Computational Statistics and Data Analysis 93 (2016) 146–161

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Latent profile analysis with nonnormal mixtures: A Monte Carlo examination of model selection using fit indices

Grant B. Morgan [a,*], Kari J. Hodge [a,1], Aaron R. Baggett [b,1]

[a] Department of Educational Psychology, Baylor University, One Bear Place #97301, Waco, TX, 76798-7301, USA

[b] Department of Psychology, University of Mary Hardin-Baylor, Box 8014, Belton, TX, 76513-8014, USA

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form 27 February 2015
Accepted 28 February 2015
Available online 10 March 2015

Keywords:
Mixture model
Model selection
Nonnormal data

ABSTRACT

The performances of fit indices used for model selection in cross-sectional mixture modeling with nonnormally distributed indicators were examined in two studies using Monte Carlo methods. Simulation conditions were selected to mirror conditions found in educational and psychological research. The design factors under investigation were: indicator distribution, number of indicators, sample size, and profile prevalence. All models contained five, ten, or 15 continuous indicators with varying departures from normality. The fit indices examined were Akaike's information criterion (AIC), corrected Akaike's information criterion (AICc), consistent Akaike's information criterion (CAIC), Bayesian information criterion (BIC), sample size-adjusted Bayesian information criterion (SSBIC), Draper's information criterion (DIC), integrated classification likelihood criterion with Bayesian-type approximation (ICL), entropy, and the adjusted Lo–Mendell–Rubin likelihood ratio test (LMR). In the first study, nonnormally distributed data were used to estimate the mixture models. No fit index uniformly identified the simulated number of profiles using nonnormal indicators. The fit indices that tended to identify the simulated number of profiles more frequently than others were BIC, SSBIC, CAIC, and LMR although the condition(s) in which this was observed varied. In the second study, the raw data were transformed using van der Waerden quantile normal scores. Despite deflating the indicator variances, the use of normal scores increased the frequency with which fit indices identified the simulated number of profiles across most conditions.

1. Introduction

Classification procedures have been used for decades by researchers interested in classifying individual cases of a heterogeneous dataset into homogeneous groups. During this time, classification methods have been applied in many disciplines, such as business, education, medicine, and the social sciences. Generally, classification refers to the process of dividing a large, heterogeneous set of observations into smaller, homogeneous groups with smaller within-group variability and greater between-group variability (Clogg, 1995; Gordon, 1981; Heinen, 1996; Muthén and Muthén, 2000). The primary challenge facing researchers is that the frequency and form of the groups underlying a complex dataset is rarely known in advance. The frequency of the groups refers to the number and size of each group, and the form refers to the group-specific

* Corresponding author. Tel.: +1 254 710 7231; fax: +1 254 710 3265.
E-mail addresses: grant_morgan@baylor.edu (G.B. Morgan), kari_hodge@baylor.edu (K.J. Hodge), abaggett@umhb.edu (A.R. Baggett).

[1] There is a supplementary material comprising the tables that contain the frequency with which each fit index identified the competing component models and a sample of the M*plus* and SAS code.

http://dx.doi.org/10.1016/j.csda.2015.02.019
0167-9473/

means, proportions, variances, and/or covariances. Both distance- and model-based classification approaches have been applied by researchers in their efforts to meaningfully structure the individual cases because the purpose of both approaches is to correctly classify similar cases into one of K subgroups.

Mixture modeling generally refers to a model-based approach that is often used to identify underlying subgroups (may also be referred to as classes or profiles depending on the analysis) whose members tend to have more similar values on the manifest variables than with members of other subgroups. The purpose of mixture modeling is often the same as other, distance-based clustering methods, but mixture models treats the underlying class variable as a categorical latent variable. As such, class membership must be measured indirectly using two or more observed, or indicator, variables, which are subject to measurement error.

