Cluster Analysis with General Latent Class Model

Dingxi Qiu
University of Miami, USA

Edward C. Malthouse
Northwestern University, USA

INTRODUCTION

Cluster analysis is a set of statistical models and algorithms that attempt to find "natural groupings" of sampling units (e.g., customers, survey respondents, plant or animal species) based on measurements. The observable measurements are sometimes called *manifest* variables and cluster membership is called a *latent* variable. It is assumed that each sampling unit comes from one of K clusters or classes, but the cluster identifier cannot be observed directly and can only be inferred from the manifest variables. See Bartholomew and Knott (1999) and Everitt, Landau and Leese (2001) for a broader survey of existing methods for cluster analysis.

Many applications in science, engineering, social science, and industry require grouping observations into "types." Identifying typologies is challenging, especially when the responses (manifest variables) are categorical. The classical approach to cluster analysis on those data is to apply the latent class analysis (LCA) methodology, where the manifest variables are assumed to be independent conditional on the cluster identity. For example, Aitkin, Anderson and Hinde (1981) classified 468 teachers into clusters according to their binary responses to 38 teaching style questions. This basic assumption in classical LCA is often violated and seems to have been made out of convenience rather than it being reasonable for a wide range of situations. For example, in the teaching styles study two questions are "Do you usually allow your pupils to move around the classroom?" and "Do you usually allow your pupils to talk to one another?" These questions are mostly likely correlated even within a class.

BACKGROUND

This chapter focuses on the mixture-model approach to clustering. A mixture model represents a distribution composed of a mixture of component distributions, where each component distribution represents a different cluster. Classical LCA is a special case of the mixture model method. We fit probability models to each cluster (assuming a certain fixed number of clusters) by taking into account correlations among the manifest variables. Since the true cluster memberships of the subjects are unknown, an iterative estimation procedure applicable to missing data is often required.

The classical LCA approach is attractive because of the simplicity of parameter estimation procedures. We can, however, exploit the correlation information between manifest variables to achieve improved clustering. Magidson and Vermunt (2001) proposed the latent class factor model where multiple latent variables are used to explain associations between manifest variables (Hagenaars, 1988; Magidson & Vermunt, 2001; Hagenaars & McCutcheon, 2007). We will, instead, focus on generalizing the component distribution in the mixture model method.

MAIN FOCUS

Assume a random sample of n observations, where each comes from one of K unobserved classes. Random variable $Y \in \{1, \ldots, K\}$ is the latent variable, specifying the value of class membership. Let $P(Y=k) = \eta_k$ specify the *prior distribution* of class membership, where

$$\sum_{k=1}^{K} \eta_k = 1.$$

For each observation $i = 1, \ldots, n$, the researcher observes p manifest variables $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})'$. Given that an observation comes from class k (i.e., $Y = k$), the *class-conditional distribution* of $\boldsymbol{X}$, denoted as $f_k(\boldsymbol{x};\theta_k)$, is generally assumed to come from common distribution families. For example, classical LCA assumes that the components of $\boldsymbol{X}$ are each multinomial and independent of each other for objects within the same class. Suppose each manifest variable takes only 2 values, hereafter labeled generically "yes" and "no", then $P(X_j = x_j \mid Y)$ is a Bernoulli trial. Let π_{jk} be the probability that someone in class k has a "yes" value to manifest variable X_j. Then the class-conditional distribution, under the assumption of class-conditional independence, is

$$f_k(\boldsymbol{x};\theta_k) = \prod_{j=1}^{p} P(x_j \mid Y=k) = \prod_{j=1}^{p} \pi_{jk}^{x_j} (1-\pi_{jk})^{1-x_j}. \quad (1)$$

This assumption greatly reduces the number of parameters that must be estimated. However, in many cases, more flexible distributions should be developed to allow for improved clustering.

Component Distributions

In general, $f_k(\boldsymbol{x};\theta_k)$ can take any component distribution. However, due to the constraint of identifiableness and the computing requirement for parameter estimation, only two component distributions, to our best knowledge, have been proposed in the literature to address the correlation structure within each cluster. Qu, Tan and Kutner (1996) proposed a random effects model that is a restricted version of the multivariate Probit model. The conditional dependence is modeled by subject-specific random variables. The manifest variables are correlated because of the correlations between the underlying normal random variables in addition to the class membership. The correlation matrix in the component distribution of the random effects model has a restricted structure which makes it less appealing (Tamhane, Qiu & Ankenman, 2006).

Tamhane, Qiu and Ankenman (2006) provide another general-purpose multivariate Bernoulli distribution, called the continuous latent variable (CLV) model, based on subject-specific uniform random variables. This proposed distribution can handle both positive and negative correlations for each component cluster. It is relatively flexible in the sense that the correlation matrix does not have any structural restrictions as in the random effects model. This approach has been applied to two real data sets (Tamhane, Qiu & Ankenman, 2006) and provided easily interpretable results.

Parameter Estimation

The model parameters (η_k and $\boldsymbol{\theta}_k$) are estimated with maximum likelihood. The probability density function of the mixture is

$$f(\boldsymbol{x};\psi) = \sum_{k=1}^{K} \eta_k f_k(\boldsymbol{x};\theta_k),$$

where the vector ψ of unknown parameters consists of the mixture proportions η_k and class-conditional distribution parameters θ_k. Under the assumption that $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are independent observations, the incomplete log-likelihood function is given by

$$\log L = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \eta_k f_k(\boldsymbol{x}_i, \theta_k),$$

which must be maximized with respect to parameters η_k and $\boldsymbol{\theta}_k$. Due to the summation inside the logarithm, direct maximization is difficult, and the expectation-maximization (EM) algorithm of Dempster, Laird and Rubin (1977) is generally used to obtain the parameter estimators. See Bartholomew and Knott (1999, pp. 137-139) for details. The EM algorithm is convenient to construct if there exist closed-form solutions to the maximum likelihood estimators (MLEs). When closed-form solutions do not exist, the more general optimization procedures, such as quasi-Newton method, will be used. Generally speaking, there are no known ways of finding starting values that guarantee a global optimum, and different starting values will often produce different local maxima. One solution to the starting-value problem is to run the optimization with multiple random starting values and select the one with the largest log-likelihood value. Commercial software package such as Knitro® and LatentGold® solve this type of optimization problem. There are also software packages in the public domain.[a]