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

and Peel, 2000). One approach based on L uses the likelihood ratio as a test statistic. A second approach imposes a penalty on the likelihood ratio based on the sample size, estimated parameters, or both. The latter approach permits comparisons of the relative fit for competing models. The indices that fall under this approach may be referred to as relative fit indices. Classification-based information criteria may also be considered when researchers select a particular model to judge the accuracy of the classification process.

In addition to fit indices, researchers must also weigh factors of parsimony and interpretability. Collins and Lanza (2010) describe parsimony as a philosophical principle that prefers simpler models over more complex models, all other things being equal. More parsimonious models require the estimation of fewer parameters.

1.2. Absolute model fit

Log-Likelihood Value

Absolute fit indices illustrate how well the K-class model fits the data irrespective of model-data fit of other models. It is more common to use the logarithm of the likelihood (log-likelihood) instead of L for greater simplicity in interpretation. Larger values of the log-likelihood ratio value (i.e., closer value is to 0), provide stronger evidence that the model fits the data. Comparing log-likelihood values of competing models is likely to prove unfruitful because the log-likelihood tends to identify models with more underlying subgroups. That is, the log-likelihood will likely indicate that fit is best for the model in which every response profile is its own subgroup. Such a model violates the parsimony principle and would also likely be difficult to interpret.

Lo–Mendell–Rubin likelihood ratio test

Traditionally, comparing nested models is common practice in SEM (Bollen, 1989). Unfortunately, the likelihood ratio test (LRT) cannot be conducted in the same manner in mixture modeling because the likelihood ratio difference between models with different numbers of classes does not often follow a χ^2 distribution (McLachlan and Peel, 2000). Furthermore, the traditional LRT of nested models examines differences between two models that specify the same number of groups but differ only with respect to the model parameterization. Competing mixture models may specify different numbers of underlying profiles. Models that differ in the number of underlying mixtures necessarily have different model parameterizations, which yields the traditional test of nested models inappropriate.

Lo et al. (2001) extended the work of Vuong (1989) by applying the Vuong test to allow comparisons between competing models with different numbers of components. The Lo–Mendell–Rubin likelihood ratio test (LMR) is a global test of model fit in which the LRT distribution is approximated, thus making the comparison of neighboring class models possible (Lo et al., 2001). The LMR compares the improvement in fit between neighboring class models (i.e., compares a model with $k-1$ classes [H_0] against one with k classes [H_1]) and provides a p-value that can be used to evaluate whether the improvement in fit for the inclusion of one additional class is statistically significant with $1-\alpha\%$ confidence (Nylund et al., 2007). The likelihood ratio under the null hypothesis from Lo et al. (2001, p. 771) is expressed:

$$\mathrm{LR} = \mathrm{LR}\left(\hat{\theta}, \hat{\gamma}; x\right) = L_f(\hat{\theta}; x) - L_g(\hat{\gamma}; x) = \sum_{j=1}^{n} \log \frac{f(X_j; \hat{\theta})}{g(X_j; \hat{\gamma})}. \tag{3}$$

An ad hoc adjustment was also made to the LMR to improve the accuracy of inferences based on the test.

$$2\mathrm{LR}^* = \frac{2\mathrm{LR}}{1 + (p-q)\log n^{-1}}. \tag{4}$$

It is important to note that Jeffries (2003, p. 991) pointed out that the conditions for the LMR theorem are not generally satisfied in the mixture modeling context. When the null hypothesis of LMR is true, the parameters of the additional component hypothesized under the alternative hypothesis do not exist. Therefore, they do not have a unique maximum in parameter space, which is a violation of one of the assumptions presented in Lo et al. (2001). Despite this flaw, LMR has been widely used and has been shown to be effective in recovering the number of underlying components (Morgan, 2015; Nylund et al., 2007; Tofighi and Enders, 2007).[2]

1.3. Relative fit indices

Relative fit indices are interpreted in relation to the fit estimates from other models. These indices are computed by imposing a penalty on the log-likelihood function when more parameters are estimated and/or when fewer subjects are included in the analysis. The Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and slightly modified versions of these criteria offer comparative evidence to evaluate different solutions (Muthén, 2001; Vermunt and Magidson, 2002).

[2] We wish to thank an anonymous reviewer for her/his constructive comments on this topic.

Both AIC and BIC are based on the value of −2 times the log-likelihood of the model and are adjusted for the number of parameters in the model (Akaike, 1977). The AIC can be defined as:

$$\text{AIC} = -2\log L + 2p, \tag{5}$$

where p is the number of free model parameters. Derived from the AIC, the consistent AIC (CAIC) penalizes the value of −2 times the log likelihood of the model for the number of free model parameters using the sample size (Bozdogan, 1987). The CAIC is defined as:

$$\text{CAIC} = -2\log L + p\left[\log(n) + 1\right], \tag{6}$$

where p is the number of free parameters and n is the sample size. The adjustment made to CAIC increases the penalty for model complexity, but it is not as commonly used as AIC as an information-based fit criteria.

An additional AIC-based criterion, the corrected AIC (AICc) has also been proposed based on the work of Hurvich and Tsai (1989). It imposes an additional penalty as a function of the sample size. The AICc is defined as:

$$\text{AICc} = \text{AIC} + \frac{(2(p+1)(p+2))}{(n-p-2)}, \tag{7}$$

where p is the number of free parameters and n is the sample size.

In addition to adjusting for the number of parameters, BIC adjusts for the sample size, yielding larger values as sample size increases, with all other factors held constant (Schwarz, 1978). The penalized likelihood under BIC is defined as:

$$\text{BIC} = -2\log L + p\log(n), \tag{8}$$

where p is the number of free parameters and n is the sample size. Sclove (1987) proposed an adjustment to BIC that modifies the sample size used in the BIC definition in order to reduce the sample size penalty. The sample size adjusted BIC (SSBIC) is defined as:

$$\text{SSBIC} = -2\log L + p\log\left[\frac{(n+2)}{24}\right], \tag{9}$$

where p is the number of free parameters and n is the sample size.

Draper's information criterion (DIC) (Draper, 1995) is based on a slight modification to BIC. The modification is likely omitted in BIC due to its trivial influence asymptotically, but Draper (1995) suggested that the modification may improve BIC in finite samples. The DIC is defined as:

$$\text{DIC} = -2\log L + p\left(\log\left(\frac{n}{2\pi}\right)\right), \tag{10}$$

where p is the number of free parameters and n is the sample size. For all information-based criteria, comparatively lower values indicate better fitting models (Muthén and Muthén, 2010) although all criteria may not be minimized for the same model (Collins and Lanza, 2010).

Prior simulation studies have shown that AIC tends to overestimate the number of underlying mixtures (Bacci et al., 2014; Celeux and Soromenho, 1996; Henson et al., 2007; Koehler and Murphree, 1988; Morgan, 2015; Nylund et al., 2007; Soromenho, 1994; Yang, 2006), despite being a frequently reported index. The BIC provides information for both selecting models with varying numbers of mixtures and selecting between competing models that are parameterized differently, such as restrictions made to the covariance matrices (McLachlan and Peel, 2000). When the value of $\log(n)$ is greater than two, BIC tends not to overestimate the number of underlying mixtures as the AIC does because it imposes a stronger penalty to L. Under smaller sample size conditions, BIC has been shown to underestimate the number of classes using sample sizes (Celeux and Soromenho, 1996; Morgan, 2015; Tofighi and Enders, 2007). Morgan (2015) found the BIC and SSBIC tended to identify true number of classes with higher frequency than other criteria except under conditions with very rare classes. Yet, when rare classes were present, none of the indices examined performed well (Morgan, 2015). Nylund et al. (2007) found that BIC generally outperformed all other information-based criteria, including AIC, CAIC, and SSBIC. Yang (2006) found SSBIC to outperform AIC, CAIC, BIC, and other information-based criteria not discussed here in models with categorical indicators and at least 50 subjects per latent class. These findings suggest that SSBIC may be related to class enumeration and sample size in order to maintain a certain subject-to-class ratio. Bacci et al. (2014) found criteria ability of detecting the true number of latent Markov states was strongly affected certain design factors. BIC and CAIC performed poorly with higher levels of uncertainty in allocating observations into latent states, whereas AIC and AICc were less affected by high uncertainty and low persistence (Bacci et al., 2014). Increasing sample size and number of time points positively influence detection behavior of all criteria (Bacci et al., 2014). Although Bacci et al. (2014) did include AICc in their examination of latent states, AICc and DIC have not been as systematically studied as the other information criteria presented in this section.