There are a number of major benefits of mixture methods over distance-based clustering methods. First, mixture models can easily accommodates variables measured on different scales (i.e., mixed metric data). Morgan (2015) showed using Monte Carlo methods that statistical fit indices were effective under many conditions at recovering the true number of classes using a combination of dichotomous and continuous class indicators. Second, mixture modeling approaches recognize that there may be some uncertainty associated with the classification of each case. That is, each vector of observations, $\mathbf{y_i}$, is assigned to group k based on the estimated posterior probability ($\hat{p}_{ik}$). Letting $\hat{\Phi}$ represent the maximum likelihood estimates of the mixture of profile-specific joint distributions of indicators covariance matrices, $\hat{\pi}_k$ represent the estimated profile prevalence, and $\hat{\theta}_k$ represent the profile-specific means, variances, and covariances, the posterior probabilities can be defined as:

$$\hat{p}_{ik} = \Pr(\text{individual } i \in \text{group } k|\mathbf{y_i}; \hat{\Phi}) = \frac{\hat{\pi}_k f_k(\mathbf{y_i}|\hat{\boldsymbol{\theta}}_\mathbf{k})}{\sum_{k=1}^{K} \hat{\pi}_k f_k(\mathbf{y_i}|\hat{\boldsymbol{\theta}}_\mathbf{k})}, \tag{1}$$

for $k = 1, \ldots, K$. Next, $\mathbf{y}_i$ is assigned to group k if

$$\hat{\pi}_{ik} > \hat{\pi}_{ik'}, \tag{2}$$

for $k = 1, \ldots, K$, where $k \neq k'$ (Hunt and Jorgensen, 2003).

The third major benefit of mixture modeling is the flexibility it offers for model estimation. The researcher has the option to freely estimate or constrain any of the model parameters though most restrictions are concerned with elements of the covariance matrix (Vermunt, 2004). A fourth benefit is the availability of indices of model-data fit. A number of studies have investigated fit index performance, but the conditions studied to this point may not generalize to some of the conditions that some researchers are likely to encounter. Thus, model selection through statistical criteria can be viewed as an unresolved issue in mixture modeling. There are many fit indices available in mixture modeling, and each fit index provides slightly different information regarding the model-data fit. The fit indices examined are discussed below.

Procedures that may be included under a mixture modeling umbrella include mixture likelihood approach to clustering (McLachlan and Basford, 1988; Everitt, 1993), model-based clustering (Banfield and Raftery, 1993), finite mixture modeling (McLachlan and Peel, 2000), and latent variable mixture modeling (Henson et al., 2007; Bartolucci et al., 2013). More recently, Bauer and Curran (2004) presented structural equation mixture modeling as integrative framework that may accommodate both categorical and continuous latent variable models. Mixture analysis based on categorical indicators are commonly referred to as latent class analysis, and analysis that employs continuous indicators is commonly referred to as latent profile analysis.

Bauer and Curran (2004) provided an excellent discussion of relationships between popular latent variable models that primarily rely on categorical and/or continuous data. For example, they noted the analytic similarity of latent profile models and common factor models for the first and second order moments with regard to the decomposition of the covariance matrix. They also provided a conceptual and analytic comparison between finite normal mixture modeling and latent profile models. Under finite normal mixture modeling, the within-group distributions of mixture indicators are assumed to be normally distributed. Under latent profile models, the indicators need not be normally distributed, but the model assumes that indicators are locally independent for theoretical reasons. Conceptually, the latent variable in finite normal mixture models is a moderator whereas it is an explanatory variable in latent profile models.

Many simulation-based investigations of mixture model selection are based on within-group normality (Dolan and van der Maas, 1998; Everitt, 1981; Lo et al., 2001; Lubke and Neale, 2006; McLachlan and Peel, 2000; Morgan, 2015; Nylund et al., 2007). As an initial investigation, we chose to focus on the extent to which the true number of underlying profiles that were nonnormally distributed could be recovered using latent profile analysis. Then, building on the ideas explored in Milligan and Cooper (1988), we examined the potential impact standardization of indicators would have on model selection aided by fit indices.

1.1. Model selection using fit indices

In general, mixture model fit indices reflect absolute model fit, relative fit, classification certainty, and validation (Collins and Lanza, 2010). The likelihood index (L) serves as the primary basis for model selection in mixture modeling (McLachlan