Computational Statistics and Data Analysis 110 (2017) 87–102

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Mixture models for mixed-type data through a composite likelihood approach

Monia Ranalli *, Roberto Rocci

Department of Statistics, The Pennsylvania State University, USA
Department of Economics and Finance, Tor Vergata University, Rome, Italy

ARTICLE INFO

Article history:
Received 5 February 2016
Received in revised form 17 November 2016
Accepted 31 December 2016
Available online 6 January 2017

Keywords:
Mixture models
Mixed-type data
Composite likelihood
EM algorithm

ABSTRACT

A mixture model is considered to classify continuous and/or ordinal variables. Under this model, both the continuous and the ordinal variables are assumed to follow a heteroscedastic Gaussian mixture model, where, as regards the ordinal variables, it is only partially observed. More specifically, the ordinal variables are assumed to be a discretization of some mixture variables. From a computational point of view, this creates some problems for the maximum likelihood estimation of model parameters. Indeed, the likelihood function involves multidimensional integrals, whose evaluation is computationally demanding as the number of ordinal variables increases. The proposal is to replace this cumbersome likelihood with a surrogate objective function that is easier to maximize. A composite approach is used, in particular the original joint distribution is replaced by the product of three blocks: the marginal distribution of continuous variables, all bivariate marginal distributions of ordinal variables and the marginal distributions given by all continuous variables and only one ordinal variable. This leads to a surrogate function that is the sum of the log contributions for each block. The estimation of model parameters is carried out maximizing the surrogate function within an EM-like algorithm. The effectiveness of the proposal is investigated through a simulation study and two applications to real data.

1. Introduction

Nowadays, modern complex data structure raises new challenges in likelihood-based inference methods, where complex means that the dependence between observed variables is difficult to model. The situation is even more complicated when mixed-type data (continuous and ordinal variables) are present. Furthermore, in many applications, the standard assumption of homogeneity is not reliable. Indeed, in several fields, such as e.g. in economics, social sciences or genomics, the population is composed of a finite number of subpopulations. In other words, there exists a cluster structure underlying the data. The aim of cluster analysis is discovering groups within a set of objects, such that homogeneous clusters differ considerably from each other. The literature on clustering has been mainly developed for continuous data. Clustering ordinal variables is a lively field of research, but the amount of work done is still relatively small. The challenge to model ordinal

* Corresponding author at: Department of Statistics, The Pennsylvania State University, USA.
E-mail address: monia.ranalli@psu.edu (M. Ranalli).

http://dx.doi.org/10.1016/j.csda.2016.12.016
0167-9473/

data is mainly due to the lack of metric properties. For this reason, it is still common to analyze ordinal data following a naive approach whereby their nature is ignored. Ranks are treated as interval-scaled, and thus clustering techniques developed for continuous data are applied. Finite mixture of Gaussians (see e.g. Lindsay, 1995 and McLachlan and Peel, 2000) represents the most used and well-known model-based clustering for continuous data. They have been intensively used in many fields and with different purposes (e.g. unsupervised, semi-supervised and supervised classification). Their success is mainly due to their simplicity to be fitted and interpreted. According to a clustering point of view, they provide a coherent strategy for classifying data accounting for uncertainties through probabilities. On the other hand, the most common model-based clustering for categorical data is latent class analysis (Goodman, 1974) and some constrained versions that have been provided for ordinal data (see e.g. DeSantis et al., 2008). These models are finite mixtures arising from the local independence assumption. They consider the cluster membership as a nominal latent factor and assume that the manifest variables are independent given that factor. Of course, this model is inadequate when there are dependences among the manifest variables within the clusters. Such inadequacy can be solved following the Item Response Theory (IRT) approach. The ordinal variables are assumed to be independent given a set of latent continuous variables. The latter having a clustering structure, for example, they can be distributed as a finite mixture of Gaussians (FMG) (Cagnone and Viroli, 2012; McParland et al., 2014). Another way to overcome the within-independence limitation is to consider a FMG that allows dependences within clusters to be modeled by means of the covariance matrices. Following the Underlying Response Variable (URV) approach used in latent variables models, the FMG can be adapted to ordinal data by assuming that the observed variables are a categorization of underlying non-observable continuous variables distributed as a FMG (see for example Lubke and Neale, 2008; Ranalli and Rocci, 2016a; Ranalli and Rocci, 2015). However, even if it is possible to cluster via a model based approach continuous or ordinal variables separately, combining both into a common framework may raise some issues. Assuming that the data arises from a finite mixture model, the main problem is represented by the choice of the parametric joint distribution for the continuous and ordinal variables. From a practical point of view, it may be easy treating variables as they were all continuous; this means ignoring the nature of ordinal variables and considering the ranks as continuous variables. However, this naive approach, although it could give some useful clues about the cluster structure, it could lead to some biased inferential conclusions.