1.4. Classification uncertainty

Classification-based criteria require that individuals are grouped to determine how well the model works to classify cases. Entropy-based measures provide the degree of certainty in classification procedures with one index. These measures are computed as the maximum of the probability density distribution underlying the mixture model (Akaike, 1977). A commonly used entropy-based measure referred to as relative entropy, E_k, was defined by Ramaswamy et al. (1993):

$$E_k = 1 - \frac{\sum_{i=1}^{N}\sum_{k=1}^{K}\left(-\pi_{(k|y_i)}\log\pi_{(k|y_i)}\right)}{n\log(K)}. \tag{11}$$

It should be noted that the $\pi_{(k|y_i)}$ in Eq. (9) corresponds to p_{ik} in Eq. (1).

Relative entropy is bounded by 0 and 1 with larger values indicating a greater degree to which latent classes are distinguishable by the data and the model (Muthén, 2004). One disadvantage of using an entropy-based index is that it can be negatively impacted by misclassification due to chance when more mixtures are allowed (Collins and Lanza, 2010). That is, there is greater opportunity for misclassification with more mixtures, which may result in decreases in entropy-based estimates. Entropy assumes that the estimated model is correct, and there is no agreed upon lower bound for acceptable entropy values (Pastor et al., 2007).

Another classification-based information criterion is the integrated classification likelihood criteria (Biernacki et al., 1998). It can be estimated using a BIC-type approximation (ICL-BIC). The approximation is discussed here because it is much easier to compute, and when class sizes are sufficiently large the performance of the ICL-BIC differs very little from the more accurate version (Biernacki et al., 1998). The ICL-BIC is defined as:

$$\text{ICL-BIC} = -2\log L + 2O_k + p\log(n), \tag{12}$$

where p is equal to the number of free parameters, O_k is the raw entropy, and n is the sample size. Raw entropy, O_k, is the term in the numerator of Eq. (11). When the raw entropy term is not included, ICL-BIC is equivalent to BIC.

The ICL-BIC incorporates a penalty for poor class separation, model complexity, and sample size. Like AICc and DIC, ICL-BIC has not been as widely examined in simulation studies as the other indices reported here though a several studies have included it. McLachlan and Ng (2000) included ICL-BIC under very limited conditions (i.e., three samples of $n_1 = 625$; $n_2 = 300$; $n_3 = 200$), and ICL-BIC identified the true number of underlying classes in all instances. In another study, Fonseca and Cardoso (2005) examined multiple measures of fit with mixture analyses and found that ICL-BIC outperformed other indices studied (i.e., BIC, CAIC, AICc, L, AIC, Entropy, CLC, and two variants of AIC not discussed here [AIC3 and AICu]).

While previous research has offered some guidance when choosing fit indices to interpret for use with mixture models, additional questions remain. Many other fit indices exist outside of those produced by popular software packages (e.g., CAIC, AICc, DIC), and these have not been quite as systematically investigated. Furthermore, researchers often collect data that are nonnormally distributed. Recent studies have begun looking at the utility of mixtures of nonnormal distributions (e.g., skew-t and skew-normal) (Lee and McLachlan, 2013; Murray et al., 2014; Vrbik and McNicholas, 2014). We examined a potential alternative strategy modeling nonnormal mixture when nonnormality is nonsubstantive. First, we examined the extent to which model fit statistics could recover the true number of underlying classes with nonnormally distributed indicators. Second, we examined the extent to which transforming nonnormal indicator distributions using van der Waerden quantile normal scores impacts the fit indices' model selection performances. The use of nonnormally distributed underlying profiles has important implications for any analyses that may follow the selection of a particular profile-enumeration model. Although the primary focus of these studies is the identification of underlying profiles through model fit comparisons, the implications for subsequent analysis are discussed following the presentation of findings. We should emphasize here that our focus is on the frequency with which the fit indices identify the model with the same number of components as were generated. These so-called "automatic methods" (Hennig and Liao, 2013) should be used with caution in practice and should always be considered in conjunction with one's guiding theory, interpretability, and meaningfulness.

2. Method

The current study focuses on mixture models for which indicators have different continuous distributions. Thus, the models examined here may also be referred to as latent profile models. As such, we use the term "profile" instead of "mixtures" or "classes" in the presentation and discussion of the findings. The general latent profile model can be written:

$$f(\mathbf{y_i}|\boldsymbol{\Phi}) = \sum_{k=1}^{K}\pi_k f_k\left(\boldsymbol{y_i}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\right), \tag{13}$$

where $\mathbf{y_i}$ denotes the profile of scores for case i across the set of variables, Φ is a mixture of the class-specific joint distributions of the indicators, K is the number of underlying classes, π_k denotes the prior probability of belonging to profile k (i.e., profile prevalence), and $\boldsymbol{\mu_k}$ is the profile mean vector for profile k, and $\boldsymbol{\Sigma_k}$ is the covariance variance for profile k.

The within-component distribution, f_k, was a normal distribution. In the current study, no class-specific covariances were estimated between indicators.

The simulation and estimation was conducted in Mplus (version 6.12; Muthén and Muthén, 2010), and the estimation was automated using the *MplusAutomation* package (Hallquist, 2011) in R (version 2.15.1; R Development Core Team, 2010). Robust maximum likelihood estimation produces standard errors of parameter estimates that are robust to nonnormality, thus producing standard errors corrected for attenuation (Muthén and Muthén, 2010). This estimation technique is readily available in Mplus, which was used. The number of random starts was set to 500 for each solution.

2.1. Measures of model fit

Although many measures of model fit have been proposed, we focused on measures of fit that are provided by Mplus (i.e., LL, AIC, BIC, SSBIC, entropy, LMR) or can be computed from Mplus output (i.e., CAIC, AICc, DIC, ICL-BIC). For each set of models tested per replication, the model with the highest log-likelihood value was selected as the best fitting model according to the log-likelihood. For each LMR test, a calculated p-value reports the probability of observing a particular likelihood ratio or greater if K is equal $k - 1$. The p-value was compared against a Type I error rate (α) of 5%. Neighboring profile model comparisons stopped when the k-profile did not result in better model fit (i.e., $p \geq \alpha$) than the $k - 1$ profile model. For AIC, BIC, SSBIC, CAIC, AICc, DIC, and ICL-BIC, the model with the smallest values on each index was selected as the best fitting model of those tested for each replication. The mean and standard deviation of the entropy values will be computed for the three class solution within each condition.

2.2. Empirical conditions

To create conditions for the study that mirror empirical research situations, a review of applied studies employing mixture models was conducted. The selection of studies was identified by searching the *Elton B Stephens Company (EBSCO)*, *Education Full Text*, *Education Resources Information Center (ERIC)*, *PsycINFO*, and *PsycARTICLES* databases. Keywords used to locate relevant studies were: latent class analysis, latent profile analysis, mixed mode clustering, finite mixture models, and latent class clustering. The search was also limited to include recent (2005–2010) peer-reviewed, full text articles available through a large southeastern university library. It should be noted that searching large databases does not ensure that all relevant studies will be discovered. Unpublished technical reports or masters theses, for example, are likely to be underrepresented or missed.

Studies were selected that reported the use of mixture models in peer-reviewed education and psychology journals. The modal number of identified classes in the papers using mixture modeling was three although the true number of underlying classes cannot be determined from this information. The sample sizes ranged from approximately 200 to slightly less than 2500 and used between three and 26 indicators of class membership. The five number summary for the sample sizes was 196, 386, 689, 1429, 2381 and for indicators was 3, 5.5, 10, 15, 26.

