ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Variable assessment in latent class models[☆]

Q. Zhang[*], E.H. Ip

Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston Salem, NC, USA

HIGHLIGHTS

- Two measures are proposed for assessing continuous and discrete variables in LCA.
- Both measures are either in closed form or straightforward to compute.
- Both measures perform reasonably well compared to existing measures such LRT and F_{st}.
- Both absolute and relative interpretations of one measure are possible.

ARTICLE INFO

Article history:
Received 24 July 2013
Received in revised form 19 February 2014
Accepted 20 February 2014
Available online xxxx

Keywords:
Latent class analysis
Variable selection
Mixed data type
Total variation
Posterior gradient
Cross entropy
Kolmogorov distance

ABSTRACT

The latent class model provides an important platform for jointly modeling mixed-mode data—i.e., discrete and continuous data with various parametric distributions. Multiple mixed-mode variables are used to cluster subjects into latent classes. While the mixed-mode latent class analysis is a powerful tool for statisticians, few studies are focused on assessing the contribution of mixed-mode variables in discriminating latent classes. Novel measures are derived for assessing both absolute and relative impacts of mixed-mode variables in latent class analysis. Specifically, the expected posterior gradient and the Kolmogorov variation of the posterior distribution, as well as related properties are studied. Numerical results are presented to illustrate the measures.

1. Introduction

Heterogeneous data types have become commonplace in many sciences. In the medical sciences, clinical studies often collect data that are continuous (e.g., blood pressure), binary (whether or not the subject has diabetes), ordinal (severity level of a disease), categorical (medication used), and other types such as count and time-to-event. The identification of clinically meaningful phenotypes in the population using a heterogeneous data type is thus an important area of research. Everitt (1988, 1993) referred to heterogeneous data types as mixed-mode data in the context of latent class and mixture analysis, in which multiple data types are used as indicators for putting similar objects into groups (see also Lawrence and Krzanowski, 1996; Vermunt and Magidson, 2002). The terms latent class model and mixture are used interchangeably. The idea here is to cluster a vector of mixed-mode responses $\mathbf{Y} = (Y_i)$ for indicators $i = 1, \ldots, m$ into S distinct latent classes $Z = 1, \ldots, S$. There are at least two general approaches for mixed-mode latent class analysis (MM-LCA). The first approach is to relate the manifested categorical response to an underlying multivariate Gaussian distribution such that continuous

[☆] Some mathematical derivation results are included as annexes in the electronic version of this article.

[*] Correspondence to: Department of Biostatistical Sciences, Wake Forest School of Medicine, Medical Center Blvd, Winston Salem, NC 27157, USA. Tel.: +1 336 716 5141; fax: +1 336 716 6427.

E-mail addresses: qizhang@wfubmc.edu, qizhang@wakehealth.edu (Q. Zhang), eip@wakehealth.edu (E.H. Ip).

http://dx.doi.org/10.1016/j.csda.2014.02.017
0167-9473/© 2014 Elsevier B.V. All rights reserved.

Please cite this article in press as: Zhang, Q., Ip, E.H., Variable assessment in latent class models. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.02.017

normal variables and categorical variables can be jointly modeled (Joreskog, 1973; Shi and Lee, 2000). As pointed out by Dunson (2003), the underlying Gaussian approach has limitations, one of which is that it cannot easily accommodate general data types such as counts. An alternative is to use the generalized linear mixed-model approach proposed by Sammel et al. (1997) and later extended by Moustaki and Knott (2000), Dunson (2003), Daniels and Normand (2006), Yang and Dunson (2010), and Cai et al. (2011). This approach can accommodate any mixture of outcomes from an exponential family. Under the assumption of conditional independence given latent class Z, the likelihood for an individual subject in an MM-LCA can be expressed as:

$$f(\mathbf{Y}|\theta) = \sum_{z=1}^{S} \alpha_z \prod_{i=1}^{m} p_{iz}(y_i|\theta_z), \tag{1}$$

where θ contains the vector of parameters θ_z for each individual class z, which has a prior probability $\alpha_z = p(Z = z)$. Within an exponential family framework, different link functions can be specified for the conditional distribution p_{iz} for different data types.

