PERFORMANCE OF INFORMATION CRITERIA FOR MODEL SELECTION IN A LATENT GROWTH CURVE MIXTURE MODEL

Satoshi Usami*

ABSTRACT

Novel simulation studies are performed to investigate the performance of likelihood-based and entropy-based information criteria for estimating the number of classes in latent growth curve mixture models, considering influences of true model complexity and model misspecification. Simulation results can be summarized as (1) Increased model complexity worsens the performance of all criteria, and this is salient in Bayesian Information Criteria (BIC) and Consistent Akaike Information Criteria (CAIC). (2) The classification likelihood information criterion (CLC) and integrated completed likelihood criterion with BIC approximation (ICL.BIC) frequently underestimate the number of classes. (3) Entropy-based criteria correctly estimate the number of classes more frequently. (4) When a normal mixture is incorrectly fit to non-normal data including outliers, although this seriously worsens the performance of many criteria, BIC, CAIC, and ICL.BIC are relatively robust. Additionally, overextracted classes with trivially small mixture proportions can be detected when the sample size is large. (5) When there is an upper bound of measurement, although this worsens the performance of almost all criteria, entropy-based criteria are robust. (6) Although no single criterion is always best, ICL.BIC shows better performance on average.

1. Introduction

In the behavioral sciences, latent growth curve models (LGCMs) (McArdle, 1988; Meredith & Tisak, 1984, 1990) have become a standard tool to model repeated measures data, and many extensional models have been developed for applied research purposes (for details, see Bollen & Curran, 2004, 2006; McArdle, 2009). LGCMs are mathematically subsumed into structural equation modeling (SEM) or, more generally, into generalized linear latent and mixed models (GLLAMMs) (Rabe-Hesketh, Skrondal, & Pickles, 2004; Skrondal & Rabe-Hesketh, 2004, 2007), thereby enabling us to effectively use various SEM techniques such as multi-group analysis methods and model fit indices.

One reason for the recent widespread use of LGCMs in the behavioral sciences is undoubtedly their extension to finite mixtures (McLachlan & Peel, 2000), that is, latent growth curve mixture models (LGCMMs: e.g., Bauer & Curran, 2003; Leiby, Sammel, Ten Have, & Lynch, 2009; Nagin, 1999; Neelon, Swamy, Burgette, & Miranda, 2011; Song, Lee & Hser, 2009). LGCMMs assume population heterogeneity behind repeated measures data, and enable us to detect multiple developmental trajectories that differ in mean and covariance structure, and also allow use of profile analysis (i.e., nominal logistic regression analysis) to investigate the relationship between the classification result to classes and subject profiles.

In applying LGCMMs, a difficult but sometimes intriguing question is estimation of the number of classes. Specifically, how many classes should researchers extract from the repeated measures...
data? Unfortunately, the likelihood ratio test, which is one of the most famous model selection tools, cannot be used to estimate the number of classes, because regularity conditions are not satisfied among models with different numbers of classes (e.g., Bauer & Curran, 2003, 2004; Henson, Reise, & Kim, 2007). As a result, although several methods such as the Vuong-Lo-Mendell-Rubin (VLMR) test (Lo, Mendell, & Rubin, 2001), parametric bootstrap method (McLachlan & Peel, 2000) and the skewness and kurtosis (SK) test (Muthén & Muthén, 1998–2010) have been developed to estimate the number of classes, using information criteria seems to be the most common procedure in actual studies because of its familiarity and utility (e.g., Henson et al., 2007; Nylund, Asparouhov & Muthén, 2007; Vrieze, 2012).

As Henson et al. (2007) clearly summarized, in finite mixture modeling several information criteria can be classified into likelihood-based statistics and entropy-based statistics. As introduced in the next section, famous information criteria such as Akaike Information Criterion (AIC) (Akaike, 1973) and Bayesian Information Criterion (BIC) (Schwarz, 1978) are classified as likelihood-based statistics. Even if they differ in their exact definition of a "good model", different information criteria do have the same aim of identifying good models (Acquah, 2010), and then the performance of these criteria for model selection has been compared in the mixture-modeling context (e.g., Bauer & Curran, 2003; Henson et al, 2007; Lubke & Neale, 2006; Nylund et al, 2007; Vrieze, 2012). Vrieze (2012) reviewed the features of AIC and BIC, and compared the performance of the AIC and BIC by a novel simulation study and showed that BIC outperformed the AIC when (i) the true model is among the candidate models considered, (ii) the true model is simple (i.e., the small number of classes), (iii) degree of separation (i.e., mean differences among classes) is large, (iv) sample size is large, (v) mixture proportion is not extreme (i.e., class sizes are not very disparate). And so far, few studies have compared the performance of entropy-based statistics. Exceptionally, Henson et al (2007) performed a large simulation study that considers both likelihood-based statistics and entropy-based statistics, and noted that the Classification Likelihood information Criterion (CLC; McLachlan & Peel, 2000) and Integrated Completed Likelihood criterion with BIC approximation (ICL.BIC; Biernacki, Celeux & Govaert, 2000) both performed well in distinguishing two- and three-class models, although they cannot be used to compare models to a one-class model.

These findings were undoubtedly useful in actual research, but several important points have not been clearly investigated so far. The first is the sensitivity of each information criterion to model misspecification. One of the most popular model misspecifications is to fit normal-mixtures to data that do not satisfy normality within classes (e.g., Bauer & Curran, 2003, 2004). We often face such data in actual analysis due to various reasons such as features of variables and measurement methods. As for measurement methods, for example in gerontological research that intends to classify the elderly by memory test scores, test scores may be left-censored because the memory test consists of relatively difficult items. However, even in this case many applied researchers seem to prefer normal-mixtures for classification because normal-mixture modeling is undoubtedly highly popular among mixture modeling methods and this procedure is very convenient in most statistical software packages. The use of normal-mixtures can be validated if extracted classes pragmatically reflect the true structure of developmental trajectories well, even if classes are overestimated or a part of estimated trajectories are biased. Bauer and Curran (2003,2004) technically noted that the number of extracted classes and estimates of parameters are sensitive to model misspecification, and compared the performance of both likelihood-based and entropy-based statistics by generating non-normal data under various conditions of skewness and kurtosis of univariate variable. However, this simulation study is very limited in its variety of the model complexities, degrees of separation, mixture proportions and types of misspecification.

The second problem relates to the performance of entropy-based information criteria. Although
the results in Henson et al (2007) are undoubtedly useful, unfortunately their research did not focus on several conditions including sample size and model complexity. Especially, as for model complexity, the true number of classes is limited to two in Henson et al (2007) and this setting prevents differentiation between underestimation and correct estimation in entropy-based information criterion since they cannot be used to compare models to a one-class model.

The present research therefore describes two simulation studies. In the first simulation, we investigate the performance of both likelihood-based and entropy-based information criteria for estimating the unknown number of classes in LGCMs under different sample sizes, degrees of separation, mixture proportions, and the true number of classes. Investigations focusing on LGCM are required considering the fact that many applied researchers have shown interest in the use of mixture modeling for longitudinal data analysis, while in Henson et al (2007), the mixture of factor analysis model was considered. There models are formally different, but LGCM is based on the confirmatory factor analysis model, so the numbers of parameters (and complexity of the model) do not largely differ between models and comparing the result of current simulation result with that of Henson et al (2007) is reasonable. In the second simulation, we mainly investigate the influence of model misspecification. Namely the normal-mixture models are arbitrarily fitted to data that fail to satisfy normality within classes by generating outliers and censored data. As a data generation model, a two-variate latent change score model (Hamagami & McArdle, 2007; McArdle & Nesselroade, 1994; McArdle, 2001; McArdle & Hamagami, 2001; McArdle et al., 2004; McArdle, 2009), which is a generalized expression of the two-variate standard LGCM as shown later and has almost the similar model complexity with factor analysis model, is used. Latent change score models recently have attracted much attention due to their utility in investigating influential relationships among variables (see Section 3). An additional feature of this model is its generality and ability to express various developmental trajectories, making finite mixtures in latent change score models a promising procedure for future applications in behavioral research (e.g., McArdle, 2009).

This paper is organized as follows. Section 2 and Section 3 introduce the two-variate latent change score model and various information criteria. Section 4 and Section 5 address the procedures and results of two simulation studies. An actual data example is provided in Section 6. The final section presents a discussion and directions for future research.

