One-Mode Classification of a Three-Way Data Matrix

Maurizio Vichi
University of Chieti

Abstract: One-mode classification (OMC) of a three-way data set (matrix) $X$ is the automatic hierarchical classification of one mode (units or variables or occasions) of $X$ on the basis of the other two. In this paper the case of OMC of units according to variables and occasions is discussed. OMC is the synthesis of a set of hierarchical classifications $\Delta$ obtained from $X$; e.g., the OMC of units is the consensus (synthesis) among the set of dendrograms individually defined by clustering units on the basis of variables, separately for each given occasion of $X$. However, because $\Delta$ is often formed by a large number of classifications, it may be unrealistic that a single synthesis is representative of the entire set. In this case, subsets of similar (homogeneous) dendrograms may be found in $\Delta$ so that a consensus representative of each subset can be identified. This paper proposes, PARtition and Least Squares Consensus cLassifications Analysis (PARLSCA) of a set of $r$ hierarchical classifications $\Delta$. PARLSCA identifies the best least-squares partition of $\Delta$ into $m$ ($1 \leq m \leq r$) subsets of homogeneous dendrograms and simultaneously detects the closest consensus classification (a median classification called Least Squares Consensus Dendrogram (LSCD)) for each subset. PARLSCA is a generalization of the problem to find a least-squares consensus dendrogram for $\Delta$. PARLSCA is formalized as a mixed-integer programming problem and solved with an iterative, two-step algorithm. The method proposed is applied to an empirical data set.

Keywords: Three-way data set; Cluster analysis; Consensus classification.

---

I wish to thank Allan D. Gordon, Phipps Arabie and the anonymous referees for their helpful suggestions and comments that greatly improved the quality of this paper.

Author’s address: Maurizio Vichi, University of Chieti, Department of Metodi Quantitativi e Teoria Economica, Viale Pindaro 42, 65127 Pescara, Italy; e-mail: Vichi@dmqte.unich.it
1. Introduction

When a multivariate phenomenon is observed on different occasions, elements of the data set collected are identified according to three modes: units (rows), variables (columns), and occasions (layers), and then arranged into a three-way array (matrix) $X = [x_{ijk}]$. In this paper occasions are different times when the data are collected, and units remain the same over time, defining a three-way longitudinal data set. Large data structures of this kind can be difficult to comprehend, and methods of synthesizing and extracting relevant information are necessary. It is often useful to classify elements of one mode hierarchically on the basis of the other two. This approach is referred to as One-Mode Classification (OMC) of a three-way data set. Three types of OMC can be considered: (a) OMC of units according to variables and occasions, (b) OMC of variables according to units and occasions, and (c) OMC of occasions on the basis of units and variables. OMC may be seen as a consensus (synthesis) of a set of hierarchical classifications\(^1\) $\Delta$ obtained by the three-way data set. For example, OMC of units of $X$ is a synthesis of the set of hierarchical classifications obtained by clustering units of $X$ according to variables and repeating this classification separately for each distinct time period (i.e., by clustering units starting from frontal slices of $X$). This paper will emphasize (a) the OMC of units.

A first and simple way to achieve OMC of units is to rearrange adjacent layers of the three-way matrix into a large pooled two-way matrix and to analyze it with a hierarchical technique. However, the relevant information on the dynamics of units over different times is thus lost. Furthermore, such an analysis cannot reveal if units remain stable in the same group over different times or even whether and how many times units change cluster membership.

Less information is lost when more than one consensus classification is used to summarize $\Delta$. For this reason classification techniques for three-way data sets should be able to determine more than one consensus classification, each one summarizing one of the different classification structures present in the three-way data set.

When one of the modes is time, one might expect that units do not change cluster membership at two close times and that dendrograms maintain the same topological structure ($n$-tree), with possibly different linkage lengths between cluster levels. These dendrograms are called "topologically

\(^1\) The terms hierarchical classification, dendrogram, and ultrametric matrix will be used interchangeably throughout the paper.