Statistical Methodology 32 (2016) 107–121

Contents lists available at ScienceDirect

Statistical Methodology

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

# Latent class analysis of incomplete data via an entropy-based criterion

Chantal Larose [a], Ofer Harel [b,*], Katarzyna Kordas [c], Dipak K. Dey [b]

[a] School of Business, State University of New York at New Paltz, New Paltz, NY, USA
[b] Department of Statistics, University of Connecticut, Storrs, CT, USA
[c] University of Bristol, Senate House, Tyndall Avenue, Bristol, BS8 1TH, UK

ARTICLE INFO

*Article history:*
Received 23 July 2015
Received in revised form
5 April 2016
Accepted 27 April 2016
Available online 10 May 2016

*Keywords:*
Entropy
Latent class analysis
Missing data
Model selection
Multiple imputation

ABSTRACT

Latent class analysis is used to group categorical data into classes via a probability model. Model selection criteria then judge how well the model fits the data. When addressing incomplete data, the current methodology restricts the imputation to a single, pre-specified number of classes. We seek to develop an entropy-based model selection criterion that does not restrict the imputation to one number of clusters. Simulations show the new criterion performing well against the current standards of AIC and BIC, while a family studies application demonstrates how the criterion provides more detailed and useful results than AIC and BIC.



## 1. Introduction

Latent class analysis (LCA) [16] is a model-based clustering methodology for categorical data. Variables in a data set are sometimes called "manifest" variables, while the unknown vector of class membership is the "latent" variable. LCA breaks the data into classes (e.g., clusters) via two parameters: latent class probabilities and conditional probabilities. The former dictates how likely it is that a record belongs to each class, while the latter describes the probability of a particular variable having a particular value given that it is in a certain class. LCA assumes that the relationships between manifest variables are accounted for by their class membership. Thus, conditioning on class membership makes manifest variables independent.

* Corresponding author.
*E-mail address:* ofer.harel@uconn.edu (O. Harel).

http://dx.doi.org/10.1016/j.stamet.2016.04.004
1572-3127/

Our goal is to develop a new model selection criterion in order to utilize methods for clustering incomplete categorical data using MI without having to limit ourselves to a single number of clusters. To do so, we first prove that the entropy of an LCA model decreases to zero as the number of classes increases to the number of unique records. We then use this knowledge to construct our criterion.

There are methods for clustering categorical data using entropy [4,25], but they do not address incomplete data. There is also an entropy-based criterion for mixture model data, but it was applied to complete, normal mixture model data [7]. There are also ways to cluster incomplete categorical data using multiple imputation and latent class analysis (LCA) [17] using the fraction of missing information (FMI) as an LCA model selection criterion [18]. However, the methodology sets a fixed number of classes prior to imputation. We propose a new, entropy-based model selection criterion method for the case where the manifest variables are incomplete and class membership is unknown, which allows the use of LCA with multiple imputation without having to set a number of clusters beforehand.

Entropy [7,34,10] is a way to measure the variability, or chaos, in a stochastic system. Entropy depends on the probability density or mass function of the variables or model. One can calculate entropy of a mixture model, and thus entropy has been used as a model selection criterion in clustering scenarios. Typically, it is combined with the log-likelihood [6,5], although it has been used on its own [25,4]. Between two competing model based cluster solutions, the one with lower entropy means there is less variability within each cluster, and thus the clusters are more homogeneous. We are interested in looking at entropy itself as a model selection criterion.

Model selection in clustering also chooses the number of clusters. Existing model selection criteria have penalties to avoid choosing too many clusters, and overfitting the data. To determine what sort of penalty to introduce to our new model selection criterion, we need to understand how entropy of an LCA model behaves as the number of classes increases. Thus, we prove that the entropy of an LCA model with $G$ classes goes to zero as $G$ goes to the number of unique records in the data. Fruhwirth-Schnatter [15] describes entropy of a mixture model as equaling zero if each record belongs to its cluster with probability one. Realistically, this is not likely to happen unless every record has its own cluster. We are unaware of a proof that shows entropy equaling zero when the number of classes approaches the number of unique records. Therefore, we begin by providing such a proof.

Using a number of classes equal to the number of unique records is akin to over-fitting the model; it tells you almost nothing about the grouping patterns in your data. Since we seek to build an entropy-based model selection criterion, we introduce a penalty function, aimed at choosing the best number of classes before encountering the tailing-off effect in entropy, which occurs as more and more unnecessary classes are used.

Moreover, we are interested in the performance of an entropy-based criterion as a model selection tool after multiple imputation has been implemented. BIC or AIC are often used to choose a model, though they do not take into account the need for a well-separated cluster solution [15]. In addition, the performance of BIC and AIC breaks down after multiply imputing data sets in a regression context [8]. This leaves the field open for a new model selection criterion. Therefore, we set out to build an entropy-based model selection criterion which outperforms BIC and AIC after multiple imputation, while considering more than one number of classes at a time.

The paper is organized as follows. Sections 2–4 discuss latent class analysis and model selection, entropy, and missing data respectively. Section 5 details LCA entropy, and showcases our proof that the entropy of an LCA model goes to zero as the number of classes approaches the number of possible unique records. Section 6 describes the methodology of Harel et al. [17], and how we propose to extend the methodology. Section 7 contains the simulation and data application studies. Section 7.1 presents our simulation study, in which we compare our entropy-based criterion to AIC and BIC. Section 7.2 demonstrates an application of our entropy-based criterion, and compared the results to those obtained by AIC and BIC. Section 8 wraps up the paper with our conclusions and directions for future work.

## 2. Latent class analysis

Clustering is the categorization of records into bunches (e.g., clusters) in order to describe grouping patterns in the data set. Latent Class Analysis (LCA) is a model-based clustering method which