2.3. Population models

Use of a population with a known structure allows for investigation of the performance of fit indices when true characteristics are known. The population structure used in the current study is a mixture model with known number of profiles, profile prevalence, and profile-specific continuous indicator distributions. The population structure also includes local independence. Samples were generated from these structures and thus contain an element of random error. Models were then fit to the data to determine the extent to which the true structure was recovered.

Based on the information from the review of applied studies, each simulated model included 5, 10, or 15 indicators. The sample sizes employed were of sizes 400, 800, and 1200, which are very close to the quartiles from the empirical research conditions. These indicator and sample size conditions yielded subject-to-parameter ratios from 6.5:1 (i.e., 400 cases with 62 free parameters) to 54.5:1 (i.e., 1200 cases with 22 free parameters).

Although the number of classes found in the applied literature ranged from two to eight classes, the majority of selected models contained three classes or profiles. It should be noted that the profile enumeration solutions found in the empirical research may not be representative of the true number of underlying profiles. The profile enumeration was limited to three.

Profile prevalence indicates the proportion of the overall sample attributed to each profile. A review of the three-class models revealed three general patterns of prevalences. There was generally a profile within which the majority of the cases were classified along with two smaller profiles. The prevalences used for the three profile population models for profiles 1, 2, and 3 respectively were 0.59 — 0.26 — 0.15, 0.45 — 0.40 — 0.15, and 0.89 — 0.08 — 0.03.

Indicator distributions

Five distributions were used for the mixture indicators and were distinguished on the basis of skewness and kurtosis. The first distribution condition used normally distributed continuous indicators. The remaining conditions were fully crossed between skewness of 0.75 and 1.25 and kurtosis of 1.75 and 3.75. The means of continuous indicators for profile 1, 2, and 3 were set to 1, 0, and −1, respectively, and the variances for each distribution were set to one. These skewness and kurtosis values were used by Flora and Curran (2004) in their examination of nonnormally distributed factor indicators in confirmatory factor analysis with ordinal variables and a single underlying population.

Table 1
Fleishman coefficients for desired indicator distributions.

Desired skewness	Actual skewness	Desired kurtosis	Actual kurtosis	*B*	*C*	*D*
0.00	0.00	0.00	0.00	1.00	0.00	0.00
0.75	0.69	1.75	1.42	0.89	0.10	0.03
0.75	0.69	3.75	2.86	0.77	0.09	0.07
1.25	1.19	1.75	1.52	1.00	0.25	−0.02
1.25	1.12	3.75	2.85	0.82	0.16	0.05

Note. The coefficients were entered into the Fleishman power method formula: $Y = A + BX + CX^2 + DX^3$, where Y is the transformed value, X is the value from a standard normal distribution, and A equals $-C$. The B, C, and D coefficients are rounded to two decimal places. Nine decimal places were used in the actual transformation.

Fleishman's power method (Fleishman, 1978) was applied in order to obtain indicator distributions with the specified levels of skewness and kurtosis. We used the following general steps to generate the data.

1. Generated a three-component mixture model where each within-component distribution was standard normal.
2. Used Fleishman coefficients to transform each within-component distribution to have the desired levels of skewness and kurtosis, on average.
3. Added a constant to two of the component distributions to specify separation.

When creating normal mixtures, step 2 was excluded. We should note that generating a mixture of standard normal distributions and later adding a constant is equivalent to generating normal mixture model in which the components have different means. The coefficients for each distribution shape are provided in Table 1 as well as the observed skewness and kurtosis in the generated distributions. The levels of skewness and kurtosis present in the generated data were slightly lower than the respective generating parameters.

In the second simulation study, all values in the joint-distributions from the first simulation study were transformed with van der Waerden quantile normal scores. These scores were computed as:

$$Z_i = \Phi^{-1}\left(\frac{R(X_i)}{N+1}\right), \tag{14}$$

where Φ is the cumulative distribution function of the standard normal distribution, $R(X_i)$ is the rank of the value for the ith person (lowest score is assigned a rank of 1), and N is the sample size. Midranks were used if ties were present in the data. The two- through five-profile models were then fit using the normal scores instead of the raw data. We should note a side effect of using this type of score transformation. The use of normal scores may reduce component separation. For example, in condition 1 (i.e., $(\pi_1 = 0.59, \pi_2 = 0.26, \pi_3 = 0.15;\ \mu_1 = 1, \mu_2 = 0, \mu_3 = -1;\ \sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 1,\ n = 400;\ sk = 0;\ ku = 0;)$ of the current study the expected mean and variance of the observed raw score distribution for each indicator are respectively:

$$\mu = \sum_{k=1}^{K} \pi_k \mu_k = 0.59 * 1 + 0.26 * 0 + 0.15 * -1 = 0.44 \tag{15}$$

$$\begin{aligned}\sigma^2 &= \sum_{k=1}^{K} \pi_k (\mu_k - \mu)^2 + \sum_{k=1}^{K} \pi_k \sigma_k^2 = (0.59 * (1 - 0.44)^2) + (0.26 * (0 - 0.44)^2) \\ &\quad + (0.15 * (-1 - 0.44)^2) + (0.59 * 1 + 0.26 * 1 + 0.15 * 1) = 1.55.\end{aligned} \tag{16}$$

Following normal score transformation, the observed distribution for each indicator is standard normal. Therefore, the variability of the joint-distribution was reduced from 1.55 to 1.0 by applying normal score transformation. Although the classes may still be distinguished by their means, the overall variability in the data has been restricted, which may adversely affect those fit indices that are sensitive to profile separation.

2.4. Summarizing conditions

Models were fit testing two through five profile solutions for each replication, which resulted in 2000 outputs per condition. All analyses were conducted using R (version 2.15.1; R Development Core Team, 2010). Replications that provide implausible values or did not converge were deemed unusable. As in previous studies (e.g., Flora and Curran, 2004; Yang-Wallentin et al., 2010), these solutions were removed prior to analysis given that they do not provide useful information (Forero et al., 2009). Descriptive information (e.g., proportions of selected models) was adjusted accordingly based upon the number of usable replications for a given cell.

For each converged solution tested, the fit measures outlined above were recorded and used to identify the best approximating model for each replication by applying interpretation rules for each measure. Next, for each set of competing

Table 2
Overall model selection rates (percentage of fitted models that recovered the simulated profiles) by indicator distribution.

Mixtures of raw scores					
Index	Indicator distribution				
	$Sk = 0$; $Ku = 0$	$Sk = 0.75$; $Ku = 1.75$	$Sk = 0.75$; $Ku = 1.75$	$Sk = 1.25$; $Ku = 3.75$	$Sk = 1.25$; $Ku = 3.75$
AIC	33	0	0	0	0
CAIC	70	44	25	22	15
AICc	65	0	0	0	0
BIC	72	43	21	17	9
SSBIC	87	0	0	0	0
DIC	82	16	2	2	0
ICL-BIC	47	50	36	47	26
LMR	77	50	47	30	41
Mixtures of standardized scores					
Index	Indicator distribution				
	$Sk = 0$; $Ku = 0$	$Sk = 0.75$; $Ku = 1.75$	$Sk = 0.75$; $Ku = 3.75$	$Sk = 1.25$; $Ku = 1.75$	$Sk = 1.25$; $Ku = 3.75$
AIC	16	5	4	0	2
CAIC	60	67	63	68	68
AICc	42	24	17	6	13
BIC	64	70	66	68	69
SSBIC	78	61	40	18	29
DIC	73	75	65	54	58
ICL-BIC	64	70	66	68	69
LMR	79	73	71	70	67

models, the number of profiles in the best fitting solution for each index was compared against the number of profiles simulated. When the number of profiles between the best fitting and true models match, this replication assigned a "1" and mismatching profile enumeration models were assigned a "0". This process was repeated for each fit index for each replication within the study conditions. The mean, which is equal to the proportion of matching solutions, was reported for each fit index. Here, higher mean values indicate higher rates of concordance between the number of profiles in the models identified as best fitting and true models.