2. Latent Change Score Model

Here we refer to the explanation in McArdle et al (2004, p298–299) and McArdle (2009). Let \( y_{jt} \) and \( x_{jt} \) be the observed variables at occasion \( t (1, \ldots, T) \) for each person \( j (1, \ldots, J) \) who belongs to the \( c (1, \ldots, C) \)th latent class, respectively. In the finite mixtures in a latent change score model, firstly \( y_{jt} \) and \( x_{jt} \) are decomposed into the latent true scores (\( f_{yjt} \) and \( f_{xjt} \)) and errors (\( \epsilon_{yjt} \) and \( \epsilon_{xjt} \)). Namely,

\[
y_{jt} = f_{yjt} + \epsilon_{yjt}, \quad x_{jt} = f_{xjt} + \epsilon_{xjt}.
\]

(1)

Here, \( \epsilon_{yjt} \sim N(0, \psi_y^2) \) and \( \epsilon_{xjt} \sim N(0, \psi_x^2) \) are assumed, respectively. Assuming that \( y_{jt} \) and \( x_{jt} \) are measured over a defined interval of time, \( f_{yjt} \) and \( f_{xjt} \) are additionally expressed via latent change scores (\( \Delta f_{yjt} \) and \( \Delta f_{xjt} \)) as follows:

\[
f_{yjt} = f_{yj(t-1)} + \Delta f_{yjt}, \quad f_{xjt} = f_{xj(t-1)} + \Delta f_{xjt},
\]

(2)

for \( t \geq 2 \). For clarity, in the discussion below we use notations \( f_{yj} \) and \( f_{xj} \) as intercept factor scores that indicate \( f_{yj} = f_{yjt} \) and \( f_{xj} = f_{xjt} \), respectively. Finally, \( \Delta f_{yjt} \) and \( \Delta f_{xjt} \) are modeled by the linear form
\[
\Delta_{xy}^{\Delta} = S_{xy}^{\Delta} + \beta_{xy}^{\Delta} f_{xy(t-1)}^{\Delta} + \gamma_{xy}^{\Delta} f_{xy(t-1)}^{\Delta} + \delta_{xy}^{\Delta}, \quad \Delta_{xy}^{c} = S_{xy}^{c} + \beta_{xy}^{c} f_{xy(t-1)}^{c} + \gamma_{xy}^{c} f_{xy(t-1)}^{c} + \delta_{xy}^{c} \quad (3)
\]

Here, \(S_{xy}^{\Delta}\) and \(S_{xy}^{c}\) are factor scores that express constant change. So, these are generally referred to as slope factor scores, though in the equations \(S_{xy}^{\Delta}\) and \(S_{xy}^{c}\) work as intercepts for latent change scores \(\Delta_{xy}^{\Delta}\) and \(\Delta_{xy}^{c}\). For \(\xi_{c} = (f_{xy}^{c}, S_{xy}^{c}, f_{xy}^{c}, S_{xy}^{c})^{\prime}\), \(\xi_{c} \sim N(\psi, \Phi_{c})\) are assumed, where

\[
\psi = \begin{pmatrix}
\mu_{0}^{x} \\
\mu_{0}^{y} \\
\mu_{1}^{x} \\
\mu_{1}^{y}
\end{pmatrix}, \quad \Phi_{c} = \begin{pmatrix}
\phi_{11} & \phi_{12} & \phi_{13} \\
\phi_{21} & \phi_{22} & \phi_{23} \\
\phi_{31} & \phi_{32} & \phi_{33}
\end{pmatrix}
\]

\(\mu_{x}, \mu_{y}, \mu_{0}, \mu_{1}\) and \(\psi_{x}, \psi_{y}, \psi_{0}, \psi_{1}\) denote the initial level means and the averages of constant change for participants in the \(c\)th class for their respective variables. And diagonal elements of \(\Phi_{c}\) (i.e., \(\phi_{11}, \phi_{22}, \phi_{33}\)) represent the corresponding variances. \(\phi_{xy}, \phi_{yx}\) are autoregressive parameters for describing proportional changes of respective variables for the \(c\)th class. \(\phi_{xy}, \phi_{yx}\) are coupling parameters for the \(c\)th class to describe loading of the factor scores \(f_{xy}, f_{yx}\) to subsequent changes in the other factor scores \(f_{yx}, f_{xy}\), respectively. Therefore, coupling parameters play an important role in investigating influential relationships among variables. Finally, \(\delta_{xy}, \delta_{yx}\) indicate disturbances and \(\delta_{xy} \sim N(0, \omega_{xy}^{2}), \delta_{yx} \sim N(0, \omega_{yx}^{2})\) are assumed, respectively. To promote understanding, a path diagram of a latent change score model (for the case \(T = 4\)) is shown in Figure 1. From model expressions (1)–(3), it can be noted that the latent change score model is an extension of an autoregressive model that includes covariates and uses latent variables, and that this model reduces to a standard (multivariate) latent growth curve model when both autoregressive parameters and coupling parameters are equal to 0 under reparameterized error and residual variances.

If each person’s latent class membership information is completely given, observed data \(z_{j} = (y_{j1}, x_{j1})'\) have a conditional multinormal distribution whose mean and variance-covariance correspond to the given class, thus mathematically this equates to the expression of multi-group analysis (e.g., Hoshino, 2001). Actually, each participant is probabilistically assigned to each class, and \(z_{j}\) follows a finite mixture of normal distributions with unconditional form. That is,

\[
z_{j} \sim \sum_{c=1}^{C} p_{c} N\left(\mu_{c}(\Theta), \Sigma_{c}(\Theta)\right). \quad (5)
\]

\(\mu_{c}(\Theta), \Sigma_{c}(\Theta)\) are means and covariances for the \(c\)th latent class, and \(p_{c}\) (\(p = (p_{1}, p_{2}, \ldots, p_{C})\)) is the unknown mixture proportion of the membership for the \(c\)th latent class such that \(p_{c} > 0\) and \(\sum_{c=1}^{C} p_{c} = 1\). Details of \(\mu_{c}(\Theta)\) and \(\Sigma_{c}(\Theta)\) using matrix notation are provided in Usami (2012). From equation (5) the likelihood function \(L(\Theta | Z)\) for data \(Z = (X', Y')'\) randomly drawn from the mixture can be expressed as

\[
L(\Theta | Z) = \prod_{j=1}^{J} \left( \sum_{c=1}^{C} p_{c} (2\pi)^{-C/2} | \Sigma_{c}^{-1/2} \right) \exp \left[ -\frac{1}{2} (z_{j} - \mu_{c})\prime (\Sigma_{c})^{-1} (z_{j} - \mu_{c}) \right]. \quad (6)
\]

Here, \(X = (x_{1}, x_{2}, \ldots, x_{J})'\) and \(Y = (y_{1}, y_{2}, \ldots, y_{J})'\). When \(z_{j}\) follows a finite mixture of normal distributions expressed in equation (5), mean and variance-covariance structure for the whole data set (expressed as \(\mu_{z}(\Theta)\) and \(\Sigma_{z}(\Theta)\), respectively) can be expressed as

\[
\mu_{z}(\Theta) = \sum_{c=1}^{C} p_{c} \mu_{c}(\Theta), \quad \Sigma_{z}(\Theta) = \sum_{c=1}^{C} p_{c} p_{c'} (\mu_{c}(\Theta) - \mu_{c'}(\Theta)) (\mu_{c'}(\Theta) - \mu_{c'}(\Theta))' + \sum_{c=1}^{C} p_{c} \Sigma_{c}(\Theta)
\]

by using \(\mu_{c}(\Theta)\) and \(\Sigma_{c}(\Theta)\) (Bauer & Curran, 2004).
3. Definition of information criteria

In this section we briefly introduce the information criteria used in this study, referring to Bauer and Curran (2003), Henson et al. (2007) and Vrieze (2012).

3.1. (i) likelihood-based information criteria

Information criteria classified in this category are generally defined as taking the form a trade-off between fit (i.e., log-likelihood) and model complexity (i.e., the number of parameters). The general form can be expressed as

$$-2 \log L(\hat{\Theta}|Z) + \alpha(J)k$$

where $\hat{\Theta}$ is the maximum likelihood estimate, and $\log L(\hat{\Theta}|Z)$ is the log of the likelihood given the data $Z$, and $k$ is the total number of estimated model parameters. $\alpha(J)$ is a penalty coefficient that...
may depend on total sample size $J$. AIC and BIC are defined as:

$$AIC = -2 \log L(\hat{\Theta}|Z) + 2k,$$

$$BIC = -2 \log L(\hat{\Theta}|Z) + k \log J.$$  

Obviously, $\alpha(J) = 2$ and $\alpha(J) = \log J$ in AIC and BIC, respectively. AIC is an estimator of mean expected log-likelihood and it aims to find the best approximating model to the unknown data generating process. BIC relates to a first-order approximation of Bayes factor given data $Z$ and is designed to identify the true model (e.g., Lee, 2007). When $J \geq 8$, BIC favors a simpler model than AIC. The performance of these two famous information criteria has been theoretically and empirically compared so far (e.g., Kuha, 2004; Vrieze, 2012), and Vrieze (2012) summarized statistical literature on these criteria and noted as follows:

(a) Asymptotically, the BIC is consistent, in that it will select the true model if, among other assumptions, the true model is among the candidate models considered. The AIC is not consistent under these circumstances.

(b) When the true model is not in the candidate model set the AIC is efficient, in that it will asymptotically choose whichever model minimizes the mean squared error of prediction/estimation. The BIC is not efficient under these circumstances.

(c) Unlike the BIC, the AIC also has a minimax property, in that it can minimize the maximum possible risk in finite sample sizes.

(Vrieze, 2012, p228)

Although BIC seems to be most widely used and recommended among several information criteria (e.g., Jedidi et al., 1997; Leiby et al., 2009; Nagin, 1999; Nylund et al., 2007; Steinley & Brusco, 2011), there are now various types of information criteria, and different information criteria have quite different properties that require different assumptions. So, no single information criterion is always best (e.g., Henson et al, 2007), and applied researchers and methodologists alike will benefit from improved understanding of the asymptotic and finite-sample behavior of various information criteria according to features of specified models and obtained data (Vrieze, 2012).

