Computational Statistics and Data Analysis 54 (2010) 3057–3069

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Robust clusterwise linear regression through trimming

L.A. García-Escudero *, A. Gordaliza, A. Mayo-Iscar, R. San Martín

Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, Prado la Magdalena s/n, 47005 Valladolid, Spain

ARTICLE INFO

Article history:
Received 12 December 2008
Received in revised form 2 July 2009
Accepted 3 July 2009
Available online 8 July 2009

ABSTRACT

The presence of clusters in a data set is sometimes due to the existence of certain relations among the measured variables which vary depending on some hidden factors. In these cases, observations could be grouped in a natural way around linear and nonlinear structures and, thus, the problem of doing robust clustering around linear affine subspaces has recently been tackled through the minimization of a trimmed sum of orthogonal residuals. This "orthogonal approach" implies that there is no privileged variable playing the role of response variable or output. However, there are problems where clearly one variable is wanted to be explained in terms of the other ones and the use of vertical residuals from classical linear regression seems to be more advisable. The so-called TCLUST methodology is extended to perform robust clusterwise linear regression and a feasible algorithm for the practical implementation is proposed. The algorithm includes a "second trimming" step aimed to diminishing the effect of leverage points.

1. Introduction

Most of non-hierarchical clustering methods are based on the idea of forming clusters around "objects". These objects are geometrical structures which represent the typical behavior of data belonging to each group. The search for these objects is often done by minimizing a criterium based on distances of data points to their closest objects. When data come from a population model, the objects become population features which are interesting to be understood and estimated.

The first kind of objects considered in the literature of Cluster Analysis were "centers", that gave rise to the well-known *k*-means method (McQueen, 1967; Hartigan and Wong, 1979) as well as to its robust version, the trimmed *k*-means method (Cuesta-Albertos et al., 1997). However, the presence of clusters in a data set is sometimes due to the existence of certain relations among the measured variables which may be different depending on some unknown or hidden factors. Thus, observations can be grouped in a natural way around more complex objects which can adopt the form of linear or nonlinear structures. Clustering around linear affine subspaces and more general manifolds is receiving a lot of attention in the literature. This is partly motivated by the existence of interesting fields of application like computer vision, pattern recognition, tomography, fault detection, etc., where this type of cluster usually appears. A common feature in all these application fields is that noisy data frequently appear. Therefore protection against outliers is a desirable property to be required for any designed procedure.

García-Escudero et al. (2009) contains a large collection of references that reflect the state of art on clustering around linear structures. That paper also presented a new approach aimed at performing robust clustering around linear subspaces. The approach is based on the minimization of a trimmed sum of orthogonal residuals and it may be seen as a robust version for the Linear Grouping Algorithm introduced by Van Aelst et al. (2006). It may be also viewed as a robust extension of the classical Principal Components Analysis (PCA) to the Cluster Analysis setup (see Serneels and Verdonck (2008) for a recent review on robust PCA method).

* Corresponding author. Tel.: +34 983 186313; fax: +34 983 433111.
E-mail addresses: lagarcia@eio.uva.es (L.A. García-Escudero), alfonsog@eio.uva.es (A. Gordaliza), agustinm@eio.uva.es (A. Mayo-Iscar), rsmartin@eio.uva.es (R. San Martín).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.07.002

The "orthogonal" choice for the residuals implies that there is no privileged variable to be used as a response or output variable. However, there are situations where the interest clearly rests on explaining one variable in terms of other ones and the use of classical linear regression residuals seems to be more advisable. This would allow for addressing classical tasks in Regression Analysis like, for instance, understanding the role of each explanatory variable, forecasting the response variable, validating the model through analysis of residuals, and so on. To this different approach we can call "clusterwise linear regression".

Several approaches related to clusterwise linear regression can be found in the literature. Hosmer (1974) and Lenstra et al. (1982) are examples of early references devoted to the "two regression lines" case. In the econometric and chemometric settings we can find the "switching regression model" which was introduced from a maximum likelihood point of view in Goldfeld and Quandt (1976) and has also been later addressed through a Bayesian approach (e.g., Hurn et al. (2003)). In the Machine Learning community literature, we can also find the "multiple model estimation" approach (see Cherkassky and Ma (2005), and references therein). In these two mentioned approaches, the emphasis is mainly put on aspects related to the fit of the model instead of clustering aspects. Another interesting reference is Hennig (2003) where a "fixed point" approach was analyzed. Finally, Neykov et al. (2007) has recently introduced a mixture fitting approach based on trimmed likelihood that will be discussed later.

