Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 659–675
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Latent class models for mixed variables with applications in Archaeometry

Irini Moustaki*, Ioulia Papageorgiou

Department of Statistics, Athens University of Economics and Business, 76 Patission Street, Athens 104 34, Greece

Received 17 February 2004; received in revised form 3 March 2004; accepted 3 March 2004

Abstract

Latent class models are used in social sciences for classifying individuals or objects into distinct groups/classes based on responses to a set of observed indicators. The latent class model for mixed binary and metric variables (Br. J. Math. Statist. Psych. 49 (1996) 313) is extended to accommodate any type of data (including ordinal and nominal) and its use in Archaeometry for classifying archaeological findings/objects into groups is discussed. The models proposed are estimated using a full maximum like-lihood with the EM algorithm. Two data sets from archaeological findings are used to illustrate the methodology.

Keywords: Latent class models; Nominal, ordinal, metric and mixed variables; EM algorithm; Archaeometry

1. Introduction

One of the main problems in Archaeology is classification of objects found in excavations such as ceramic sherds, artefacts, etc. The criterion for grouping in the context we are interested in is the origin of the objects. Provenance is an important issue for Archaeology researchers, as this is a first step to derive conclusions about the structure of the communities in ancient years. Conclusions are drawn with respect to civilization, level of the manufacturing techniques used, and also import–export of goods, ability of move and relations between them.

The most widely used classification techniques by Archaeometry scientists is hierarchical clustering. This class of procedures involves choosing a measure of (dis)similarity

* Corresponding author. Tel.: +302108203543; fax: +302108203113.
E-mail addresses: moustaki@aueb.gr (I. Moustaki), ioulia@aueb.gr (I. Papageorgiou).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.03.001

between pairs of cases in a sample to be clustered, and choosing an algorithm for clustering cases hierarchically on the basis of the (dis)similarity coefficient. Both choices can be made in many different ways, leading to a large number of possible ways of clustering a data set.

Such approaches are essentially heuristic, and have been contrasted with model-based approaches to clustering (see Fraley and Raftery, 1999). Fraley and Raftery (1999) use a mixture of multivariate normals and that can be considered a special case of the latent class model for mixed variables that will be developed here. Heuristic methods dominate archaeometric practice. Papageorgiou et al. (2001) and Baxter (2001) have compared some approaches to grouping data used in archaeometry that are model-based. A model-based method is understood to be one in which explicit assumptions are made about the form of the probability density function describing the population from which the observed data are considered to be a random sample. Clustering and inferences about the numbers of clusters and cluster membership are based on estimation of the unknown parameters in the probability model used.

One could summarize the potential merits of model-based methodologies in contrast to distribution-free methodologies.

(1) Cases are assigned to clusters based on probabilities estimated from a model. Within that process outliers can be identified. Initial assignment to a cluster can be based on archaeological rather than statistical grounds and model-based statistical methods may then be used to assess whether or not such a group is also chemically coherent. A variant of this (Glascock, 1992) is to determine initial groups statistically using an heuristic method, and then to 'refine' these using probabilistic calculations that assume the groups are multivariate normal.

(2) In many compositional studies variables may be highly correlated within groups leading to clusters that are ellipsoidal in p-dimensional space. Heuristic clustering methods typically impose spherical structure on the data and can fail to recognize the true structure. One common method of cluster analysis, Ward's method, often used in an heuristic manner, can be shown to be a special case of a model-based method that not only assumes that clusters are spherical but also of equal size (volume). This difficulty is well-known but resolving it is not easy (Harbottle, 1976). Krzanowski and Marriott (1995) observe that 'most methods not specifically distribution-based are inefficient at finding strongly elliptical clusters'. In principle, therefore, distribution- or model-based methods provide a way of addressing a problem that has been an issue ever since multivariate methods began to be applied to the analysis of compositional data.

(3) The output from many heuristic methods of cluster analysis is typically presented in the form of a dendrogram. Judgements about the number of clusters are usually made on the basis of subjective interpretation of the dendrogram. Apart from the subjectivity involved there are several difficulties here. The appearance of a dendrogram is affected by the scale of the data, choice of (dis)similarity measure and clustering algorithm used, and is often not easy to interpret. Furthermore, a clear separation into distinct clusters on a dendrogram does not guarantee that they are genuinely distinct (Baxter, 1994, p. 161). A potential advantage of model-based