Numerous information criteria have been proposed based on either AIC or BIC so far, and these examples include consistent AIC (CAIC; Bozdogan, 1987) and adjusted BIC (aBIC, also referred to as sBIC as an abbreviation of Sample Size adjusted BIC (Sclove, 1987)). CAIC is an information criterion under which AIC satisfies asymptotical consistency. CAIC is defined as

$$CAIC = -2 \log L(\hat{\Theta}|Z) + k(\log J + 1),$$

and as a result the difference between BIC and CAIC is just sample size (i.e., $\alpha(J) = \log J + 1$). aBIC is defined by setting $\alpha(J) = \log \frac{J+2}{24}$. Namely,

$$aBIC = -2 \log L(\hat{\Theta}|Z) + k \log \frac{J+2}{24}.$$  

aBIC is based on the Rissanen Information Criteria for autoregressions (Henson et al, 2007).

3.2. (ii) Entropy-based information criteria

Entropy-based information criteria are characteristic to mixture modeling because the ambiguity of assignment to clusters is highly related to model evaluation. For example, if there are two
different models that show similar model complexity and log-likelihood, likelihood-based information criteria values should also be close between models, leading to difficulty in choosing the better model. In entropy-based information criteria two such models can show very different information criteria values, if one model can classify subjects with less ambiguity (the assignment probability is near 0 or 1 in each person) while the other model can not.

Celeux and Soromenho (1996) therefore considered that the log-likelihood of a model $\log L(\Theta|Z)$ can be decomposed as

$$\log L(\Theta|Z) = \log L_c(\hat{\Theta}|Z) + EN(\hat{\tau}).$$

(13)

$\log L_c(\hat{\Theta}|Z)$ is classification maximum likelihood (CML) that denotes likelihood in the case where classification is conducted perfectly (i.e., all subjects are classified to classes with probabilities 0 or 1, with no ambiguity). And the entropy $EN(\hat{\tau})$ indicates the ambiguity of the actual assignment. Here, $EN(\hat{\tau})$ is defined as

$$EN(\hat{\tau}) = -\sum_{j=1}^{J} \sum_{c=1}^{C} \hat{\tau}_{jc} \log \hat{\tau}_{jc},$$

(14)

where $\hat{\tau}_{jc}$ is an estimate of posterior probability for subject $j$ in latent class $c$. Here, $C$ is the number of latent classes. As is obvious in this equation, as assignment is recovered well, $\hat{\tau}_{jc}$ takes values near 0 or 1 and then $EN(\hat{\tau})$ closes to 0.

The classification likelihood information criterion (CLC; McLachlan & Peel, 2000) is an information criterion derived directly from the previous equation as

$$CLC = -2 \log L(\Theta|Z) + 2EN(\hat{\tau}).$$

(15)

Entropy is another entropy-based information criterion proposed by Ramaswamy, DeSarbo, Reibstein and Robinson (1993) defined as the ratio of observed $EN(\hat{\tau})$ to the maximum possible values of $EN(\hat{\tau})$ (i.e., $J \log C$):

$$Entropy = 1 - \frac{EN(\hat{\tau})}{J \log C}.$$ 

(16)

In Entropy, the model that takes the largest value is selected.

 Celeux and Soromenho (1996) proposed the normalized entropy criterion (NEC) by considering the ratio of $EN(\hat{\tau})$ to the difference between log-likelihoods for models ($C \geq 2$) under evaluation and the model with $C = 1$ as:

$$NEC = \frac{EN(\hat{\tau})}{\log L(\Theta|Z) - \log L'(\Theta|Z)}.$$ 

(17)

Here, $\log L'(\Theta|Z)$ is a log likelihood for the model with $C = 1$. As Biernacki, Celeux and Govaert (1999) noted, the value of NEC is sometimes defined as 1 when a single class is assumed (models with $C = 1$).

Lastly, the integrated completed likelihood criterion with BIC approximation (ICL.BIC; Biernacki, Celeux & Govaert, 2000) is an information criterion defined by the following equation;

$$ICL.BIC = -2 \log L(\Theta|Z) + k \log J + 2 EN(\hat{\tau}).$$

(18)

ICL.BIC is a combined form of BIC and CLC with penalized log-likelihood through sample size, the number of parameters, and $EN(\hat{\tau})$. Therefore, this information criterion tends to favor simpler models than BIC and CLC.
Entropy-based information criteria have advantages in that they can include entropy $EN(\hat{\tau})$ for model selection, although this value cannot be defined when a single class is assumed (models with $C = 1$; Henson et al., 2007). Thus, except for NEC, the use of these entropy-based information criteria is restricted to models with $C \geq 2$.

4. Simulation Study I (no model misspecification)

This section describes the procedures and results of the simulation study to investigate the performance of information criteria when model misspecification has not occurred.

4.1. Simulation design

Data generation model

A two-variate latent change score model is used as data generation model in this study.

True number of classes and upper bounds for the number of classes estimated

In this simulation, we set two conditions (a) and (b) for the true number of classes and an upper bound for the number of classes estimated. In a condition (a), these values were set as 2 and 3, respectively, due to the convenience of interpreting the influence of factors such as mixture proportion and degree of separation. This design is the same as that of Henson et al. (2007). For condition (b), these values were set as 3 and 4, respectively. This condition is more desirable in evaluating the performance of entropy-based information criteria because such information criteria cannot be defined under a model with a single class.

Specifying parameters and degree of separation among classes

We set five conditions for the degree of separation by setting different effect sizes (a value equal to the mean difference between classes divided by equal standard deviations on the first occasion) among classes. More precisely, equal covariance structures are assumed among classes, and only the mean structure is set differently among classes. This equal variance-covariance structure assumption is often effective in the actual analysis (e.g., Preacher, Wichman, MacCallum, & Briggs, 2008) because it serves as a preliminary tool for specifying the alternative model for subsequent analysis with a lower risk of a local optimal solution. While this assumption may be too strong in actual behavioral data, it is effective in this simulation because we can more clearly and concisely evaluate the performance of each information criterion.

Specifically, under condition (a), we set model parameters as $\mu_{1x}^1 = 10, \mu_{3x}^1 = 0, \mu_{1y}^1 = 10, \mu_{3y}^1 = 0, \gamma_i^1 = \gamma_i^2 = 0.112, \phi_{11}^1 = \phi_{21}^1 = \psi_{11}^2 = \psi_{21}^2 = 0.6, \beta_i^1 = \beta_i^2 = \beta_{y}^2 = 0.224$, $\omega_{y1}^1 = \omega_{y2}^1 = \omega_{x1}^2 = \omega_{x2}^2 = 0.1$, and

$$\Phi^1 = \Phi^2 = \begin{pmatrix} 0.400 & 0.035 & 0.160 & 0.017 \\ 0.035 & 0.075 & 0.017 & 0.015 \\ 0.160 & 0.017 & 0.400 & 0.035 \\ 0.017 & 0.015 & 0.035 & 0.075 \end{pmatrix}.$$ 

Additionally, we set parameters regarding mean vectors in the second class as $\mu_{1x}^2 = 10 + 0.5Q$, $\mu_{3x}^2 = 2, \mu_{1y}^2 = 10 + 0.5Q, \mu_{3y}^2 = 2$ by setting different $Q(1, \cdots, 5)$ to control the degree of separation between classes. Because the variance of observed variables $x$ and $y$ on the first occasion is $\phi_{11}^1 + \psi_{11}^2 = 1$, the different values of $Q(1, \cdots, 5)$ correspond to the different effect sizes 0.5, 1.0, $\cdots$, 2.5 on the first occasion. This parameter setting implicitly assumes that data are sampled with high reliability on each occasion (i.e., error and residual variance are small relative to the variance of observed variables) and generates nonlinear developing trajectories. Under condition (b), the same parameter...
values were set for the respective classes, except for the mean vector parameters. Namely, we set $\mu^0_{1x} = 10$, $\mu^1_{1y} = 10$, $\mu^0_{2y} = 0$, $\mu^1_{2y} = 10 + 0.5Q$, $\mu^2_{3y} = 0$, $\mu^1_{3y} = 10 + 2 \times 0.5Q$, $\mu^3_{3y} = 0$, $\mu^1_{4y} = 10 + 2 \times 0.5Q$ and $\mu^3_{3y} = 0$ with different values of $Q$.

Table 1 provides variance-covariance (and correlation) matrices of the whole data set $\Sigma(\Theta)$ calculated from equation (7) for the cases of $Q = 1.5$ and $p = (p_1, p_2) = (0.5, 0.5)$ in condition (a). As is shown in these matrices, parameters specified in this simulation reflect typical longitudinal data such as physical function and intellectual ability where variances of variables increase as time goes by and the correlations indicate medium to large values. Additionally, the amount of change of variances becomes larger as time goes by, and correlations become larger for pairs of later variables. Variance-covariance matrices that indicate these features can be observed in many longitudinal data. For example, as is expressed in Bollen and Curran (2006, p.69), the results of the National Longitudinal Survey of Youth (NLSY) shows a similar variance-covariance matrix for data on children’s reading recognition subscore of the Peabody Individual Achievement Test (PIAT).

Table 1: Variance-covariance matrices of whole data calculated from model parameters.

