ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

A hierarchical modeling approach for clustering probability density functions

Q1 Daniela G. Calò, Angela Montanari, Cinzia Viroli *

Department of Statistics, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy

ARTICLE INFO

Article history:
Received 29 March 2012
Received in revised form 23 April 2013
Accepted 23 April 2013
Available online xxxx

Keywords:
Maximum likelihood
Mixture modeling
Pdf clustering

ABSTRACT

The problem of clustering probability density functions is emerging in different scientific domains. The methods proposed for clustering probability density functions are mainly focused on univariate settings and are based on heuristic clustering solutions. New aspects of the problem associated with the multivariate setting and a model-based perspective are investigated. The novel approach relies on a hierarchical mixture modeling of the data. The method is introduced in the univariate context and then extended to multivariate densities by means of a factorial model performing dimension reduction. Model fitting is carried out using an EM-algorithm. The proposed method is illustrated through simulated experiments and applied to two real data sets in order to compare its performance with alternative clustering strategies.

1. Introduction

The problem of analyzing distributions pertaining to continuous variables (i.e. cumulative distribution functions or probability density functions) may arise in different research domains. To consider only a few examples, we can cite the study of age distributions across different world countries on a given year (Delicado, 2011), the characterization of computer images by the distribution of the respective gray-scale pixel values (Spellman et al., 2005), the comparison among the distributions of collagen fibril diameters observed in different mice strains (Chervoneva et al., 2012), or the idea of performing customer segmentation after each customer has been represented by the distribution of the item unit price across his purchases (Sakurai et al., 2008).

In general, distributions are commonly used in exploring and modeling complex data sets in order to ensure that, after database aggregation, information is preserved. In fact, representing the generic "object" of interest (namely, in the above-mentioned examples: world country, image, mouse, customer, respectively) using a distribution is more informative than using classical representations, such as the average, when characteristics like variability, skewness or modality are of specific interest. This is leading to the emergence of a new methodological setting in data analysis, in which a data point is represented by a whole distribution (see Noirhomme-Fraiture and Brito, 2011).

Within this context, the present work focuses on the problem of clustering a set of J "objects" into homogeneous clusters with respect to the corresponding distribution of a continuous random vector $\mathbf{y} \in R^p$ ($p \geq 1$). More precisely, we suppose that for each object j a set of n_j observations of $\mathbf{y}$ is available, which is taken as a random sample from the corresponding unknown parental probability density function f_j, with $j = 1, \ldots, J$. For this reason, the objects to be clustered will be termed as "pdf-objects" in the following.

The methods that have been proposed in the literature for clustering a set of observed distributions mainly consist in specifying a suitable metric to quantify divergence between two arbitrary distributions and then applying a classical

* Corresponding author. Tel.: +39 051 2098250; fax: +39 051 232153.
E-mail address: cinzia.viroli@unibo.it (C. Viroli).

0167-9473/$ – see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.csda.2013.04.013

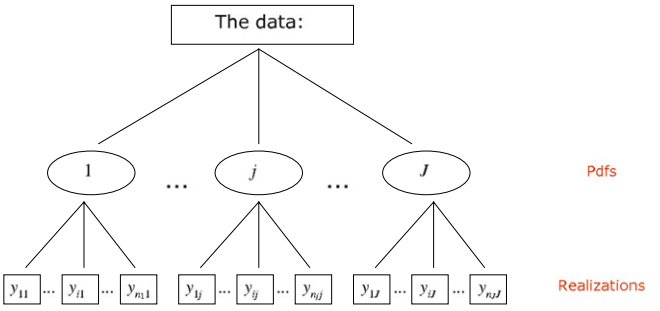

Fig. 1. Hierarchical structure.

hierarchical or partitive clustering method. In this field, Irpino and Verde (2008) have carried out an extensive work, with special attention to histogram data. Alternatively, Terada and Yadohisa (2010) have proposed a non-hierarchical clustering method that works with empirical cumulative distribution functions, in order to avoid that the number of histogram bins (or the range of bins) affects the results of clustering.