In this paper we extend the TCLUST methodology in García-Escudero et al. (2008) to the context of robust clusterwise linear regression. The TCLUST methodology was there introduced to perform robust clustering assuming multivariate normal clusters with different sizes and different covariance matrices. The approach relies on modeling the data through an adaptation of the "spurious-outliers model" (Gallegos and Ritter, 2005). In the here proposed extension, the TCLUST methodological flexibility will allow for different scatters for the regression errors together with different group weights. A constraint on the error term scatters is also needed in order to avoid singularities on the objective function defining the problem. Apart from dealing with a clusterwise regression problem, the main difference with the previously developed TCLUST methodology relies on how k scatter parameters are now controlled instead of the eigenvalues of k covariance matrices. The possibility for a further "second" trimming is now also considered.

The structure of the paper is as follows. Section 2 is devoted to present the proposed methodology explaining the role of all its ingredients. An algorithm inspired by this methodology is outlined in this section. A detailed description of the algorithm can be found in the Appendix. The importance of imposing a scatter similarity constraints is discussed in Section 3. An additional "second" trimming step is also presented in Section 4. This step is designed to improve the protection against outliers in x which is an interesting feature in Regression Analysis. That further trimming allows to diminish the effect of some leverage points that (although they do not break down the procedure) could entail important biases in the determination of the underlying linear structures. This second trimming is also appropriate to avoid classification errors that sometimes occur due to the artificial elongations of influence zones of linear clusters. A simulation study is carried out in Section 5. Some applications based on real data sets are developed in Section 6. These examples illustrate the interest of the presented methodology to perform allometric studies in Biology. Finally, some conclusions and further directions are presented in Section 7.

2. Proposed methodology

Let us assume that the total number of groups k to search for is fixed in advance and that we want to be able to handle the presence of a fixed proportion α of outlying observations. We consider a data set $\{X_i\}_{i=1}^n$ with $X_i = (y_i, x_i')'$ where $y_i \in \mathbb{R}$ is the value for the response variable and $x_i \in \mathbb{R}^p$ are the values of the explanatory variables. If X_i is being assigned to group j then we assume that (conditional to the value of x_i) the value of the response variable y_i was generated by a $N(x_i'\beta_j, \sigma_j^2)$ distribution, i.e. a normal distribution with mean $x_i'\beta_j$ and variance σ_j^2, for some unknown parameters $\beta_j \in \mathbb{R}^p$ and $\sigma_j^2 > 0$. Some weights π_j satisfying $\sum_{j=1}^k \pi_j = 1$ for the groups will be also considered in this approach.

As it was done in García-Escudero et al. (2008), we can modify the "spurious-outlier" model in Gallegos (2002) and Gallegos and Ritter (2005) considering a "likelihood" for the clusterwise linear regression problem as follows:

$$\left[\prod_{j=1}^{k}\prod_{i\in R_j}\pi_j f(y_i; x_i', \beta_j, \sigma_j)\right]\left[\prod_{i\notin R} g_i(y_i; x_i')\right], \tag{1}$$

with

$$f(y; x', \beta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-(y - x'\beta)^2/(2\sigma^2)), \tag{2}$$

$R = R_1 \cup \cdots \cup R_k \subset \{1, \ldots, n\}$, $R_r \cap R_s = \emptyset$ for $r \neq s$ and such that $\#R = n - [n\alpha]$. The $g_i(\cdot; x_i')$'s are probability density functions in $\mathbb{R}$ (conditional to the value of x_i') satisfying the condition

$$\arg\max_{\mathcal{R}} \max_{\beta_j, \sigma_j} \prod_{j=1}^{k}\prod_{i\in R_j}\pi_j f(y_i; x_i', \beta_j, \sigma_j) \subseteq \arg\max_{\mathcal{R}} \prod_{i\notin \cup_{j=1}^k R_j} g_i(y_i; x_i')$$

where $\mathcal{R}$ stands for the set of all partitions of the indices $\{1, \ldots, n\}$ onto $k+1$ groups: $R_1, \ldots, R_k$ such that $\#\{R_1 \cup \cdots \cup R_k\} = n - [n\alpha]$ and another additional group with $[n\alpha]$ data points. This condition is quite reasonable for the g_i's whenever a fraction