- **(a) $Q=1$ and $(p_1, p_2) = (0.5, 0.5)$ in condition (a).**

$$
\Sigma(\Theta) = \\
\begin{pmatrix}
1.062 & 0.625 & 0.838 & 1.116 & 0.222 & 0.340 & 0.503 & 0.728 \\
0.476 & 1.623 & 1.430 & 1.959 & 0.341 & 0.528 & 0.799 & 1.180 \\
0.489 & 0.675 & 2.767 & 3.011 & 0.504 & 0.799 & 1.216 & 1.807 \\
0.487 & 0.691 & 0.814 & 4.946 & 0.727 & 1.178 & 1.806 & 2.683 \\
0.209 & 0.260 & 0.294 & 0.317 & 1.063 & 0.626 & 0.838 & 1.118 \\
0.259 & 0.326 & 0.377 & 0.416 & 0.477 & 1.622 & 1.431 & 1.958 \\
0.293 & 0.377 & 0.439 & 0.488 & 0.489 & 0.675 & 2.768 & 3.012 \\
0.318 & 0.416 & 0.489 & 0.542 & 0.487 & 0.691 & 0.814 & 4.946
\end{pmatrix}
$$

*Elements in the lower triangular matrix exhibit corresponding correlation coefficients.

- **(b) $Q=5$ and $(p_1, p_2) = (0.5, 0.5)$ in condition (a).**

$$
\Sigma(\Theta) = \\
\begin{pmatrix}
2.560 & 2.625 & 3.510 & 4.682 & 1.721 & 2.344 & 3.175 & 4.294 \\
0.792 & 4.293 & 4.998 & 6.721 & 2.343 & 3.201 & 4.367 & 5.941 \\
0.799 & 0.879 & 7.532 & 9.374 & 3.180 & 4.370 & 5.979 & 8.168 \\
0.798 & 0.885 & 0.932 & 13.440 & 4.300 & 5.946 & 8.166 & 11.713 \\
0.672 & 0.707 & 0.724 & 0.733 & 2.561 & 2.628 & 3.510 & 4.685 \\
0.707 & 0.745 & 0.768 & 0.782 & 0.792 & 4.298 & 4.998 & 6.723 \\
0.723 & 0.768 & 0.794 & 0.812 & 0.800 & 0.879 & 7.526 & 9.367 \\
0.732 & 0.782 & 0.812 & 0.832 & 0.799 & 0.885 & 0.932 & 13.432
\end{pmatrix}
$$

*Elements in the lower triangular matrix exhibit corresponding correlation coefficients.

**Total sample size**

We manipulated five conditions for total sample size $J = 100, 200, 400, 800, 1600$.

**Mixture proportion**

Under condition (a), we set two conditions for mixture proportions, $p = (p_1, p_2) = (0.5, 0.5)$, and $(0.8, 0.2)$. Additionally, $p = (p_1, p_2, p_3) = (1/3, 1/3, 1/3)$ and $(0.45, 0.45, 0.10)$ are set for
condition (b).

**Data generation and estimation of parameters**

In data generation, firstly the sample size for each class was sampled using total sample size \( J \) and mixture proportions \( p \) via binomial and multinomial distributions for the conditions (a) and (b), respectively. Then a simulation dataset was generated through the mean and covariance structure \((\mu_c(\Theta), \Sigma_c(\Theta))\) calculated using specified parameters. These procedures were conducted by using R software (R Core Developing Team, 2011). Additionally, parameter estimation was performed via OpenMx (Boker et al., 2011) based on maximum likelihood estimation, which runs within the R program. All simulation programs are available upon request.

In total we set 2 (true number of classes and upper bound of classes to be estimated) \( \times \) 5 (degree of separation) \( \times \) 5 (total sample size) \( \times \) 2 (mixture proportion) = 100 conditions and generated 100 datasets for respective conditions, thus yielding 10,000 datasets. For these datasets, mixtures in the bivariate change score model with different numbers of classes are fit and parameters are estimated.

### 4.2. Results

For each information criterion, an estimation was considered correct if it showed a superior statistic in the model with \( C = 2 \) in the condition (a) and with \( C = 3 \) in the condition (b), respectively. Specifically, the smallest AIC, BIC, CAIC, aBIC, NEC, CLC, or ICL.BIC reflected the best fitting model, whereas largest values of Entropy reflected the best fitting model (Henson et al., 2007). To begin, we investigated the overall performance of information criteria in detecting classes under the various conditions, focusing on condition (a). Next, the results for condition (b) were examined to confirm consistency of the results under condition (a) and to investigate the performance of entropy-based information criteria more intensively.

#### Overall results

Figure 2 shows the average results of proportion of selected number of classes for each information criterion in 5,000 datasets. Well-known information criteria such as AIC, BIC, and CAIC show similar performances as observed in prior research. Namely, while AIC underestimated the number of classes (favoring the model with \( C = 1 \)) in 16% of cases, it most frequently overestimated the number of classes among likelihood-based information criteria and favored the model with \( C = 3 \) in 33% of cases. As expected from their definition, BIC and CAIC showed very similar performance, overestimating classes just once in 5,000 datasets, but underestimating classes in 54% of cases. These results are consistent with Bauer & Curran (2003), Henson et al (2007) and Vrieze (2012), who pointed out that AIC tends to overestimate the number of classes, and that BIC (and CAIC) tends to favor a smaller number of classes. aBIC showed features intermediate between AIC and BIC (CAIC), favoring the model with \( C = 1 \) and \( C = 3 \) by 26% and 21%, respectively, mainly due to the fact that aBIC takes a lower penalty as compared with BIC and CAIC, while it takes a larger penalty than AIC. These results show that likelihood-based information criteria did not give consistent estimation results when detecting classes, and the performance appears to be highly sensitive to controlled factors such as degree of separation and sample size, as noted in Henson et al (2007) and Vrieze(2012).

We next investigate the performance of entropy-based information criteria. Note that in the model with \( C = 1 \), no entropy-based information criterion other than NEC can be defined (as a convention, in this study NEC is defined as 1 in the model with \( C = 1 \)). From Figure 2, all entropy-based information criteria showed overextraction of classes in large proportions, and this was obvious in Entropy that favored the model with \( C = 3 \) 43% in cases. On the other hand, CLC and ICL.BIC were stable and rarely overestimated the number of classes (15% and 5% of cases, respectively).
Fig. 2: The average results of proportion of selected number of classes for 5,000 data sets for each information criteria.

These results regarding CLC and ICL.BIC were also indicated in Henson et al. (2007). On the other hand, NEC underestimated the class and favored the $C = 1$ model 47% in cases.

**The influence of degree of separation**

The proportions of models correctly favoring the model with $C = 2$ under different degrees of separation are illustrated in Figures 3. As shown in Figure 3, the degree of separation significantly influenced the performance of information criteria in detecting classes. Especially in the range from $Q = 2$ to $Q = 4$ (the effect size in the first occasion from 1.0 to 2.0), almost all information criteria showed obvious performance improvements, selecting the model with $C = 2$ in larger proportion. For example, when $Q = 4$ and $Q = 5$, all information criteria except for AIC and aBIC showed more than 80% or 90% preference to the model with $C = 2$, respectively. Thus, the information criteria may show poor performance in detecting classes, showing a biased estimation unless the degree of separation is large enough to detect classes. Even when $Q = 3$, which is often observed and anticipated in actual datasets, overall correct specification rates are not large enough in all information criteria. Although CLC and ICL.BIC favored the $C = 2$ model in almost all cases, even when the
degree of separation was small; this may be because the proportion of underestimation was removed since these information criteria cannot be numerically defined in the model with $C = 1$. This point will be investigated in the results for condition (b).

When investigating similar results in each level of degree of separation, AIC showed inconsistent preference and both underestimated and overestimated the classes when $Q = 1, 2$, while only overestimation was observed when $Q = 3, 4, 5$. aBIC showed a similar pattern, an inconsistent preference that both underestimated and overestimated the classes when $Q = 1, 2$, while only overestimation is observed when $Q = 3, 4, 5$. Interestingly, while BIC and CAIC favored the model with $C = 1$ in almost all datasets when $Q = 1, 2$, this rarely occurred when $Q = 3, 4, 5$. More specifically, both BIC and CAIC selected the model with $C = 1$ in just 15% and 1% of the cases when $Q = 4, 5$, respectively, and showed the largest proportion of preference for the model with $C = 2$ among all likelihood-based information criteria. A similar tendency was also observed in entropy-based information criteria. Entropy, CLC, and ICL.BIC overestimated the number of classes when the degree of separation was small, but these biases dramatically improved as the degree of separation became large (about $Q = 4$). These significant influences of degree of separation were consistent with prior research (e.g., Henson et al, 2007; Nylund et al, 2007; Steinley & Brusco, 2011; Vrieze, 2012).

**The influence of sample size**

The proportions correctly favoring the model with $C = 2$ under different sample sizes are illustrated in Figure 4. In Figure 4, almost all information criteria correctly favored the model with $C = 2$ as the sample size increased, although this influence is clearly smaller as compared with that of degree of separation. Exceptionally, NEC showed poor performance with large sample sizes, favoring the model with $C = 1$ as sample size increased. While not as obvious as with degree of separation, the influence of sample size can be observed in BIC, CAIC, and aBIC, which directly consider sample size by definition. In particular, as Henson et al. (2007) pointed out, when the sample size is small ($J = 100, 200, 400$) BIC and CAIC tend to underestimate classes and favor the model with $C = 1$, and are inferior to AIC and aBIC. From theoretical points of view this is natural since BIC and CAIC have a consistency property and are not guaranteed to select the true model when sample size is small (Vrieze, 2012).

