ is the check loss function defined by $\rho_p(x) = x(p - I(x < 0))$ and $I(\cdot)$ denotes the usual indicator function. Koenker and Machado (1999) showed that there is a direct relationship between minimizing (1) and the maximum likelihood theory using independently distributed asymmetric Laplace variable with density

$$\operatorname{ald}(y_t|\beta, \sigma, p) = \frac{p(1-p)}{\sigma} \exp\left\{ -\rho_p\left(\frac{y_t - x_t'\beta}{\sigma} \right) \right\} \tag{2}$$

where $\sigma > 0$ and $0 < p < 1$ represents the skewness parameter that can be used directly to model any quantile of interest.

According to the finite mixture framework theory we define the likelihood of our mixture model for a single vector $\mathbf{y}$ as

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, p|\mathbf{y}) = \sum_{k=1}^{K} \alpha_k \prod_{t=1}^{T} \operatorname{ald}_1(y_t|\beta_k, \sigma_k, p) = \sum_{k=1}^{K} \alpha_k \operatorname{ALD}(\mathbf{y}|\beta_k, \sigma_k, p) \tag{3}$$

where $\boldsymbol{\beta} = (\beta_1', ..., \beta_K')'$, $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_K)'$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)'$ is the vector of the mixing proportions for the K clusters which satisfy the conditions $0 < \alpha_k < 1$ and $\sum_{k=1}^{K} \alpha_k = 1$.

We now consider a set $\mathcal{Y} = \{\mathbf{y}_i, i = 1, ..., n\}$ of n vectors $\mathbf{y}_i = (y_{i1}, ..., y_{iT})'$ of independent observations. and we want to split the data set $\mathcal{Y}$ into K clusters. According the mixture model (3) the cluster membership $c_i \in \{1, ..., K\}$, where $c_i = k$ indicates that the ith vector $\mathbf{y}_i$ belongs to cluster k is a multinomial random variable with parameter $\boldsymbol{\alpha}$.

We adopt a Bayesian approach to make inference on the model parameters $\boldsymbol{\psi} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\sigma}')'$. Moreover it is possible to get the posterior probability of membership of a single vector, $\Pr(c_i = \cdot|\mathcal{Y})$. In doing this we first note that

Kozumi and Kobayashi (2011) represent the density (2) as a location scale mixture of Gaussian distributions i.e.

$$y_t = x_t'\beta + \theta w_t + \omega\sqrt{\sigma w_t}\nu_t \tag{4}$$

where $\nu_t \sim N(0,1)$, and w is an exponential random variable with $E(w) = \sigma$. Here ν and w are mutually independent and $\theta = (1-2p)/\{p(1-p)\}$ and $\omega^2 = 2/\{p(1-p)\}$.

Equation (4) constitutes the first stage of a hierarchical Bayesian model where the prior distribution on the cluster specific parameters and as well as the mixing proportions are specified as conjugate priors to having closed form conditional posterior densities which are easy to sample from in a MCMC algorithm.

A conjugate prior for the mixing proportions $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)'$ is the Dirichlet distribution, $\boldsymbol{\alpha} \sim D(\zeta_1, ..., \zeta_K)$. A straightforward prior for β_k is the multivariate Gaussian distribution, $\mathcal{N}(b_0, \Sigma_0)$ where by setting $b_0 = 0$ and $\Sigma_0 = aI$, for $a \gg 0$, leads to an improper prior. Finally we propose the inverse gamma distribution, $\mathrm{IG}(s_0, d_0)$, as the prior for σ_k where the shape and scale parameters, s_0 and d_0 respectively, are known.

Musau (2021) gives a complete account on how we can devise an MCMC algortihm for sampling from the posterior distribution of $\boldsymbol{\psi}$.

3 Numerical results

We exemplify our proposal with a clustering problem for functional data. We consider the well-known Canadian temperature dataset available in the R package `fda`. The dataset consists of the daily measured temperatures from 35 Canadian weather stations across the country.

Under functional data framework (Ramsay & Silverman, 2005), daily temperature data, y_t, can be described by a linear combination of $L = 65$ cubic spline basis functions, $y_t \simeq \sum_{j=1}^{L} \beta_j B_j(t) = x_t'\beta$, with knots which are equally distributed over the range of time.

The funHDDC clustering algorithm (Bouveyron & Jacques, 2011) on this data selects $K = 4$ as the optimal number of clusters. Figure 1 (left panel) summarize the resulted clusters.

For each of the 35 stations we randomly introduce outliers (y_t=0) at 10% of the total observation points. This distorts the general trend of the data, as shown in right panel of Figure 1, making reconstruction of the clusters difficult.

We apply our mixture model setting $p = 0.5$, i.e. we consider a robust median regression and we compare its performance in reconstructing the 4

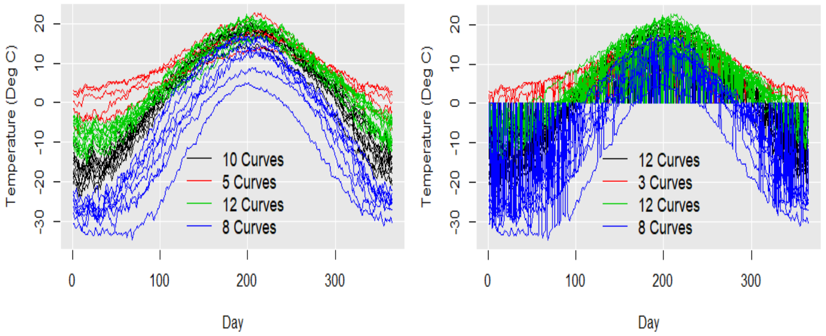

Figure 1. *Clustering of the 35 temperature curves as obtained by funHDDC algorithm (left panel) and results with curves contaminated by outliers (right panel).*

clusters with the previous algorithm, leading to a perfect agreement. These results generally indicate a good performance of our proposed algorithm when clustering data characterized by outlying observations.

References

BOUVEYRON, C., & JACQUES, J. 2011. Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, **5**, 281–300.

FRALEY, C., & RAFTERY, A.E. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

KOENKER, R., & BASSETT, G. 1978. Regression quantiles. *Econometrica*, **46**, 33–50.

KOENKER, R., & MACHADO, J. A.F. 1999. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, **94**, 1296–1310.

KOZUMI, H., & KOBAYASHI, G. 2011. Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, **81**, 1565–1578.

MUSAU, V.M. 2021. *Model-based Clustering Using Quantile Regression*. Ph.D. thesis, University of Padua, Italy.

RAMSAY, J.O., & SILVERMAN, B.W. 2005. *Functional Data Analysis*. Springer, New York.