These approaches have been developed for both univariate (i.e. $p = 1$) and multivariate settings. However, dealing with multidimensional distributions is more challenging, since a relatively large number of observations is required due to the curse of dimensionality; in addition, the straightforward extension of univariate methods to the multivariate settings involves computationally demanding procedures. A common solution to the latter problem is to assume independence and classify the objects using marginal distributions. Alternatively, Vrac et al. (2011) have proposed a method for clustering a set of estimated multivariate distribution functions based on the use of copula analysis so that the relationship between the observed variables is taken into account.

The present paper introduces a novel approach to the problem of clustering a set of pdf-objects, whose main feature is to rely on a hierarchical mixture modeling of the data rather than on some heuristic clustering procedure. In addition, the solution designed for the case of multivariate densities has the desirable feature of allowing dimension reduction by assuming a generative factorial model for the observed variables.

The rest of the paper is organized as follows. Section 2 introduces the proposed method in the case $p = 1$; an example on a real data set is reported in Section 3 for illustrative purposes. Section 4 presents the extension of the method to $p > 1$. Section 5 reports the EM-algorithm used to fit the proposed model; identifiability conditions are also discussed. The performance of the method on simulated and real data is evaluated in Sections 6 and 7, respectively. A final section contains some concluding remarks.

2. The proposed approach: the univariate setting

Let $\{y_{11}, \ldots, y_{i1}, \ldots, y_{n_1 1}\}, \ldots, \{y_{1j}, \ldots, y_{ij}, \ldots, y_{n_j j}\}, \ldots, \{y_{1J}, \ldots, y_{iJ}, \ldots, y_{n_J J}\}$ denote J samples of variable $y \in R$, each set being randomly drawn from a different parental distribution j (with $j = 1, \ldots, J$), referred to as "pdf-object" j. As shown in Fig. 1, these data have an intrinsic two-level hierarchical structure, in which pdf-objects correspond to *higher-level units* (referred to by the index j), and the random realizations from the different pdfs represent *lower-level units* (those nested within the higher-level unit j are denoted by the index i, ranging from 1 to n_j). If $\mathbf{y}$ denotes a p-dimensional variable, the multilevel framework can still be adopted by interpreting the p values observed on each lower-level unit as nested univariate observations; this amounts to adding an extra level of nesting at the bottom of the hierarchy. In general, under the multilevel perspective, the problem of clustering a set of univariate or multivariate pdfs can be naturally recast as a problem of clustering higher-level units.

Among the models for describing multilevel data (see, for example, Skrondal and Rabe-Hesketh, 2004), those based on finite mixture modeling (McLachlan and Peel, 2000) are particularly appealing for the purpose of partitioning the set of higher-level units (see, for example, Vermunt and Magidson, 2005 and Vermunt, 2007). This is due to the fact that finite mixtures provide both a sound statistical framework for cluster analysis (where each component in the mixture is assumed to correspond to a cluster) and as a semi-parametric tool for estimating unknown distributional shapes (provided that the number of components is taken sufficiently large to yield an accurate estimate), as an alternative to nonparametric density estimation (see, for instance, Golyandina et al., 2012 and references therein). This is particularly attractive for our purpose of building a model for the hierarchical data depicted in Fig. 1, as clustering and density estimation are the two main issues pertaining to the higher and the lower level of the hierarchy, respectively.

Therefore, we propose to assume two separate finite mixture models simultaneously, each one corresponding to a distinct level of the hierarchy, as described in the following. For notational convenience, we present the proposed model by introducing latent multinomial random variables s and r to indicate component membership for higher- and lower-level units, respectively.

– At the higher level, sample $\mathbf{y}_j = [y_{ij}]_{i=1}^{n_j}$ from the j-th pdf-object is assumed to be drawn from one of L different latent classes or sub-populations (L being fixed but unknown), with respective unknown prior probabilities w_l, $l = 1, \ldots, L$,

Please cite this article in press as: Calò, D.G., et al., A hierarchical modeling approach for clustering probability density functions. Computational Statistics and Data Analysis (2013), http://dx.doi.org/10.1016/j.csda.2013.04.013