![Fig. 3: Proportions that favored the model with $C=2$ in each degree of separation ($Q$).](image)
When investigating similar results in each level of sample size, AIC and aBIC favored the $C = 2$ model most frequently when $J = 100$ and $J = 200$, 400, 800, 1600, respectively, while AIC showed both overestimation and underestimation even with large sample sizes. This can also be explained by the non-consistency property of AIC. Additionally, while aBIC favored models with both $C = 1$ and $C = 2$ more frequently as sample size increased, only the proportion for the model with $C = 2$ was increased in AIC. These results again were consistent with prior research (e.g., Henson et al, 2007). On average, although aBIC was the most accurate in detecting classes among the likelihood-based statistics, this performance was limited to cases where degree of separation is large ($Q = 3, 4, 5$), or cases where the sample is sufficiently large ($J \geq 200$).

**The influence of mixture proportion**

Figure 5 illustrates the proportion correctly favoring the model with $C = 2$ under different con-

---

---
ditions of mixture proportions. As Figure 5 shows, the influence of mixture proportions is minimal as compared with degree of separation and sample size, again agreeing with Henson et al. (2007). Therefore, even when the mixture proportion is small as a part of whole classes, this does not directly degrade the performance of information criteria if the total sample size is more than $J = 100$. However, specified true mixture proportions in this simulation were not extreme and AIC and aBIC may perform much better than BIC and CAIC when class sizes are quite disparate (Vrieze, 2012), which is a likely error under this design.

**Results for condition (b) (where the true number of classes is 3)**

Here, we focus on the results for condition (b), where the true number of classes and the upper bound of classes estimated are 3 and 4, respectively. The main purpose of these conditions is to evaluate the consistency of the results under condition (a) and to investigate the performance of entropy-based information criteria that could not be closely investigated in Henson et al. (2007). Figure 6 illustrates the average results of proportion of selected number of classes for each inform-

![Figure 6: The average results of proportion of selected number of classes for 10,000 data sets for each information criteria in the condition (b) (True number of class is set to 3).](image-url)
mation criterion in condition (b) for 5,000 datasets. AIC underestimated the class and favored the model with $C = 1$ and $C = 2$ in 11% and 14% of cases, respectively, and overextracted classes (favoring the model with $C = 4$) in 27%, the largest value among likelihood-based information criteria. Additionally, while BIC and CAIC overestimated the number of classes in no cases, they underestimated the number of classes in 69%, a larger value favoring the model with $C = 3$. aBIC also showed consistent results for condition (a), overestimating and underestimating the number of classes in 18% and 36% of cases, respectively. When comparing the results Figure 2 and Figure 6, it is clear that proportions of correct specification of classes become much worse in BIC and CAIC, while AIC and aBIC showed slight declines. This reveals the drawback of BIC and CAIC in that they show poor performance when the true model is assumed to be complex (i.e., true number of classes is large).

Next, we focus on the performance of entropy-based information criteria. Most entropy-based information criteria rarely favored the model with $C = 4$. Specifically, this occurred in only 1% of cases under ICL.BIC, but in 37% of cases under Entropy. Additionally, NEC often underestimated the number of classes and favored the model with $C = 1$ in 40% of cases, indicating the inappropriateness of defining NEC as 1 when $C = 1$. Henson et al. (2007) also noted this point. CLC and ICL.BIC often underestimated the number of classes as BIC and CAIC did, and favored the model with $C = 2$, implying the superior results of CLC and ICL.BIC shown in Figure 2 (and Figure 2 in Henson et al., 2007, p215) were seriously biased.

Figure 7 shows the performance of each information criterion through different degrees of separation. Figure 7 again indicates that degree of separation strongly affects the mixture detection performance in all criteria. And the proportion of correct estimations is lower as compared with the results of condition (a) because the true model becomes complex. Additionally, although the difference is not meaningfully large, entropy-based information criteria performed better than likelihood-based information criteria in some conditions. Namely, on average entropy-based information criteria correctly estimated models more frequently than did likelihood-based information criteria when $Q = 4, 5$. Especially ICL.BIC shows better performance than any other criterion when

![Fig. 7: Proportions that favored the model with $C=3$ in each degree of separation in the condition (b).](image-url)
degree of separation is largest ($Q = 5$). We omit estimation results with different sample sizes and mixture proportions because nearly identical tendencies as under condition (a) were observed.

In summary, even when all model assumptions including normality of data are satisfied, all information criteria showed both overestimation and underestimation of classes under different conditions of degree of separation, sample size, and mixture proportions. Namely, different information criteria have quite different properties that require different assumptions. So, no single information criterion is always best (Henson, et al, 2007), and applied researchers and methodologists alike will benefit from improved understanding of the asymptotic and finite-sample behavior of various information criteria according to features of specified models and obtained data (Vrieze, 2012).

In the condition (a), the overall results were consistent with Henson et al (2007), and significant influences of degree of separations were clearly observed. Although different data generation models were used between the present study and Henson et al (2007), these models shared almost the same degree of model complexity (i.e., the number of parameters are not much different) and this fact likely contributed to yielding such consistent results. The results in the condition (b), which is one of the major contribution of the present study, can be summarized as: (1) Increased model complexity worsen the performance of all information criteria, and this is salient in BIC and CAIC. (2) The influence of degree of separation is dominant, and CLC and ICL.BIC also frequently underestimate the classes when it is not sufficiently large. Therefore the superior results of CLC and ICL.BIC shown in Figure 2 of the present study and Henson et al (2007, p.215) are seriously biased since the true model is too simple. (3) On average entropy-based information criteria correctly estimate models more frequently than likelihood-based information criteria, and this is salient when degree of separation is large. Especially, ICL.BIC shows better performance than any other criterion when degree of separation is large.

5. Simulation Study II (Model misspecification)

This section investigates the performance of information criteria when model misspecification is occurred. This section is divided into two subsections, and the different factors (i.e., outliers and an upper bound of measurement) that cause non-normality of data within classes are investigated in each subsection.

5.1. The influence of outliers on detecting classes

5.1.1. Simulation design

Since we have already investigated the average performance of each information criterion in the previous section, here we limit the simulation design and focus on condition (a), the degree of separation for $Q = 4$, sample sizes for $J = 200, 400, 800$, and mixture proportion for $(\rho_1, \rho_2) = (0.5, 0.5), (0.8, 0.2)$. Additionally, to simplify the investigation, the same simulation datasets are used to investigate the effects of both outliers and the upper bound of measurement value. As for parameters, we set $\mu_{i1}^x = 20, \mu_{i2}^x = 5, \mu_{i1}^y = 30, \mu_{i2}^y = 10, \mu_{i1}^z = 30, \mu_{i2}^z = 5, \mu_{i1}^s = 40, \mu_{i2}^s = 10, \gamma_1^x = \gamma_2^x = \gamma_1^y = \gamma_2^y = 0.10, \psi_1^{x2} = \psi_1^{y2} = \psi_1^{z2} = \psi_1^{s2} = 15, \beta_1^x = \beta_1^y = \beta_1^z = \beta_1^s = 0.20, \omega_1^{x2} = \omega_1^{y2} = \omega_1^{z2} = \omega_1^{s2} = 2$, and

$$
\Phi^1 = \Phi^2 = \begin{bmatrix} 10.00 & 0.63 & 4.00 & 0.32 \\ 0.63 & 1.00 & 0.32 & 0.20 \\ 4.00 & 0.32 & 10.00 & 0.63 \\ 0.32 & 0.20 & 0.63 & 1.00 \end{bmatrix}.
$$

— 32 —
Namely, degree of separation in the first occasion is controlled as \((40 - 30)/\sqrt{25} = (30 - 20)/\sqrt{25} = 2.0\). Under this parameter setting, the degree of separation and other features such as (non)linearity of latent curves and correlation matrices of the whole data were set to remain almost the same as in the former simulation study. Figure 8 shows the latent curves specified in this section.

Fig. 8: Specified developing trajectories in the simulation II.

Simulation data including outliers are generated via a multivariate \(t\) distribution with specified means and variance and covariance structures \((\mu_z^t(\Theta), \Sigma_z^t(\Theta))\). Specifically, we assume \(z_j\) follows a finite mixture of \(t\) distributions as

\[
z_j \sim \sum_{c=1}^{C} p_t(v, \mu_z^c(\Theta), \sqrt{\frac{v-2}{v}} \Sigma_z^c(\Theta)),
\]

where \(v\) denotes the degree of freedom. The term \(\sqrt{\frac{v-2}{v}}\) is a constant to generate data with an unbiased variance-covariance matrix. In this simulation, we set three conditions of \(v\) as \(v = 5, 10, 20\). Here, the lower the value \(v\) becomes, the more frequently outliers occur. Then, the data generation model with \(v = 20\) is relatively close to a multivariate normal distribution. Additionally, the normal distribution (where \(v = \infty\)) was also included as a data generation model for reference level.