3. Results

A total of 67,500 datasets were successfully generated (135 conditions × 500 replications = 67,500). All simulated datasets were examined, and the general structure of the samples closely approximated that specified by the population model. Given that researchers typically do not know the true number of profiles underlying a dataset, various profile solutions were tested. Models were fit to the data using the MIXTURE option in Mplus (v6.12).

3.1. Study #1

Convergence and model summaries were recorded for each fitted model. Convergence rates were 100% for the models fit to the data. The overall three-profile model selection frequency for each fit index across all indicator distribution conditions is presented in Table 2. When considered across all cells of the design, there was a sharp decline in the performance of the fit indices' ability to identify the number of simulated profiles when the profile indicators deviated from normality even if only slightly. The percentages of true and selected models with matching numbers of profiles for all cells in the design with normally distributed indicators are presented in Table 3. The percentages of true and selected models with matching numbers of profiles for all cells with nonnormally distributed indicators were collapsed and are presented in Table 4.

Information criteria

Akaike's Information Criterion (AIC) and its variants. Overall, the AIC and its variants functioned poorly as a criterion for model selection, but their performance was largely influenced by the indicator distributions. When the indicator distributions followed a normal distribution, CAIC was most frequently identified the three-profile solution at 70.2% followed by AICc at 64.7% and AIC at 32.6%. The frequency deteriorated very quickly as the indicator distributions deviated from normality. When indicators were not normally distributed, AIC did not identify the number of simulated profiles (i.e. 3) in *any* of the generated datasets. When skewness was 0.75, the AICc identified three latent profiles in less than 0.5% of the datasets. As the degree of nonnormality increased, CAIC dropped to 14.8% when skewness and kurtosis were 1.25 and 3.75, respectively. The effect of nonnormality was mitigated when more indicators were present. With 15 normally distributed indicators, CAIC identify the three profiles in 88.4% of datasets as opposed to 43.9% of datasets with 5 normally distributed indicators. When indicator distributions were nonnormal AIC and AICc were of little to no value for model selection regardless of sample size

Table 3
Accuracy of all fit indices for all normally distributed indicators cells (rounded to the nearest percentage point).

Profile Prev.	No. of Ind.	Sample size	Fit index							
			AIC	*AICc*	*CAIC*	*BIC*	*SSBIC*	*DIC*	*ICL-BIC*	*LMR*
$\pi_1 = 0.59$ $\pi_2 = 0.26$ $\pi_3 = 0.15$	5	400	42	62	8	17	77	56	0	46
		800	44	56	56	71	97	93	0	79
		1200	43	51	89	96	99	100	0	83
	10	400	36	87	100	100	94	100	50	96
		800	30	62	100	100	100	100	85	95
		1200	28	50	100	100	100	100	96	95
	15	400	28	98	100	100	96	100	100	97
		800	23	79	100	100	100	100	100	98
		1200	20	60	100	100	100	100	100	98
$\pi_1 = 0.45$ $\pi_2 = 0.40$ $\pi_3 = 0.15$	5	400	45	64	29	48	85	81	0	65
		800	46	56	94	96	98	100	0	86
		1200	45	50	100	100	99	100	0	86
	10	400	34	89	100	100	95	100	89	95
		800	29	58	100	100	100	100	100	94
		1200	23	45	100	100	100	100	100	96
	15	400	28	98	100	100	97	100	100	97
		800	20	77	100	100	100	100	100	97
		1200	20	56	100	100	100	100	100	98
$\pi_1 = 0.89$ $\pi_2 = 0.08$ $\pi_3 = 0.03$	5	400	30	35	0	0	19	2	0	13
		800	34	39	0	0	19	5	0	19
		1200	41	45	0	1	27	7	0	29
	10	400	36	76	1	3	71	27	0	19
		800	34	65	15	31	96	74	0	70
		1200	33	53	57	76	99	97	0	92
	15	400	33	96	7	22	93	73	8	40
		800	26	78	77	92	100	99	58	97
		1200	27	62	99	100	100	100	90	99

Table 4
Accuracy of all fit indices for all nonnormally distributed indicators cells (rounded to the nearest percentage point).

Profile Prev.	No. of Ind.	Sample size	Fit index							
			AIC	*AICc*	*CAIC*	*BIC*	*SSBIC*	*DIC*	*ICL-BIC*	*LMR*
$\pi_1 = 0.59$ $\pi_2 = 0.26$ $\pi_3 = 0.15$	5	400	0	0	28	21	0	5	30	32
		800	0	0	6	2	0	0	27	21
		1200	0	0	0	0	0	0	22	13
	10	400	0	0	67	43	0	8	33	73
		800	0	0	25	15	0	1	39	50
		1200	0	0	11	4	0	0	42	32
	15	400	0	1	95	83	0	24	85	85
		800	0	0	68	43	0	10	71	68
		1200	0	0	36	24	0	4	62	49
$\pi_1 = 0.45$ $\pi_2 = 0.40$ $\pi_3 = 0.15$	5	400	0	0	29	20	0	3	30	34
		800	0	0	5	2	0	0	29	24
		1200	0	0	0	0	0	0	26	17
	10	400	0	0	75	52	0	9	51	78
		800	0	0	32	18	0	2	44	62
		1200	0	0	16	8	0	0	40	46
	15	400	0	1	96	87	1	32	86	86
		800	0	0	79	59	0	14	70	81
		1200	0	0	54	32	0	8	61	68
$\pi_1 = 0.89$ $\pi_2 = 0.08$ $\pi_3 = 0.03$	5	400	0	0	17	13	0	2	28	28
		800	0	0	2	0	0	0	24	21
		1200	0	0	0	0	0	0	19	12
	10	400	0	0	28	20	0	6	25	26
		800	0	0	12	7	0	0	21	21
		1200	0	0	2	0	0	0	14	14
	15	400	0	0	16	35	0	12	19	34
		800	0	0	28	19	0	1	38	36
		1200	0	0	14	4	0	0	39	21

or number of indicators. When AIC, AICc, and CAIC failed to identify a three-profile solution, they all tended to overestimate the number of profiles.

Bayesian Information Criterion (BIC) & Sample Size-Adjusted Bayesian Information Criterion (SSBIC). The BIC seemed to be most heavily influenced by the degree of nonnormality and the number of variables included in the model. When nonnormality was most extreme, BIC identified three profiles in only 9.2% replications. Across indicator distributions, BIC frequency of selecting the three-profile solution increased from 14.7% to 30.5% to 52.5% as the number of indicator variables increased from 5 to 10 to 15, respectively. The SSBIC was also influenced by nonnormality and the number of indicator variables, but the nonnormality effect was detrimental to SSBIC three-profile model selection. When the profile indicators were generated from normal distributions, SSBIC identified three profiles in 68.7%, 94.9%, and 98.4% of replications when there were 5, 10, or 15 indicators, respectively. Yet, when the indicator variables followed a nonnormal distribution SSBIC correctly identified three profiles for 0.3% or fewer solutions regardless of how many indicators were used. When the indicators were normally distributed, both BIC and SSBIC underestimated the number of underlying profiles. When indicators were nonnormally distributed, both BIC and SSBIC overestimated the number profiles.

Draper's Information Criterion (DIC). The performance of DIC was similar to other indices included in the study with regard to nonnormality. As the degree of nonnormality increased, the DIC performance decreased substantially—from 81.9% with normally distributed indicators to 0.2% with indicators with skewness and kurtosis of 1.25 and 3.75, respectively. In the cells with normally distributed indicators, DIC identified the solution with the highest number of solutions of those tested at least 60.0% of the time and often more than 90.0% of the time. When 15 normally distributed indicators and sample sizes of 1200 were used, DIC identified three profiles in 100% of the conditions, regardless of the mixing weight distribution. When nonnormally distributed indicators were used, the DIC identified three profiles less than a third of the time.

Classification uncertainty

Of all indices examined, entropy seemed to be the least impacted by nonnormality. The three cells with the highest mean entropy values ($E_k \geq 0.972$) all had 15 indicators and profile prevalences of 0.89 (π_1), 0.08 (π_2), and 0.03 (π_3). They had sample sizes of 400 or 800, and the indicator distributions were nonnormal ($ku = 3.75$). The cells with the lowest mean entropy values ($E_k \leq 0.657$) differed only by sample size. All three cells had normally distributed indicators, five indicators, and profile prevalences of 0.45 (π_1), 0.40 (π_2), and 0.15 (π_3).