One question that arises from the generalized mixed-model approach for latent class analysis and latent variable in general is how the different types of data "impact" the likelihood. It is possible that one data type "overwhelms" another data type in the likelihood and becomes dominant in defining the structure of the latent class model. Because data values are not measured on the same scale, it is not easy to promptly assess the impact of a variable on the overall likelihood. This question is directly related to a second question: if only a limited number of mixed-mode indicators can be included in a latent class analysis, which variables should be selected for maximally "discriminating" between the classes? Interestingly, the latter question can also be reformulated as a variable-selection problem and solved by a search algorithm using criteria such as the BIC (Raftery and Dean, 2006; Dean and Raftery, 2010).

Two measures are proposed for assessing a variable's contribution to the classification of latent classes. In LCA class labels are not known a priori; the term classification here refers to the extent to which a variable contributes to discriminating the classes, or in the case the class label is known (e.g., in a simulation setting) the accuracy in retrieving class membership. The first measure, the expected posterior gradient (EPG), measures the absolute contribution of a variable to MM-LCA. The second measure, based on the Kolmogorov variation of the posterior distribution (KVP), can be interpreted in terms of the relative contribution of a variable by comparing classification accuracies with and without the variable in the MM-LCA. Interestingly, both measures can be related to the statistical distance between the prior distribution $p(z)$ and the posterior distribution $p(z|y)$. There are several advantages in using the EPG and KVP. First, they both have strong theoretical foundations, which will be described in the following two sections under the heading "Justification of measures". Second, the measures can be universally applied to all kinds of mixed-mode data—continuous, discrete, and count data. Furthermore, computationally the two measures are straightforward to compute and closed form solutions are available for EPG. For the remainder of the paper, Section 2 describes the EPG measure, and the procedure of how the measure can be derived and used in practice. Section 3 describes KVP and specifically its relation to the total variation measure, which is commonly used in the image processing literature. In Section 4, two numerical examples of MM-LCA are provided to illustrate the proposed methods. A brief discussion is given in Section 5.

2. Expected posterior gradient for variable assessment

Consider an S-class latent class model that includes both continuous and discrete random variables, $Y = (Y_1, \ldots, Y_m)$, with class-conditional distributions of normal, exponential, Gamma, Poisson, ordinal, or binomial distributions, given the latent random variable $Z \in \mathcal{S},\ \mathcal{S} = \{1, \ldots, S\}$. Class-conditional independence is assumed among all the variables—i.e.,

$$p(Y_1 = y_1, \ldots, Y_m = y_m|Z = z) = \prod_{i=1}^{m} p(Y_i = y_i|Z = z). \tag{2}$$

Denote the posterior probability $p(Z = z|Y_1 = y_1, \ldots, Y_m = y_m)$ by τ_z, the class-conditional probability $p(Y_i = y_i|Z = z)$ by $\pi_{y_i|z}$, and let $\pi_{y|z} = \prod_{i=1}^{m} \pi_{y_i|z}$. These quantities are related by the Bayes formula:

$$\tau_z = \frac{\alpha_z \pi_{y|z}}{p(y)}, \tag{3}$$

where $p(y)$ is the marginal probability of observing the outcome vector, $y = (y_1, \ldots, y_m)$, and

$$p(y) = \sum_{z=1}^{S} \alpha_z \pi_{y|z}. \tag{4}$$

The EPG measure for assessing the impact of variable y_i on the MM-LCA is denoted by B_i and its definition is given by:

$$B_i = \sum_{z \in \mathcal{S}} \alpha_z \left| E_y \left(\frac{\partial \log(\tau_z)}{\partial y_i} \right) \right|. \tag{5}$$