Computational Statistics and Data Analysis 71 (2014) 159–182

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Model-based clustering via linear cluster-weighted models

Salvatore Ingrassia [a,*], Simona C. Minotti [b], Antonio Punzo [a]

[a] Department of Economics and Business, University of Catania, Corso Italia 55, 95129 Catania, Italy

[b] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy

ARTICLE INFO

Article history:
Received 17 June 2012
Received in revised form 7 February 2013
Accepted 7 February 2013
Available online 14 February 2013

Keywords:
Cluster-weighted model
Mixture models with random covariates
Model-based clustering
Multivariate *t* distribution

ABSTRACT

A novel family of twelve mixture models with random covariates, nested in the linear *t* cluster-weighted model (CWM), is introduced for model-based clustering. The linear *t* CWM was recently presented as a robust alternative to the better known linear Gaussian CWM. The proposed family of models provides a unified framework that also includes the linear Gaussian CWM as a special case. Maximum likelihood parameter estimation is carried out within the EM framework, and both the BIC and the ICL are used for model selection. A simple and effective hierarchical–random initialization is also proposed for the EM algorithm. The novel model-based clustering technique is illustrated in some applications to real data. Finally, a simulation study for evaluating the performance of the BIC and the ICL is presented.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

In direct applications of finite mixture models (see Titterington et al., 1985, pp. 2–3), we assume that each mixture-component represents a group (or cluster) in the original data. The term "model-based clustering" has been used to describe the adoption of mixture models for clustering or, more often, to describe the use of a family of mixture models for clustering (see Fraley and Raftery, 1998; McLachlan and Basford, 1988). An overview of mixture models is given in Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Peel (2000), and Frühwirth-Schnatter (2006).

This paper focuses on data arising from a real-valued random vector $\left(Y, \boldsymbol{X}'\right)' : \Omega \rightarrow \mathbb{R}^{d+1}$, having joint density $p\left(y, \boldsymbol{x}\right)$, where Y is the response variable and $\boldsymbol{X}$ is the vector of covariates. Standard model-based clustering techniques assume that Ω can be partitioned into G groups $\Omega_1, \ldots, \Omega_G$. As for finite mixtures of linear regressions (see, e.g., Leisch, 2004; Frühwirth-Schnatter, 2006, Chapter 8) we assume that, for each Ω_g, the dependence of Y on $\boldsymbol{x}$ can be modeled by

$$Y = \mu\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right) + \varepsilon_g = \beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{x} + \varepsilon_g,$$

where $\boldsymbol{\beta}_g = \left(\beta_{0g}, \boldsymbol{\beta}'_{1g}\right)'$, $\mu\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right) = E\left(Y|\boldsymbol{X} = \boldsymbol{x}, \Omega_g\right)$ is the linear regression function and ε_g is the error variable, independent with respect to $\boldsymbol{X}$, with zero mean and finite constant variance σ^2_g, $g = 1, \ldots, G$. However, as highlighted in Hennig (2000), finite mixtures of linear regressions are inadequate for most of the applications because they assume *assignment independence*: the probability for a point $\left(y, \boldsymbol{x}'\right)'$ to be generated by one of the mixture components has to be the same for all covariates values $\boldsymbol{x}$. In other words, the assignment of the data points to the clusters has to be independent of the covariates.

Here, differently from finite mixtures of linear regressions, we assume random covariates having a parametric specification. This allows for *assignment dependence*: the covariate distributions of the mixture components can also be

* Corresponding author. Tel.: +39 095 7537732; fax: +39 095 7537610.
E-mail addresses: s.ingrassia@unict.it (S. Ingrassia), simona.minotti@unimib.it (S.C. Minotti), antonio.punzo@unict.it (A. Punzo).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2013.02.012

distinct. In the framework of mixture models with random covariates, the cluster weighted model (CWM; Gershenfeld, 1997), with equation

$$p(y,\boldsymbol{x}) = \sum_{g=1}^{G} \pi_g p\left(y,\boldsymbol{x}|\Omega_g\right) = \sum_{g=1}^{G} \pi_g p\left(y|\boldsymbol{x},\Omega_g\right) p\left(\boldsymbol{x}|\Omega_g\right), \tag{1}$$

also called the saturated mixture regression model by Wedel (2002), constitutes a reference approach to model the joint density. In (1), normality of both $p\left(y|\boldsymbol{x},\Omega_g\right)$ and $p\left(\boldsymbol{x}|\Omega_g\right)$ is commonly assumed (see, e.g., Gershenfeld, 1997; Punzo, 2012). Alternatively, Ingrassia et al. (2012) propose the use of the t distribution, which provides more robust fitting for groups of observations with longer than normal tails or noise data (see, e.g., Zellner 1976, Lange et al. 1989, Peel and McLachlan 2000, McLachlan and Peel 2000, Chapter 7, Chatzis and Varvarigou 2008, and Greselin and Ingrassia 2010). In particular, the authors consider

$$p\left(y|\boldsymbol{x},\Omega_g\right) = h_t\left(y|\boldsymbol{x};\boldsymbol{\xi}_g,\zeta_g\right) = \frac{\Gamma\left(\frac{\zeta_g+1}{2}\right)}{\left(\pi\zeta_g\sigma_g^2\right)^{\frac{1}{2}}\left\{1+\delta\left[y,\mu\left(\boldsymbol{x};\boldsymbol{\beta}_g\right);\sigma_g^2\right]\right\}^{\frac{\zeta_g+1}{2}}} \tag{2}$$

