International Journal of Approximate Reasoning 91 (2017) 160–178

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Clusterwise linear regression modeling with soft scale constraints ☆

Roberto Di Mari [a,*], Roberto Rocci [b], Stefano Antonio Gattone [c]

[a] Department of Economics and Business, University of Catania, Italy
[b] Department of Economics and Finance, University of Rome Tor Vergata, Italy
[c] Department of Philosophical and Social Sciences, Economics and Quantitative Methods, University G. d'Annunzio, Chieti-Pescara, Italy

ARTICLE INFO

Article history:
Received 10 November 2016
Received in revised form 9 September 2017
Accepted 10 September 2017
Available online 18 September 2017

Keywords:
Clusterwise linear regression
Adaptive constraints
Regression equivariance
Plausible bounds
Soft estimators
Constrained EM algorithm

ABSTRACT

Constrained approaches to maximum likelihood estimation in the context of finite mixtures of normals have been presented in the literature. A fully data-dependent soft constrained method for maximum likelihood estimation of clusterwise linear regression is proposed, which extends previous work in equivariant data-driven estimation of finite mixtures of normals. The method imposes soft scale bounds based on the homoscedastic variance and a cross-validated tuning parameter c. In our simulation studies and real data examples we show that the selected c will produce an output model with clusterwise linear regressions and clustering as a most-suited-to-the-data solution in between the homoscedastic and the heteroscedastic models.

1. Introduction

Let $\{(y_i, \mathbf{x}_i)\}_n = \{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$ be a sample of n independent units, where y_i is the outcome variable and $\mathbf{x}_i$ are the J covariates. A clusterwise linear regression model assumes that the density of $y_i|\mathbf{x}_i$ is given by

$$f(y_i|\mathbf{x}_i; \boldsymbol{\psi}) = \sum_{g=1}^{G} p_g f_g(y_i|\mathbf{x}_i; \sigma_g^2, \boldsymbol{\beta}_g) = \sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi\sigma_g{}^2}} \exp\left[-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_g)^2}{2\sigma_g^2} \right], \tag{1}$$

where G is the number of clusters, $\boldsymbol{\psi} = \{(p_1, \ldots, p_G; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G; \sigma_1^2, \ldots, \sigma_G^2) \in \mathbb{R}^{G(J+2)} : p_1 + \cdots + p_G = 1, p_g \geq 0, \sigma_g^2 > 0, g = 1, \ldots, G\}$ is the set of model parameters, and p_g, $\boldsymbol{\beta}_g$, and σ_g^2 are respectively the mixing proportions, the vector of J regression coefficients, and the variance term for the g-th cluster. The model in Equation (1) is also known under the name of finite mixture of linear regression models, or switching regression model [21,22,15].

The parameters of finite mixtures of linear regression models are identified if some mild regularity conditions are met [10].

The clusterwise linear regression model of Equation (1) can naturally serve as a classification model. Based on the model, one computes the posterior membership probabilities for each observation as follows:

☆ This paper is part of the Virtual special issue on Soft methods in probability and statistics, Edited by Barbara Vantaggi, Maria Brigida Ferraro, Paolo Giordani

* Corresponding author.
E-mail addresses: roberto.dimari@unict.it (R. Di Mari), roberto.rocci@uniroma2.it (R. Rocci), gattone@unich.it (S.A. Gattone).

http://dx.doi.org/10.1016/j.ijar.2017.09.006
0888-613X/

$$\mathrm{p}(g|\mathrm{y}_i) = \frac{p_g f_g(\mathrm{y}_i|\mathbf{x}_i; \sigma_g^2, \boldsymbol{\beta}_g)}{\sum_{h=1}^{G} p_h f_h(\mathrm{y}_i|\mathbf{x}_i; \sigma_h^2, \boldsymbol{\beta}_h)}, \quad (2)$$

and then classify each observation according, for instance, to fuzzy or crisp classification rules.

The problem of clustering sample points grouped around linear structures has been receiving a lot of attention in the statistical literature because of its important applications (see, for instance, [16], and references therein. For the robust literature, among the others, see [6,7]).