Overall, the performance of ICL-BIC was mixed. Like the other indices investigated, the fit as indicated by ICL-BIC deteriorated as the degree of nonnormality increased. Only when the number of indicators increased did ICL-BIC show minor resilience against nonnormality. Among the nonnormal indicator conditions, as the number of indicators increased from 5 to 10 to 15, the percentage of three-profile models selected increased to 26.1% to 34.1% to 59.0%, respectively. The profile prevalence also seemed to affect ICL-BIC performance. In two of the profile prevalence conditions ($\pi_1 = 0.59, \pi_2 = 0.26, \pi_3 = 0.15$; $\pi_1 = 0.45, \pi_2 = 0.4, \pi_3 = 0.15$), the frequency with which the ICL-BIC selected the three-profile solution increased from 45.6% to 48.3% with nonnormal indicators. In the prevalence condition with one dominant profile and two rare profiles ($\pi_1 = 0.89, \pi_2 = 0.08, \pi_3 = 0.03$), ICL-BIC identified 25.3% of the three-profile conditions with nonnormal indicators as best fitting.

Lo–Mendell–Rubin likelihood ratio test (LMR)

LMR was not as strongly impacted by the degree of nonnormality as other indices, but it did show some variability based on the design factors. Across the collapsed nonnormal conditions the rate of three-profile solution identification ranged from 12% to 87% (see Table 4). As might be expected, the LMR returned the three-profile solution more frequently as the number of indicator variables increased. With 5 and 15 indicator variables, LMR selected the three-profile model in 29.1% and 65.4% of the datasets, respectively. LMR was less likely to overestimate profile enumeration than the other fit indices.

3.2. Study #2

For the second study, we replicated the process used for the first simulation study except we estimated the mixture models using the normal score-transformed data. Convergence and model summaries were recorded for each fitted model. Convergence rates were 100% for the models fit to the data. The overall three-profile model selection percentage for each fit index across all indicator distribution conditions is presented in the bottom half of Table 2. The percentages of three-profile models selected for every cell with normally distributed indicators in the design are presented in Table 5. The percentages of three-profile models selected for every cell with nonnormality distributed indicators in the design were collapsed and are presented in Table 6.

Information criteria

Akaike's Information Criterion (AIC) and its variants. With the use of normal score transformation in the conditions with nonnormal indicators, the AIC and AICc functioned poorly as a criterion for model selection. CAIC performed very well when mixing weights were more equally distributed, sample size was larger, and/or when there were more indicators. The CAIC's ability to detect three profiles deteriorated very quickly as the transformed indicator distributions deviated from original normal distributions for AIC and AICc. The effect of nonnormality was mitigated when more indicators were present. With 15 normally distributed indicators, CAIC identify three profiles in 86.2% of datasets as opposed to 46.2% of datasets with 5 normally distributed indicators in the most extreme nonnormality condition. AIC and AICc were generally of little to no value for model selection across conditions even with normal score transformation. AIC and AICc tended to overestimate

Table 5
Frequency of three-profile model selection for all fit indices after normal score transformation for all normally distributed indicators cells (rounded to the nearest percentage point).

Profile Prev.	No. of Ind.	Sample size	Fit index							
			AIC	*AICc*	*CAIC*	*BIC*	*SSBIC*	*DIC*	*ICL-BIC*	*LMR*
		400	24	42	3	7	56	33	0	47
	5	800	6	10	29	48	76	80	0	84
		1200	1	1	71	83	73	95	0	86
$\pi_1 = 0.59$		400	13	69	96	100	83	100	20	98
$\pi_2 = 0.26$	10	800	2	15	100	100	94	100	39	100
$\pi_3 = 0.15$		1200	0	2	100	100	96	100	60	99
		400	8	96	100	100	89	100	100	99
	15	800	1	29	100	100	99	100	100	100
		1200	0	4	100	100	100	100	100	100
		400	50	68	20	35	84	73	0	69
	5	800	42	53	82	93	96	99	0	91
		1200	30	34	100	100	97	100	0	89
$\pi_1 = 0.45$		400	34	87	100	100	95	100	79	97
$\pi_2 = 0.40$	10	800	20	50	100	100	99	100	99	97
$\pi_3 = 0.15$		1200	11	29	100	100	100	100	100	99
		400	22	99	100	100	98	100	100	98
	15	800	10	66	100	100	100	100	100	98
		1200	7	36	100	100	100	100	100	98
		400	21	22	0	0	8	1	0	7
	5	800	21	26	0	0	7	1	0	14
		1200	13	18	0	0	8	1	0	24
$\pi_1 = 0.89$		400	24	39	0	0	26	4	0	22
$\pi_2 = 0.08$	10	800	19	44	1	1	59	19	0	74
$\pi_3 = 0.03$		1200	10	23	5	16	89	60	0	93
		400	24	78	0	1	70	24	0	55
	15	800	15	65	20	44	99	90	2	98
		1200	8	40	79	94	100	100	18	99

the number of profiles, which is consistent with previous research. CAIC tended to strongly underestimate the number of profiles when it failed to identify three profiles—95.9% of the identified solutions contained only two profiles.

Bayesian Information Criterion (BIC) & Sample Size-Adjusted Bayesian Information Criterion (SSBIC). BIC tended to identify three profiles more frequently than SSBIC. Across indicator distributions, BIC increased from 42.8% to 72.4% to 86.3% as the number of indicator variables increased from 5 to 10 to 15, respectively. This pattern was mirrored within each indicator distribution condition, but the frequency of selecting three profiles decreased as nonnormality increased. The SSBIC seemed to be most heavily influenced by the degree of nonnormality and the number of variables included in the model. Across all of the variable conditions, SSBIC identified the three-profile solution in 77.87% of the datasets in the normal condition whereas it only identified 29% of the three-profile solutions when nonnormality was most extreme.

The SSBIC was also influenced by nonnormality and the number of indicator variables, but the nonnormality effect was very detrimental to SSBIC's ability to detect three profiles. When the profile indicators were generated from normal distributions, SSBIC identified the three-profile solution in 56.2%, 82.4%, and 95.0% when there were 5, 10, or 15 indicators, respectively. Yet, when the indicator variables followed a nonnormal distribution SSBIC identified three profiles in 54% or fewer solutions for 5 indicators, 65.5% or fewer for 10 indicators, and 86.7% or fewer for 15 indicators.

Previous research suggests that BIC and SSBIC tend to underestimate the number of components due to imposing a more severe penalty to the log-likelihood value. This finding was supported in the current study when the indicators were normally distributed. Yet, when indicators were nonnormally distributed, BIC underestimated and SSBIC overestimated the number profiles. When BIC failed to identify three profiles, it selected the two-profile solution as best fitting in all four of the nonnormal conditions. When SSBIC failed to identify three profiles it tended to select solutions with four or five profiles.

Draper's Information Criterion (DIC). The performance of DIC was similar to other indices included in the study with regard to nonnormality. As the degree of nonnormality increased, the DIC performance decreased from 73.22% with normally distributed indicators to 58.14% with skewness and kurtosis of 1.25 and 3.75, respectively. When DIC did not identify three profiles, 20.3% contained two profiles.

Classification uncertainty

Of all indices included, entropy seemed to be the least impacted (i.e., most stable) across nonnormality, but the entropy values were all very low. In general, entropy values in the solutions with three profiles tended to be slightly higher in nonnormal conditions and within these conditions were higher with indicators with more kurtosis.

Overall, the ICL-BIC performed very well in conditions with 15 indicators and/or sample sizes of at least 800. After transforming the nonnormally distributed raw data to normal scores, ICL-BIC identified three profiles in 100% of the

Table 6
Frequency of Three-Profile Model Selection for All Fit Indices After Normal Score Transformation for All Nonnormally Distributed Indicators Cells (Rounded to the Nearest Percentage Point).