Therefore, we set 3 (conditions of total sample size) \(\times 2\) (conditions of mixture proportion) \(\times 4\) (conditions of data generation model) = 24 conditions. 100 simulation datasets were generated for each condition, yielding 2,400 datasets. Parameter estimation and model selection were performed in the same manner as in the previous simulation. Namely, normal-mixtures in latent change score models with different numbers of classes \((C = 1, 2, 3)\) are arbitrarily fit to the generated dataset.
5.1.2. Results

Figure 9 shows the average results for the performance of information criteria over all 2,400 simulation datasets. It shows that each information wrongly favored the model with \( C = 3 \) more frequently than the previous simulation. Especially criteria such as AIC and Entropy wrongly favored the model with \( C = 3 \) by 64% and 61%, respectively. It was surprising that conservative statistics such as BIC and CAIC also wrongly favored the model with \( C = 3 \) more frequently than the previous simulation. This overextracting occurred to approximate complex non-normal distributions that included outliers by normal-mixtures (Bauer & Curran, 2003). However, on average BIC, CAIC, and ICL.BIC correctly chose the model with \( C = 2 \) in 77%, 77%, and 82% of the cases, respectively, showing robustness to outliers.

Figures 10 shows proportion of favoring the model with \( C = 2 \) under different data generation
Performance of Model Selection by Information Criteria

![Figure 10: Proportions that favored the model with $C=2$ in each data generation model.](image)

models. It clearly shows that the data generation model had a large effect on the performance of each criterion. The performance was almost satisfactory when normal assumptions were strictly met (where $v = \infty$). Surprisingly, when investigating the similar results in each level of data generation model, model with $v = 20$ led every information criteria except for BIC, CAIC, and ICL.BIC to overextract classes and to favor the model with $C = 3$. More specifically, in over 50% of cases these criteria favored the model with $C = 3$ when $v = 10$, and in around 90% when $v = 5$. Even BIC, CAIC, and ICL.BIC favored the model with $C = 3$ in around 50% of cases when $v = 5$, thus indicating the salient influence of outliers on detection of classes.

As for the effects of sample size and mixture proportion, a similar tendency was observed as in the previous simulation study and these effects were less significant compared with those of the data generation model. One exception is that larger sample size resulted in lower correct estimation rates because more outliers tended to be generated. Detailed results are omitted here.

We investigated the estimated latent curves from models with different numbers of classes, to evaluate the danger of specifying incorrect models (those with $C = 1$ and $C = 3$). For example, if the latent curves estimated from the $C = 3$ model are akin to those from $C = 2$, overextraction of class is less problematic since this does not necessarily lead to wrong conclusions about data.

Table 2 shows the averages of estimated latent curves (the averages of estimated values $\bar{X}_1, \ldots, \bar{X}_4, \bar{Y}_1, \ldots, \bar{Y}_4$ across the simulation dataset) and mixture proportions from models with different numbers of classes and data generation conditions. To simplify the illustration, the results in the condition of mixture proportions $(p_1, p_2) = (0.5, 0.5)$ are focused on here. Additionally, to discriminate the labels for respective classes, the class yielding latent curves with larger values was classified as the class with higher numbers in each simulation dataset. As Table 2 indicates, averages of estimates for latent curves from the model with $C = 2$ across all data generation models are very similar to those of mean structure $\mu'_j(\Theta)$ specified in Figure 8. When seeing each estimated parameter, averages of estimates show very similar values, indicating that convergence of parameters was successfully achieved. In the model with $C = 1$, latent curves that almost equal the averages of the true latent curves are indicated in the respective data generation models. In the model with $C = 3$, estimated latent curves exhibited mixed patterns of those estimated in models with $C = 1$ and $C = 2$. 

— 35 —
Table 2: Averages of estimated latent curves in the different data generation models \((J=400, (p_1, p_2)=(0.5,0.5))\)

(a) Normal distribution

<table>
<thead>
<tr>
<th>C=1</th>
<th>P</th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
<th>Y_1</th>
<th>Y_2</th>
<th>Y_3</th>
<th>Y_4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.00</td>
<td>24.95</td>
<td>38.48</td>
<td>56.64</td>
<td>80.88</td>
<td>34.99</td>
<td>54.46</td>
<td>79.17</td>
<td>110.68</td>
</tr>
</tbody>
</table>

| C=2 | c=1 | 0.50 | 19.96 | 31.99 | 48.21 | 69.92 | 30.04 | 48.03 | 70.81 | 99.78 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | c=2 | 0.50 | 29.94 | 44.97 | 65.06 | 91.85 | 39.93 | 60.88 | 87.53 | 121.57 |

| C=3 | c=1 | 0.39 | 19.18 | 31.03 | 47.00 | 68.36 | 34.99 | 54.46 | 79.17 | 110.68 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | c=2 | 0.21 | 24.89 | 38.27 | 56.24 | 80.24 | 34.81 | 54.17 | 78.73 | 110.00 |
|     | c=3 | 0.40 | 30.46 | 45.68 | 66.00 | 93.07 | 40.51 | 61.54 | 88.32 | 122.53 |

(b) t distribution \((v=20)\)

<table>
<thead>
<tr>
<th>C=1</th>
<th>P</th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
<th>Y_1</th>
<th>Y_2</th>
<th>Y_3</th>
<th>Y_4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.00</td>
<td>24.80</td>
<td>38.26</td>
<td>56.33</td>
<td>80.46</td>
<td>34.81</td>
<td>54.27</td>
<td>78.95</td>
<td>110.33</td>
</tr>
</tbody>
</table>

| C=2 | c=1 | 0.52 | 20.01 | 32.00 | 48.19 | 69.86 | 30.01 | 48.04 | 70.85 | 99.82 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | c=2 | 0.48 | 30.02 | 45.06 | 65.18 | 91.97 | 40.01 | 61.03 | 87.74 | 121.75 |

| C=3 | c=1 | 0.43 | 19.88 | 31.23 | 47.06 | 68.30 | 29.49 | 47.34 | 69.94 | 98.67 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | c=2 | 0.15 | 25.30 | 38.81 | 56.89 | 80.93 | 34.39 | 53.91 | 78.71 | 110.32 |
|     | c=3 | 0.42 | 30.38 | 45.56 | 66.00 | 93.07 | 40.51 | 61.54 | 88.32 | 122.53 |

(c) t distribution \((v=10)\)

<table>
<thead>
<tr>
<th>C=1</th>
<th>P</th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
<th>Y_1</th>
<th>Y_2</th>
<th>Y_3</th>
<th>Y_4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.00</td>
<td>24.86</td>
<td>38.29</td>
<td>56.34</td>
<td>80.52</td>
<td>34.84</td>
<td>54.25</td>
<td>78.91</td>
<td>110.34</td>
</tr>
</tbody>
</table>

| C=2 | c=1 | 0.52 | 19.93 | 31.92 | 48.09 | 69.77 | 29.95 | 47.89 | 70.62 | 99.56 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | c=2 | 0.48 | 30.11 | 45.09 | 65.19 | 91.98 | 40.05 | 61.08 | 87.74 | 121.82 |

| C=3 | c=1 | 0.48 | 19.35 | 31.39 | 47.41 | 68.83 | 29.86 | 47.54 | 69.95 | 98.48 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | c=2 | 0.12 | 24.80 | 38.08 | 56.06 | 80.13 | 34.37 | 54.48 | 80.00 | 112.43 |
|     | c=3 | 0.41 | 30.50 | 45.57 | 65.73 | 92.68 | 40.94 | 62.46 | 89.51 | 123.77 |

(d) t distribution \((v=5)\)

<table>
<thead>
<tr>
<th>C=1</th>
<th>P</th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
<th>Y_1</th>
<th>Y_2</th>
<th>Y_3</th>
<th>Y_4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.00</td>
<td>24.98</td>
<td>38.42</td>
<td>56.54</td>
<td>80.77</td>
<td>34.92</td>
<td>54.46</td>
<td>79.21</td>
<td>110.70</td>
</tr>
</tbody>
</table>

| C=2 | c=1 | 0.50 | 19.97 | 31.88 | 48.05 | 69.73 | 29.90 | 47.94 | 70.72 | 99.66 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | c=2 | 0.50 | 30.06 | 45.08 | 65.19 | 91.98 | 40.02 | 61.08 | 87.82 | 121.90 |

| C=3 | c=1 | 0.49 | 19.86 | 31.81 | 48.03 | 69.70 | 29.94 | 47.85 | 70.48 | 99.24 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | c=2 | 0.06 | 25.89 | 39.91 | 58.17 | 81.97 | 35.83 | 55.69 | 80.72 | 112.04 |
|     | c=3 | 0.46 | 30.47 | 45.57 | 65.69 | 92.39 | 41.04 | 62.46 | 89.51 | 123.77 |

Namely, the first and the third classes in the model with \(C = 3\) exhibited almost the same latent curves as the true latent values from models with \(C = 2\), and the second curves showed the almost average of the true latent curves, as the model with \(C = 1\).

Next, we focus on mixture proportions. Table 2 clearly indicates that as outliers become influential (as \(v\) becomes small), the mixture proportion in the second class in the model with \(C = 3\) became small \((p_2 = 0.21, 0.15, 0.12, 0.06\) for the normal distribution and the three \(t\) distributions, respectively). In mixture modeling for applied research, we may face trivial and/or artifact class
extractions (e.g., Bauer & Curran, 2003). One example is the case where almost identical parameter estimates are obtained, resulting in similar latent curve estimations in at least two classes. Another example is the case where the mixture proportion of some classes is too small because it explains only a small portion of data such as outliers. In case of this simulation, if researchers can judge overextracted second classes as trivial, they can lead conclusions about estimated latent trajectories appropriately since the first and the third classes in the model with $C = 3$ exhibited almost the same latent curves as the true latent values.