and

$$p\left(\boldsymbol{x}|\Omega_g\right) = h_{t_d}\left(\boldsymbol{x};\boldsymbol{\vartheta}_g,\nu_g\right) = \frac{\Gamma\left(\frac{\nu_g+d}{2}\right)\left|\boldsymbol{\Sigma}_g\right|^{-\frac{1}{2}}}{\left(\pi\nu_g\right)^{\frac{d}{2}}\left[1+\delta\left(\boldsymbol{x},\boldsymbol{\mu}_g;\boldsymbol{\Sigma}_g\right)\right]^{\frac{\nu_g+d}{2}}}, \tag{3}$$

with $\boldsymbol{\xi}_g = \left\{\boldsymbol{\beta}_g,\sigma_g^2\right\}$, $\boldsymbol{\vartheta}_g = \left\{\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\right\}$, $\delta\left[y,\mu(\boldsymbol{x};\boldsymbol{\beta}_g);\sigma_g^2\right] = \left[y-\mu\left(\boldsymbol{x};\boldsymbol{\beta}_g\right)\right]^2/\sigma_g^2$, and $\delta\left(\boldsymbol{x},\boldsymbol{\mu}_g;\boldsymbol{\Sigma}_g\right) = \left(\boldsymbol{x}-\boldsymbol{\mu}_g\right)'\boldsymbol{\Sigma}_g^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_g\right)$. Thus, (2) is the density of a (generalized) univariate t distribution, with location parameter $\mu\left(\boldsymbol{x};\boldsymbol{\beta}_g\right)$, scale parameter σ_g^2, and ζ_g degrees of freedom, while (3) is the density of a multivariate t distribution with location parameter $\boldsymbol{\mu}_g$, inner product matrix $\boldsymbol{\Sigma}_g$, and ν_g degrees of freedom. By substituting (2) and (3) into (1), we obtain the linear t CWM

$$p(y,\boldsymbol{x};\psi) = \sum_{g=1}^{G} \pi_g h_t\left(y|\boldsymbol{x};\boldsymbol{\xi}_g,\zeta_g\right) h_{t_d}\left(\boldsymbol{x};\boldsymbol{\vartheta}_g,\nu_g\right), \tag{4}$$

where the set of all unknown parameters is denoted by $\psi = \left\{\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_G\right\}$, with $\boldsymbol{\psi}_g = \left\{\pi_g,\boldsymbol{\xi}_g,\zeta_g,\boldsymbol{\vartheta}_g,\nu_g\right\}$. Quite recent developments in CWMs are proposed by Punzo (2012), who considers polynomial regressions, and by Subedi et al. (in press) who model data with a large number of covariates.

In this paper, we introduce a family of twelve linear CWMs obtained from (4) by imposing convenient component distributional constraints. If $\zeta_g,\nu_g \to \infty$, the linear Gaussian (normal) CWM is obtained as a special case. The resulting models are easily interpretable and appropriate for describing various practical situations. In particular, they also allow us to infer if the group-structure of the data is due to the contribution of $\boldsymbol{X}$, $Y|\boldsymbol{X}$, or both.

The paper is organized as follows. In Section 2, we recall model-based clustering according to the CW approach, and give some preliminary results. In Section 3, we introduce the novel family of models. Model fitting in the EM paradigm is presented in Section 4, related computational aspects are addressed in Section 5, and model selection is discussed in Section 6. In Section 7 some applications to real data are illustrated. In Section 8 simulations for a comparison between BIC and ICL are described. Finally, in Section 9, we give a summary of the paper and some directions for further research.

2. Preliminary results for model-based clustering

This section recalls some basic ideas on model-based clustering according to the CWM approach and provides some preliminary results that will be useful for definition and justification of our family of models.

Let $\left(y_1,\boldsymbol{x}_1'\right)',\ldots,\left(y_N,\boldsymbol{x}_N'\right)'$ be a sample of size N from (4). Once $\boldsymbol{\psi}$ is estimated, the posterior probability that the generic unit $\left(y_n,\boldsymbol{x}_n'\right)'$, $n=1,\ldots,N$, comes from component Ω_g is given by

$$\tau_{ng} = P\left(\Omega_g|y_n,\boldsymbol{x}_n;\boldsymbol{\psi}\right) = \frac{\pi_g h_t\left(y_n|\boldsymbol{x}_n;\boldsymbol{\xi}_g,\zeta_g\right) h_{t_d}\left(\boldsymbol{x}_n;\boldsymbol{\vartheta}_g,\nu_g\right)}{p\left(y_n,\boldsymbol{x}_n;\boldsymbol{\psi}\right)}, \quad g=1,\ldots,G. \tag{5}$$

These probabilities, which depend on both marginal and conditional densities, represent the basis for clustering and classification.

The following two propositions, which generalize some results given in Ingrassia et al. (2012), require the preliminary definition of

$$p\left(y|\boldsymbol{x};\underset{\sim}{\pi},\boldsymbol{\xi},\zeta\right) = \sum_{g=1}^{G} \pi_g h_t\left(y|\boldsymbol{x};\boldsymbol{\xi}_g,\zeta_g\right) \tag{6}$$

Download English Version:

https://daneshyari.com/en/article/6870326

Download Persian Version:

https://daneshyari.com/article/6870326

Daneshyari.com