Profile Prev.	No. of Ind.	Sample size	Fit index							
			AIC	*AICc*	*CAIC*	*BIC*	*SSBIC*	*DIC*	*ICL-BIC*	*LMR*
		400	1	6	14	25	19	50	0	54
	5	800	0	0	68	71	12	47	0	56
		1200	0	0	76	65	6	33	0	41
$\pi_1 = 0.59$		400	1	18	100	100	31	91	82	94
$\pi_2 = 0.26$	10	800	0	1	100	99	31	82	98	87
$\pi_3 = 0.15$		1200	0	0	99	96	24	64	98	76
		400	1	59	100	100	48	98	100	97
	15	800	0	3	100	100	56	97	100	98
		1200	0	0	100	100	48	95	100	96
		400	4	10	62	75	25	70	0	65
	5	800	0	1	94	87	18	50	0	54
		1200	0	0	83	67	11	33	0	37
$\pi_1 = 0.45$		400	1	13	100	100	21	86	99	94
$\pi_2 = 0.40$	10	800	0	0	100	98	22	72	99	86
$\pi_3 = 0.15$		1200	0	0	98	94	16	53	99	71
		400	1	40	100	100	31	95	100	98
	15	800	0	1	100	100	38	91	100	97
		1200	0	0	100	100	34	82	100	92
		400	16	23	0	0	19	2	0	14
	5	800	6	9	0	0	18	5	0	25
		1200	2	3	0	0	16	7	0	30
$\pi_1 = 0.89$		400	15	51	0	1	44	13	0	19
$\pi_2 = 0.08$	10	800	6	21	6	15	70	53	0	65
$\pi_3 = 0.03$		1200	2	7	38	57	78	86	1	85
		400	14	86	2	9	80	54	4	36
	15	800	6	41	60	79	90	98	53	93
		1200	3	18	95	98	90	100	84	98

conditions with 15 indicator variables and more balanced mixing weights (see Table 6). When there were five indicators, ICL-BIC failed to return three profiles in all of the normal score-transformed conditions.

Lo–Mendell–Rubin likelihood ratio test (LMR)

LMR seemed to perform better in the normal score-transformed conditions when the indicators were nonnormally distributed originally than for data that were originally normally distributed. This is likely due to the redistribution of variance that was discussed previously. After transforming nonnormal conditions to normal scores, LMR often identified three profiles more than 80% of the time when there at least 10 indicators and sample sizes of at least 800, regardless of the mixing weight distribution. The LMR performance was more stable across all conditions with normal score transformation than the other fit indices.

A comparison of Tables 3 and 5 shows that LMR identified the simulated number of profiles with slightly higher frequency after normal score transformation despite the original data being a mixture of normal distributions. The joint distribution after normal-score transformation followed a normal distribution much more closely than the joint distribution of original values but with smaller variance. The work of which LMR is an extension was based on joint normal distribution (Vuong, 1989). The slight improvement in LMR model selection performance may be due to the fact that normal score transformation produces a normal distribution.

3.3. Comparison of simulation results

Figs. 1 through 3 were generated to allow for easier comparisons to be made between fit index performance in latent profile models using raw data and normal transformation scores. AIC and AICc performed very poorly across all simulated conditions. CAIC performed very well in most of the conditions when nonnormally distributed variables were transformed to normal scores. BIC, SSBIC, and DIC all saw a decrease in performance as indicator distributions deviated from normality, but BIC and DIC both performed fairly well when using normal score transformation. The difference in performance of ICL-BIC with raw scores versus normal scores was not as discrepant as for the other indices. This is likely due to the influence of entropy, which was fairly stable across all conditions. Generally, the performance of LMR was least impacted by nonnormal profiles, and it performed better when normal score transformation was performed. Fig. 1 compares the frequency of three-profile model selection for AIC, AICc, and CAIC across indicator distribution conditions. Fig. 2 compares the frequency of three-profile model selection for BIC, SSBIC, and DIC across indicator distribution conditions. Fig. 3 compares the frequency of three-profile model selection of ICL-BIC and LMR across indicator distribution conditions.

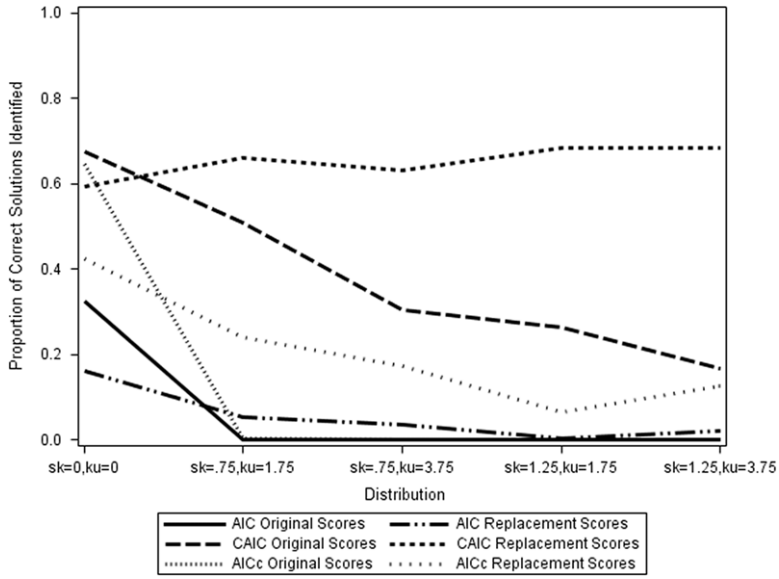

Fig. 1. Proportion of three-profile solutions selected by AIC, AICc, and CAIC by indicator distribution.

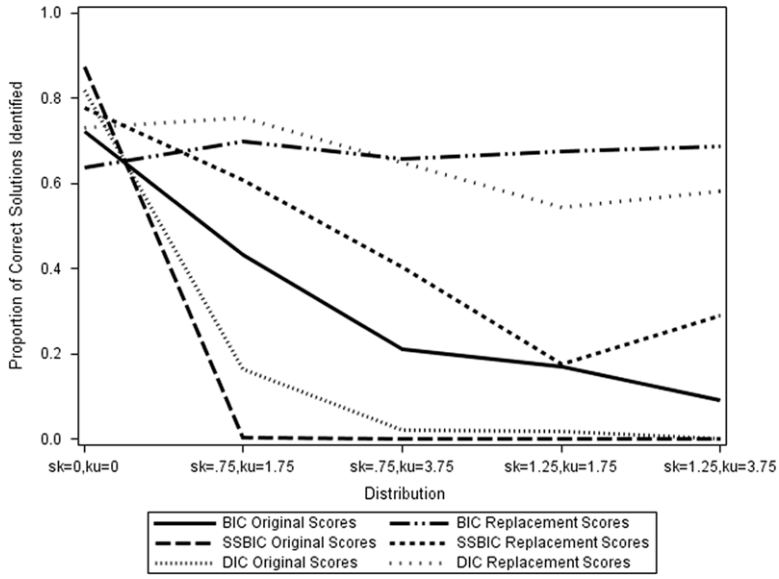

Fig. 2. Proportion of three-profile solutions selected by BIC, SSBIC, and DIC by indicator distribution.