Figure 11 shows the proportion of cases where the mixture proportion of the second class ($c = 2$) was less than three cutoff values (0.01, 0.03, and 0.05) in the model with $C = 3$ under data generation models based on normal distribution with $v = 10$. Just 10% of cases exhibited mixture proportions less than 0.01 in $J = 200$, but when $J = 800$, mixture proportions less than 0.01 were observed in 47% of cases. This result is natural, because outliers arise more frequently when the total sample size increases. Additionally, the cutoff point is strongly related to the results, indicating that larger values of cutoff point relate to larger proportions in Figure 11. Similar results are observed in other data generation models. These results imply that in actual data analysis, increasing the sample size is undoubtedly effective, because doing so not only gives larger statistical power and more precise parameter estimates with lower standard error, but also identifies classes that are artificially extracted due to model misspecification.

![Fig. 11: Proportion of cases where mixture proportion of second class ($c=2$) was less than three cut-off values in the model with $C=3$ (t distribution with $v=10$).]

5.2. The influence of measurement upper bounds on detecting classes

5.2.1. Simulation design

As noted earlier, this simulation used the same dataset generated in the previous subsection for the normal case (i.e., $v = \infty$). Then, three kinds of upper bounds of measurement $U=130, 120$ and 110, and a condition with no upper bound (i.e., $U = \infty$) are set, yielding a total of 4 conditions for upper bounds. When generated data $z$ is above $U$, this data is transformed to $U$. Namely, new simulation data $z^*$ is generated by $z^* = \min(z, U)$. Thus a total of $3(\text{total sample size}) \times 2(\text{mixture proportion}) \times 4(\text{upper bound of measurements}) \times 100 = 2,400$ simulation dataset is generated here, and normal-mixtures in latent change score models with different number of classes ($C = 1, 2, 3$) are arbitrarily fit to this dataset. Note that the simulation data for conditions with no upper bound were identical to those of the normal case ($v = \infty$) used in the previous subsection.
As means and standard deviations of trajectories in Figure 8 indicate, just part of the variables (mainly \( Y_4 \)) generated from the second class are affected by these upper bounds. Namely, when the upper bound is \( U = 130 \), almost 20% of the \( Y_4 \) data generated from the second class are expected to be transformed to 130, because the expected value and the variance (standard deviation) of \( Y_4 \) can be calculated as \( E(Y_4) = 121.75 \) and \( \sqrt{V(Y_4)} = 9.326 \), respectively. Likewise, when the upper bound is \( U = 120 \) and \( U = 110 \), almost 50% and 80% of \( Y_4 \) generated from the second class are expected to be affected and transformed to 120 and 110, respectively. Additionally, when the upper bound is \( U = 110 \), about 10% of \( Y_4 \) generated from the first class are also expected to be affected.

5.2.2. Results

Figure 12 shows the average results of the performance of information criteria in all 2,400 simulation datasets. As in the previous simulation, AIC clearly shows the largest proportion (74%)
wrongly favoring the model with $C = 3$. Surprisingly, BIC and CAIC, which showed relatively robust results to outliers in the previous simulation, overextracted classes more frequently. Overextracting occurred to approximate complex non-normal distributions that includes right-censored data from normal mixtures. Additionally, while BIC, CAIC, and ICL.BIC were relatively robust to outliers, entropy-based information criteria (Entropy, NEC, CLC, and ICL.BIC) were relatively robust to upper bounds of measurement and chose the $C = 2$ model more frequently.

Figure 13 shows the results of the performance of information criteria with different upper bounds of measurement. Figure 13 clearly indicates that while only a part of information criteria such as AIC and aBIC tend to overextract the classes when $U = 130$, almost all information criteria were sensitive to upper bounds of measurement and favored the model with $C = 3$ when $U = 120$. More specifically, on average likelihood-based and entropy-based criteria favored the model with $C = 3$ in 80–90% and 40% of cases, respectively. These results indicate that even where a part of the variables ($Y_4$) are sensitive to the upper bound of measurement, there is a strong tendency for overextraction of classes when using information criteria.

![Fig. 13: Proportions that favored the model with $C=2$ in different conditions of the upper bound.](image)

Differences in sample size were trivial in the result of model selection. On the other hand, mixture proportion is related to the performance of entropy-based information criteria since the sample size for the $c = 2$ class, which is affected by the upper bound of measurement, greatly differs according to mixture proportions $(p_1, p_2)$. These results are omitted here.

Table 3 shows the averages of estimated latent curves and mixture proportions from models with different numbers of classes and conditions of the upper bounds of measurement, focusing on the conditions that the mixture proportion is $(p_1, p_2) = (0.5, 0.5)$. Note that Table 3(a) indicates identical values to Table 2(a) because the same datasets were used in both simulations. As expected, averages of estimates for $Y_4$ in the $c = 2$ class are lower than the population values indicated in Figure 8 when an upper bound is specified and this characteristic is even more pronounced as an upper bound becomes smaller. Aside from this point, the model with $C = 2$ could reflect the true latent trajectories correctly. In the model with $C = 3$, the estimated latent curve for the third class

— 39 —

Table 3: Averages of estimated latent curves in the different conditions of the upper bound (UB) of measurement ($J=400$, $(p_1, p_2)=(0.5,0.5)$)

(a) Normal distribution (no UB)

<table>
<thead>
<tr>
<th></th>
<th>$P$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
<th>$Y_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C=1$</td>
<td></td>
<td>1.00</td>
<td>24.95</td>
<td>38.48</td>
<td>56.64</td>
<td>80.88</td>
<td>34.99</td>
<td>54.46</td>
<td>79.17</td>
</tr>
<tr>
<td>$C=2$</td>
<td>$c=1$</td>
<td>0.50</td>
<td>19.96</td>
<td>31.99</td>
<td>48.21</td>
<td>69.92</td>
<td>30.04</td>
<td>48.03</td>
<td>70.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>29.94</td>
<td>44.97</td>
<td>65.06</td>
<td>91.85</td>
<td>39.93</td>
<td>60.88</td>
<td>87.53</td>
</tr>
<tr>
<td>$C=3$</td>
<td>$c=1$</td>
<td>0.39</td>
<td>19.18</td>
<td>31.09</td>
<td>47.00</td>
<td>68.36</td>
<td>29.49</td>
<td>47.34</td>
<td>69.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.21</td>
<td>24.89</td>
<td>38.27</td>
<td>56.24</td>
<td>80.24</td>
<td>34.81</td>
<td>54.17</td>
<td>78.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.40</td>
<td>30.46</td>
<td>45.68</td>
<td>66.00</td>
<td>93.07</td>
<td>40.51</td>
<td>61.54</td>
<td>88.32</td>
</tr>
</tbody>
</table>

(b) UB=130

<table>
<thead>
<tr>
<th></th>
<th>$P$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
<th>$Y_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C=1$</td>
<td></td>
<td>1.00</td>
<td>24.71</td>
<td>38.10</td>
<td>56.09</td>
<td>80.11</td>
<td>34.67</td>
<td>54.12</td>
<td>78.67</td>
</tr>
<tr>
<td>$C=2$</td>
<td>$c=1$</td>
<td>0.53</td>
<td>20.09</td>
<td>32.10</td>
<td>48.28</td>
<td>69.93</td>
<td>30.00</td>
<td>48.03</td>
<td>70.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.47</td>
<td>29.94</td>
<td>44.91</td>
<td>64.93</td>
<td>91.63</td>
<td>39.95</td>
<td>60.99</td>
<td>87.53</td>
</tr>
<tr>
<td>$C=3$</td>
<td>$c=1$</td>
<td>0.36</td>
<td>19.43</td>
<td>31.20</td>
<td>47.02</td>
<td>68.20</td>
<td>29.28</td>
<td>46.54</td>
<td>68.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.40</td>
<td>23.15</td>
<td>36.18</td>
<td>53.67</td>
<td>77.04</td>
<td>33.05</td>
<td>52.71</td>
<td>87.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.24</td>
<td>30.50</td>
<td>45.62</td>
<td>65.85</td>
<td>92.83</td>
<td>40.51</td>
<td>61.83</td>
<td>88.48</td>
</tr>
</tbody>
</table>

(c) UB=120

<table>
<thead>
<tr>
<th></th>
<th>$P$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
<th>$Y_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C=1$</td>
<td></td>
<td>1.00</td>
<td>24.57</td>
<td>37.93</td>
<td>55.92</td>
<td>79.92</td>
<td>34.48</td>
<td>54.09</td>
<td>78.40</td>
</tr>
<tr>
<td>$C=2$</td>
<td>$c=1$</td>
<td>0.53</td>
<td>19.77</td>
<td>31.58</td>
<td>47.54</td>
<td>68.92</td>
<td>29.34</td>
<td>47.30</td>
<td>69.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.47</td>
<td>28.94</td>
<td>43.72</td>
<td>63.52</td>
<td>89.84</td>
<td>39.17</td>
<td>60.13</td>
<td>85.87</td>
</tr>
<tr>
<td>$C=3$</td>
<td>$c=1$</td>
<td>0.23</td>
<td>18.80</td>
<td>30.20</td>
<td>45.56</td>
<td>66.07</td>
<td>28.02</td>
<td>44.83</td>
<td>65.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.46</td>
<td>21.49</td>
<td>33.89</td>
<td>50.69</td>
<td>73.24</td>
<td>31.28</td>
<td>50.40</td>
<td>74.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.31</td>
<td>29.67</td>
<td>44.69</td>
<td>64.83</td>
<td>91.56</td>
<td>39.99</td>
<td>61.15</td>
<td>87.06</td>
</tr>
</tbody>
</table>