In order to estimate $\boldsymbol{\psi}$, one has to maximize the following sample likelihood function

$$\mathrm{L}(\boldsymbol{\psi}; \mathbf{y}) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi\sigma_g{}^2}} \exp\left[-\frac{(\mathrm{y}_i - \mathbf{x}_i'\boldsymbol{\beta}_g)^2}{2\sigma_g^2} \right] \right\}, \quad (3)$$

which can be done using iterative procedures like the EM algorithm [5], whose clustering can be interpreted as a fuzzy partition [9]. Unfortunately, maximum likelihood (ML) estimation of univariate unconditional or conditional normals suffers from the well-known issue of unboundedness of the likelihood function: whenever a sample point coincides with the group's centroid and the relative variance approaches zero, the likelihood function increases without bound ([14]; also the multivariate case suffers from the issue of unboundedness. See [4]). Hence a global maximum cannot be found.

Yet, ML estimation does not fail: Kiefer [15] showed that there is a sequence of consistent, asymptotically efficient and normally distributed estimators for switching regressions with different group-specific variances (heteroscedastic switching regressions). These estimators correspond, with probability approaching one, to local maxima in the interior of the parameter space. Nonetheless, although there is a local maximum which is also a consistent root, there is no tool for choosing it among the local maxima. Day [4] showed, for multivariate mixtures of normals, that potentially each sample point – or any pair of sample points being sufficiently close together, or co-planar [24] – can generate a singularity in the likelihood function of a mixture with heteroscedastic components. This gives rise, both in univariate and multivariate contexts, to a number of spurious maximizers [18].

The issue of unboundedness can be dealt with by imposing constraints on the component variances. This approach is based on the seminal work of Hathaway [8], who showed that imposing a lower bound, say c, to the ratios of the scale parameters of univariate mixtures of normals prevents the unboundedness of likelihood function. Although the resulting ML estimator is consistent and the method is equivariant under linear affine transformations of the data – that is, if the data are linearly transformed, the estimated posterior probabilities do not change and the clustering remains unaltered – the proposed constraints are very difficult to apply within iterative procedures like the EM algorithm. Furthermore, the issue of how to choose c, which controls the strength of the constraints, remains open.

For multivariate mixtures of normals, Ingrassia [12] showed that it is sufficient, for Hathaway's constraints to hold, to impose bounds on the eigenvalues of the class conditional covariance matrices. This provides a constrained solution that 1) can be implemented at each iteration of the EM algorithm, and 2) still preserves the monotonicity of the resulting EM [13]. Recently, Rocci et al. [23, RGD] proposed a constrained estimation method for multivariate mixtures of normals, being characterized by 1) fully data-dependent constraints, 2) equivariance of the clustering algorithm under change of scale in the data, and 3) ease of implementation within standard routines [12,13].

The aim of this paper is to extend the RGD constrained estimation method to clusterwise linear regression models. We demonstrate that it works very well when a conditional distribution (linear regression) is specified for each mixture component.

A correct estimation of the regression coefficients is crucial in a regression context, where the focus is not only on the cluster recovery, but on the interpretation of the estimated associations. RGD (2017) looked at how good the method was at recovering clusters: in our simulation study we also bring into focus the regression parameters, and look at the quality of the estimators in terms of mean squared error – which embeds both the bias and the variance of the estimators. In this new perspective, we can now argue, based on the evidence of our simulation studies and empirical examples, that the RGD constrained estimation method yields a final model – in terms of clustering and estimated parameters – in between the fully constrained model and the unconstrained model. How close to which of the extremes is optimally determined by maximizing a suitable objective function.

Starting from Rousseeuw and Leroy [25]'s nomenclature, the equivariance property in linear models is of three types: regression, affine and scale. Regression equivariance holds if, by adding a linear combination of the covariates to the response variable through any column vector, the model parameters are shifted by that same vector. If instead an affine transformation is applied on the covariates, affine equivariance guarantees that the model parameters are transformed accordingly. That is, the linear predictor remains the same. The third type of equivariance refers to scale changes in the response variable, in that the model parameters are rescaled such that the linear predictor and the error's standard deviation are both on the new response scale. Either equivariance property types hold for the unconstrained clusterwise linear regression. Notice that neither affine transformations of the **x**s nor shifts in the response proportional to the covariates affect the error's variance: therefore regression and affine equivariance still hold for the constrained model. Indeed, scale equivariance is no longer guaranteed in the constrained model, as constraints might prevent the error's variance to adapt to the new scale of the response variable.