4. Discussion

We investigated the performance of commonly used indices as well as several less commonly used indices of model fit in finite mixture models that vary by indicator distribution. Overall, the BIC-based indices (i.e., BIC, SSBIC, and DIC) tended to be the best performing indices in identifying three profiles out of the set of indices studied when the indicators were normally distributed. For example, when 10 or 15 indicator variables were used and profile prevalences were more balanced, each identified three profiles in nearly 100% of replications. They also tended to function well with one dominant profile and two rare profiles but needed more indicators and a larger sample size. One might expect that ICL-BIC would also perform well given its connection to the BIC. However, the entropy values within each cell were very similar regardless of the number of components. Therefore, the contribution of entropy to the ICL-BIC reduced its performance. Regardless, ICL-BIC tended to perform well with larger samples and more indicators. The penalty imposed on the log-likelihood by the DIC is less severe than BIC but more so than SSBIC. As a result, the performance of DIC was between BIC and SSBIC but tended towards SSBIC. The CAIC also performed very well with normally distributed indicators and larger samples. Furthermore, CAIC performed well in nonnormal conditions with a large number indicators and larger profile prevalences. Consistent with

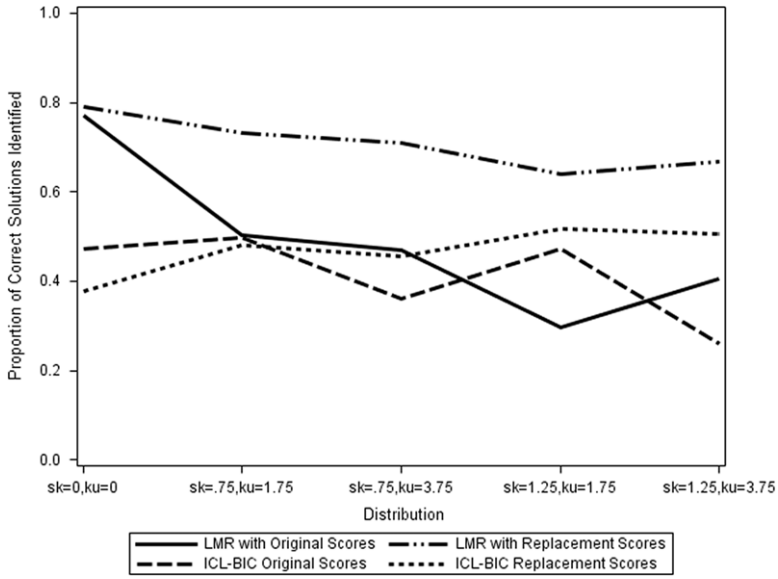

Fig. 3. Proportion of three-profile solutions selected by ICL-BIC and LMR likelihood ratio test by indicator distribution.

recommendations from other mixture modeling studies (Morgan, 2015; Nylund et al., 2007; Tofighi and Enders, 2007), BIC and/or SSBIC are good options for model section.

When indicators do not follow normal distributions, none of the indices functioned well for model selection. The negative effect nonnormality had on fit index performance was mitigated by using more indicators but not to a meaningful extent. Among the fit indices examined here, CAIC did occasionally reach a high frequency of three-profile model selection but only in select conditions. For example, in the most extreme nonnormal condition, the highest rate of three-profile model selection was returned by CAIC at 89.4% with 15 indicators, sample size of 400, and profile prevalences of 0.45 (π_1), 0.40 (π_2), and 0.15 (π_3). When sample size was increased to 1200, the frequency with which CAIC identified three profiles dropped to 0.2%, which provides support that fit indices used with nonnormal indicators may function well only under very specific conditions. Overall, it is not recommended that fit indices be used for model selection in the presence of nonnormal indicators.

When the joint-distributions of normally distributed indicators were transformed, the overall variance in the joint-distribution was reduced, which negatively affected the performance of some of the fit indices. The effect may be limited to certain situations. Other researchers may discover that normal score transformation increases the variance of the joint-distributions. It depends on the original joint-distribution. When the joint-distribution was comprised of nonnormal profiles, fit index performance was the same or better across all conditions. Under many conditions, the fit indices' ability to identify three profiles improved dramatically.

4.1. Recommendations for use of fit indices and normal score transformation in mixture modeling

With regard to the types and treatment of indicators used in mixture modeling, the findings support the use of certain fit indices when indicators are normally distributed and rare profiles are not hypothesized to underlie the data. Fit performance was greatly impacted by the degree of nonnormality of the indicators. In terms of profile prevalence, the presence of a dominant profile and a few very rare profiles were difficult to identify unless 15 normally distributed indicators are used along with a sample size of 1200. In general, the findings support researchers using fit indices to select a mixture model with the set of normally distributed indicators but not with nonnormally distributed indicators.

We should note that expectations based on one's guiding theory will greatly enhance the model selection process particularly as it relates to the prevalence of each profile. Prevalence was an important design factor. Whereas a researcher will be able to empirically examine the distributional form of the indicator variables, one's expectation of the prevalences of each underlying profile will be informed by her or his guiding theory. In studies investigating the presence and characteristics of rare profiles, one should confirm relatively normally distributed indicator variables as well as an adequate number of indicators (i.e., at least 10). When these conditions were met, SSBIC tended to function best. When rare profiles are not theorized and indicators are normally distributed, then BIC and/or SSBIC may be given more weight in model selection, particularly when there are at least 10 indicators. Both of these indices were quite effective at identifying the simulated number of profiles under such conditions. A moderate degree of separation between profiles was simulated. Fit performance is likely to improve with more separated components and decrease with components that are closer.

Theoretical expectation must also be considered with the use of normal score transformation. If nonnormal profiles are theorized, then the use of normal score transformation may provide considerable benefit if using fit indices to help with

model selection. The transformations never resulted in a decrease in fit index ability to identify the number of simulated profiles and often improved this ability greatly. Therefore, the use of normal scores may be an attractive option for some researchers. An important consideration for the use of normal score transformation is the model parameter estimates are no longer based on the original data although normal score transformation does preserve the order (i.e., relative position) of the data, which is frequently what is used for interpreting mixture model solutions.

4.2. Implications for subsequent analysis

In mixture modeling, it is common to use the components from selected model in subsequent analysis. For example, the components could be used as a predictor of some distal outcome of interest, or covariates and/or explanatory variables could be used to predict/explain component membership. Intuitively, identifying the "true" number of components may seem necessary for making correct inferences about relationships in subsequent analysis. Yet, this may not always be true. Model selection with nonnormally distributed components presents additional considerations with regard to the substantive nature of the nonnormality. If nonnormality is nonsubstantive, a model with more components may be beneficial because the additional, nonsubstantive components may absorb the nonnormality and ultimately purify the remaining, substantive components. Such purification may actually render subsequent analysis more informative. Hennig and Liao (2013) illustrate the potential drawbacks of using "automatic methods" (p. 331) for model selection, and their instructive "philosophy of clustering" discussion correctly reminds researchers to rely also on decisions they make themselves about the modeling process and to validate their results (p. 318). Hennig and Liao (2013) demonstrated that selecting more components when incorrectly assuming independence may result in more substantively meaningful components. This principle applies equally to our study regarding nonnormality. Hence, we do not mean to imply that fit indices that identify a different number of components than what was generated is necessarily bad or undesirable from a substantive perspective because that may actually be beneficial in practice. We focused solely on the statistical fit measures that are available to aid in model selection, and we strongly encourage researchers to consider alignment with one's guiding theory, meaningfulness, and interpretability.

4.3. Limitations

As with any Monte Carlo simulation study, the findings are generalizable to the conditions generated in the study. Although the selected design factors and conditions were informed by the conditions where mixture modeling was used in empirical research, the actual conditions studied will not be directly applicable to all research studies. Due to the label switching problem, accuracy of parameter estimates and correct classification of specific individuals were not included though it is recommended that future studies investigate these issues.

4.4. Future studies

As noted above, the findings may only be generalizable to study conditions that are similar to those conditions studied here. Additional research is needed to continue to develop our understanding of fit performance under various conditions that researchers are likely to encounter. First, the next study may examine the performance of another fit test, the bootstrapped likelihood ratio test, which has shown some promise in previous research (Nylund et al., 2007). Its ability to detect the true number of underlying classes with low separation and/or rare classes may be of interest to the field.

Modeling mixtures of nonnormal distributions is directly possible beginning with version 7.2 of M*plus* (Muthén and Muthén, 2014), which was released after this research was completed. Future studies may examine the frequency with which nonnormally distributed components may be identified by fit indices using the nonnormal distributions (e.g., *t*, skew-normal, and/or skew *t* distributions).

Future investigation is also warranted surrounding the parameter recovery and correct underlying group composition of mixture models. We have shown the frequent overestimation of profile enumeration. Over- and underextraction should be examined more closely with regard to its effect on decisions that result from selecting incorrect solutions. Related to our discussion in Section 4.2, future research may also examine how selecting the incorrect number of components affects subsequent analyses that incorporates the profiles as an independent variable. Given the extent to which nonnormality influenced fit performance in this study, future studies should also examine the relationship between separation and nonnormality. This study is an additional step in providing information that, when coupled with theoretical knowledge, may help researchers gain greater confidence using information from fit indices to select a mixture model.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2015.02.019.