Following the URV approach, Everitt (1988) and Everitt and Merette (1990) proposed a model according to which both the continuous and the categorical ordinal variables follow a homoscedastic Gaussian mixture model. However, as regards the ordinal variables, the mixture variables are only partially observed through their ordinal counterparts. In other words, the ordinal variables are modeled following the URV approach. This satisfies the two main requirements: dealing with ordinal data properly and modeling dependences between ordinal and continuous variables. It is interesting to note that this model can be rewritten in terms of copulas (Marbac et al., 2014). The main drawback of this model is that, in practice, it cannot be estimated through a full maximum likelihood approach, due to the presence of multidimensional integrals making the estimation time consuming. Typically, the full information maximum likelihood becomes computationally demanding even with a number of ordinal variables very low and infeasible when this number is greater than 5. Here, the proposal is to use a mixture model to classify continuous and ordinal variables. The model considered is a modified version of Everitt (1988) and can be seen as an extension to mixed-type data of the mixture model for ordinal data proposed by Ranalli and Rocci (2016a). The observed ordinal variables are considered as a categorization of an underlying mixture of normals. This means that the whole mixture is not fully observed. As in de Leon and Carrièger (2007), the joint distribution within each component can be decomposed in two factors: the first corresponds to the observed normal distribution for the continuous variables, while the second one to the distribution of the ordinal variables given the continuous ones. The latter involves multidimensional integrals, whose evaluation is computationally demanding as the number of ordinal variables increases. To overcome this issue, we propose to replace this cumbersome likelihood with a surrogate objective function, easier to maximize, that is the product of marginal likelihoods. In this way, regardless the number of variables, the multidimensional integrals are reduced to be univariate or at most bivariate integrals. Our proposals is based on the existing results within a mixture model framework (Ranalli and Rocci, 2016a, 2015). It is a composite likelihood method (Lindsay, 1988; Varin et al., 2011) where surrogate functions are defined as the product of marginal or conditional events. As we show, it is a workable compromise between statistical and computational efficiency. Indeed, the composite likelihood methods are flexible ways to create consistent estimators, which inherit the main desirable properties of the maximum likelihood estimators: asymptotically unbiased and normally distributed with the variance given by the inverse of the Godambe Information (Lindsay, 1988; Varin et al., 2011). Moreover, they have some varying degrees of robustness (Xu and Reid, 2011), they are fully efficient and identical to the full maximum likelihood estimators in exponential families under a certain closure property (Mardia et al., 2009). In general efficiency is not easy to achieve and it is strictly linked to the design issue. Under our proposal, the composite approach consists of replacing the joint likelihood with the product of three blocks of marginals: the marginal distribution of continuous variables, all bivariate marginal distributions of ordinal variables and the marginal distributions given by all continuous variables and only one ordinal variable. This leads to a surrogate function that is the sum of the log contributions for each block. The estimation of model parameters is carried out within an EM-like algorithm. The remainder of the paper is organized as follows. Section 2 introduces the clustering model, describes the estimation procedure and deals with some minor issues (classification and model selection). Section 3 sketches the necessary, but not sufficient, condition to identify the model. Some related models are described in Section 4. A simulation study is presented in Section 5. Two applications to real data have been conducted in Section 6. Finally, concluding remarks are pointed out in Section 7.