References

Akaike, H., 1977. On the Entropy Maximization Principle. North-Holland, Amsterdam, pp. 27–41.
Bacci, S., Pandolfi, S., Pennoni, F., 2014. A comparison of some criteria for states selection in the latent Markov model for longitudinal data. Adv. Data Anal. Classif. 8, 125–145.

Banfield, J.D., Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics 49 (3), 803–821.
Bartolucci, F., Farcomeni, A., Pennoni, F., 2013. Latent Markov Models for Longitudinal Data. Chapman and Hall/CRC, Boca Raton, FL.
Bauer, D.J., Curran, P.J., 2004. The integration of continuous and discrete latent variable models: potential problems and promising opportunities. Psychol. Methods 9 (1), 3.
Biernacki, C., Celeux, G., Govaert, G., 1998. Assessing a mixture model for clustering with the integrated classification likelihood. Technical report, Rhone-Alpes.
Bollen, K.A., 1989. Structural Equation with Latent Variables. John Wiley & Sons, Inc., New York.
Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52 (3), 345–370.
Celeux, G., Soromenho, G., 1996. An entropy criterion for assessing the number of clusters in a mixture model. J. Classification 13 (2), 195–212.
Clogg, C.C., 1995. Latent Class Models. Plenum, New York, pp. 311–360.
Collins, L.M., Lanza, S.T., 2010. Latent Class and Latent Transition Analysis. John Wiley & Sons, Inc., Hoboken, NJ.
Dolan, C.V., van der Maas, H.L.J., 1998. Fitting multivariate normal finite mixtures subject to structural equation modeling. Psychometrika 63 (3), 227–253.
Draper, D., 1995. Assessment and propagation of model uncertainty. J. Roy. Statist. Soc. Ser. B 45–97.
Everitt, B.S., 1981. A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. Multivariate Behav. Res. 16 (2), 171–180.
Everitt, B.S., 1993. Cluster Analysis, third ed. John Wiley & Sons, Inc., New York.
Fleishman, A.I., 1978. A method for simulating non-normal distributions. Psychometrika 43 (4), 521–532.
Flora, D.B., Curran, P.J., 2004. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis. Psychol. Methods 9, 466–491.
Fonseca, J.R.S., Cardoso, M.G.M.S., 2005. Retails Clients Latent Segments. Springer-Verlag, Covilhã, Portugal, pp. 348–358.
Forero, C.G., Maydeu-Olivares, A., Gallardo-Pujol, D., 2009. Factor analysis with ordinal indicators: a Monte Carlo study comparing DWLS and ULS estimation. Struct. Equ. Model. 16, 625–641.
Gordon, A.D., 1981. Classification. Chapman and Hall, New York.
Hallquist, M., 2011. Mplusautomation: automating Mplus model estimation and interpretation, URL: http://CRAN.R-project.org/package=MplusAutomation. R version 0.4-2.
Heinen, T., 1996. Latent Class and Discrete Latent Trait Models. Sage, Thousand Oaks, CA.
Hennig, C., Liao, T.F., 2013. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. J. Roy. Statist. Soc. Ser. C 62 (3), 309–369.
Henson, J.M., Reise, S.P., Kim, K.H., 2007. Detecting mixtures from structural model differences using latent variable mixture modeling: a comparison of relative model fit statistics. Struct. Equ. Model. 14 (2), 202–226.
Hunt, L., Jorgensen, M., 2003. Mixture model clustering for mixed data with missing information. Comput. Statist. Data Anal. 41 (3), 429–440.
Hurvich, C.M., Tsai, C., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.
Jeffries, N.O., 2003. Testing the number of components in a normal mixture. Biometrika 90 (4), 991–994.
Koehler, A.B., Murphree, E.S., 1988. A comparison of the akaike and schwarz criteria for selecting model order. Appl. Stat. 41, 187–195.
Lee, S.X., McLachlan, G.J., 2013. On mixtures of skew normal and skew t-distributions. Adv. Data Anal. Classif. 7 (3), 241–266.
Lo, Y., Mendell, N.R., Rubin, D.B., 2001. Testing the number of components in a normal mixture. Biometrika 88 (3), 767–778.
Lubke, G., Neale, M.C., 2006. Distinguishing between latent classes and continuous factors: resolution by maximum likelihood? Multivariate Behav. Res. 41 (4), 499–532.
McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. M. Dekker, New York.
McLachlan, G.J., Ng, S.K., 2000. A comparison of some information criteria for the number of components in a mixture model. Technical report. Department of Mathematics, University of Queensland, Brisbane, Australia.
McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, Inc., New York.
Milligan, G.W., Cooper, M.C., 1988. A study of standardization of variables in cluster analysis. J. Classification 5 (2), 181–204.
Morgan, G.B., 2015. Mixed mode latent class analysis: An examination of fit index performance for classification. Struct. Equ. Model. 22, 76–86.
Murray, P.M., Browne, R.P., McNicholas, P.D., 2014. Mixtures of skew-t factor analyzers. Comput. Statist. Data Anal. 77, 326–335.
Muthén, B.O., 2001. LCA and cluster analysis. Message posted to MPLUS discussion list, December 11 archived at: http://www.statmodel.com/discussion/messages/13/155.html?1077296160.
Muthén, B.O., 2004. Mplus Technical Apendices. Muthén & Muthén, Los Angeles, CA, version 3 edition.
Muthén, B.O., Muthén, L.K., 2000. Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. Alcohol. Clin. Exp. Res. 24 (6), 882–891.
Muthén, L.K., Muthén, B.O., 2010. Mplus: User's Guide, sixth ed. Muthén & Muthén, Los Angeles, CA.
Muthén, L.K., Muthén, B.O., 2014. Mplus: User's Guide, seventh ed. Muthén & Muthén, Los Angeles, CA.
Nylund, K.L., Asparouhov, T., Muthén, B.O., 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Struct. Equ. Model. 14 (4), 535–569.
Pastor, D.A., Barron, K.E., Miller, B.J., Davis, S.L., 2007. A latent profile analysis of college students' achievement goal orientation. Contemp. Educ. Psychol. 32 (1), 8–47.
Ramaswamy, V., DeSarbo, W.S., Reibstein, D.J., Robinson, W.T., 1993. An empirical pooling approach for estimating marketing mix elasticities with PIMS data. Mark. Sci. 12 (1), 103–124.
R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. (ISSN: 00905364) 6 (2), 461–464.
Sclove, S.L., 1987. Application of model-selection criteria to some problems in multivariate analysis. Psychometrika 52 (3), 333–343.
Soromenho, G., 1994. Comparing approaches for testing the number of components in a finite mixture model. Comput. Statist. 9 (4), 65–82.
Tofighi, D., Enders, C.K., 2007. Identifying the Correct Number of Classes in Growth Mixture Models. Information Age Publishing, Inc., Greenwich, CT, pp. 317–341.
Vermunt, J.K., 2004. The Sage Encyclopedia of Social Sciences Research Methods. Sage Publications, Thousand Oaks, CA, pp. 554–555. chapter Latent profile model.
Vermunt, J.K., Magidson, J., 2002. Latent Class Cluster Analysis. Cambridge University Press, Cambridge, MA, pp. 89–106.
Vrbik, I., McNicholas, P.D., 2014. Parsimonious skew mixture models for model-based clustering and classification. Comput. Statist. Data Anal. 71, 196–210.
Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333.
Yang, C., 2006. Evaluating latent class analysis models in qualitative phenotype identification. Comput. Statist. Data Anal. (ISSN: 0167-9473) 50 (4), 1090–1104. http://dx.doi.org/10.1016/j.csda.2004.11.004.
Yang-Wallentin, F., Jöreskog, K.G., Luo, H., 2010. Confirmatory factor analysis of ordinal variables with misspecified models. Struct. Equ. Model. 17 (3), 392–423.