(d) UB=110

<table>
<thead>
<tr>
<th></th>
<th>$P$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
<th>$Y_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C=1$</td>
<td></td>
<td>1.00</td>
<td>24.56</td>
<td>37.68</td>
<td>55.77</td>
<td>79.74</td>
<td>34.32</td>
<td>54.15</td>
<td>78.12</td>
</tr>
<tr>
<td>$C=2$</td>
<td>$c=1$</td>
<td>0.45</td>
<td>20.06</td>
<td>31.32</td>
<td>47.16</td>
<td>68.64</td>
<td>28.93</td>
<td>46.94</td>
<td>69.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.55</td>
<td>28.57</td>
<td>43.32</td>
<td>63.39</td>
<td>89.47</td>
<td>39.03</td>
<td>60.32</td>
<td>84.96</td>
</tr>
<tr>
<td>$C=3$</td>
<td>$c=1$</td>
<td>0.22</td>
<td>19.06</td>
<td>29.94</td>
<td>45.18</td>
<td>65.72</td>
<td>27.44</td>
<td>44.69</td>
<td>66.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.46</td>
<td>21.59</td>
<td>33.33</td>
<td>50.02</td>
<td>72.84</td>
<td>30.90</td>
<td>50.23</td>
<td>74.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.32</td>
<td>29.48</td>
<td>44.65</td>
<td>65.19</td>
<td>91.70</td>
<td>40.20</td>
<td>61.68</td>
<td>86.15</td>
</tr>
</tbody>
</table>

(c = 3) is almost equal to the second class of true latent curves, and the averages of latent curves for the first (c = 1) and second (c = 2) class in the model with C = 3 roughly correspond to the first class of true latent curves. Additionally, as lower upper bounds are specified, on average the mixture proportion for $p_2$ in the model with $C = 3$ becomes large because the variance of $Y_4$ becomes lower in the third class, causing the larger variances of $Y_4$ in the second class to explain data that distribute near the upper bound in $Y_4$. These results indicate that latent curves are extracted that less precisely reflect true latent trajectories, compared with the case of outliers (see in Table — 40 —)
Therefore, favoring the overextracted model with $C = 3$ itself can lead grave mistakes and does not enable researchers to evaluate true latent trajectories appropriately. Additionally, classes with small mixture proportions are not observed even when the sample size is large as in the previous simulation, indicating the greater difficulty of precise estimation of classes under the influence of the upper bounds of measurement. Using entropy-based criteria is therefore strongly desirable in this case.

6. An actual example using JAHEAD data

A simple example is illustrated for estimation and specification of the number of classes for mixtures in a two-variate latent change score model using JAHEAD (Japanese Study of Assets and Health Dynamics Among the Oldest Old) Data (Akiyama, Sugawara, Takeuchi & Kobayashi, 2008; Usami & Sugawara, 2012). JAHEAD is a large longitudinal survey conducted by Tokyo Metropolitan Institute of Gerontology and University of Michigan since 1987 to understand how physical ability, mental state and social relations of old people change with increasing age. The total of 8 surveys have been conducted almost every three years and almost 6000 old people have participated in surveys at least once so far.

Here, just for illustration, we use height and weight data that were collected through self report and/or records of health card from each participant from wave1-wave4 data, and ages of the old are classified into the four categories as 66–71, 72–77, 78–83, 84–89 years old (i.e., $T$ was set as $T = 4$ for consistency of this illustration with simulation study). Additionally, when two kinds of data were available for the same individual corresponding to the same age category, averages were used in this illustration. Furthermore, we restricted the data used here to the old whose weight or height data were available in at least one age category, and so that data with total $J = 4641$ were used for this illustration. Table 4 shows several descriptive statistics of height and weight data. Note that it was confirmed that the number of outliers was very small, and that no upper or lower measurement bound was specified for these data.

Mixtures in two-variate latent change score models were fit to the data. Here, due to space limitation, only estimated latent curves (and their standard deviation) for weight and height are shown in Figure 14. Although estimates of respective parameters are omitted here, it should be noted that coupling parameters (i.e., $\gamma_x$, $\gamma_y$) were not statistically significant in all classes of all fitted models. Additionally, the upper bound of the number of classes estimated is set to three since results based on the models with more than three classes showed improper solutions entailing haywood case. The values of each information criterion from the model with $C = 1$ to $C = 3$ are shown in Table 5 by sex. It is indicated that in males, many information criteria including BIC and ICL.BIC favored the model with $C = 2$, while AIC aBIC and Entropy favored the model with $C = 3$. In females, all infomration criteria except for CLC and ICL.BIC favored the model with $C = 3$, thus indicating that it seems appropriate to specify the number of classes as two or three in this data. Additionally, this judgement seems very natural since various patterns of changes of weight and height were observed in raw data.

On this point, one of the main features in these estimated latent curves is their low degree of separation. For instance, in the model with $C = 2$, effect sizes (the value equals to mean difference divided by the square root of the weighted variance) in each age category were in the range of $0.26 \sim 0.59$ and $0.05 \sim 0.32$ for height and weight in males data, respectively. As for the model with $C = 3$, the largest effect sizes were in the ranges of $0.63 \sim 0.75$ and $0.47 \sim 0.60$ between the first ($c = 1$) and the third ($c = 3$) class for height and weight, respectively. Effect sizes for data of females showed a relative larger degree of separation than those of men in both models with $C = 2$ and $C = 3$. For example, effect sizes were the largest in the range of $0.78 \sim 1.57$ for weight data between the first
### Table 4(a): Descriptive statistics for height and weight data

<table>
<thead>
<tr>
<th>Age 66-71</th>
<th>Age 72-77</th>
<th>Age 78-83</th>
<th>Age 84-89</th>
<th>Age 66-71</th>
<th>Age 72-77</th>
<th>Age 78-83</th>
<th>Age 84-89</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>155.17</td>
<td>153.92</td>
<td>153.63</td>
<td>153.37</td>
<td>53.88</td>
<td>51.47</td>
<td>49.80</td>
</tr>
<tr>
<td>minimum</td>
<td>130</td>
<td>120</td>
<td>130</td>
<td>121</td>
<td>30</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>maximum</td>
<td>182</td>
<td>178</td>
<td>175</td>
<td>175</td>
<td>105</td>
<td>93</td>
<td>78</td>
</tr>
<tr>
<td>skewness</td>
<td>0.12</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.23</td>
<td>0.42</td>
<td>0.45</td>
<td>0.30</td>
</tr>
<tr>
<td>kurtosis</td>
<td>-0.28</td>
<td>-0.02</td>
<td>-0.44</td>
<td>0.29</td>
<td>0.49</td>
<td>0.20</td>
<td>-0.18</td>
</tr>
<tr>
<td>sample size</td>
<td>1564</td>
<td>1162</td>
<td>638</td>
<td>175</td>
<td>1596</td>
<td>1206</td>
<td>680</td>
</tr>
</tbody>
</table>

### Table 4(b): Descriptive statistics for height and weight data (Male)

<table>
<thead>
<tr>
<th>Age 66-71</th>
<th>Age 72-77</th>
<th>Age 78-83</th>
<th>Age 84-89</th>
<th>Age 66-71</th>
<th>Age 72-77</th>
<th>Age 78-83</th>
<th>Age 84-89</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>161.54</td>
<td>160.69</td>
<td>160.16</td>
<td>159.44</td>
<td>57.60</td>
<td>55.47</td>
<td>54.28</td>
</tr>
<tr>
<td>minimum</td>
<td>140</td>
<td>130</td>
<td>133</td>
<td>140</td>
<td>37</td>
<td>38</td>
<td>30</td>
</tr>
<tr>
<td>maximum</td>
<td>182</td>
<td>178</td>
<td>175</td>
<td>175</td>
<td>105</td>
<td>93</td>
<td>78</td>
</tr>
<tr>
<td>skewness</td>
<td>0.08</td>
<td>-0.14</td>
<td>-0.33</td>
<td>-0.25</td>
<td>0.57</td>
<td>0.47</td>
<td>0.21</td>
</tr>
<tr>
<td>kurtosis</td>
<td>0.67</td>
<td>0.95</td>
<td>0.49</td>
<td>0.58</td>
<td>1.00</td>
<td>0.22</td>
<td>-0.11</td>
</tr>
<tr>
<td>sample size</td>
<td>713</td>
<td>516</td>
<td>306</td>
<td>89</td>
<td>722</td>
<td>523</td>
<td>310</td>
</tr>
</tbody>
</table>

### Table 4(c): Descriptive statistics for height and weight data (Female)

<table>
<thead>
<tr>
<th>Age 66-71</th>
<th>Age 72-77</th>
<th>Age 78-83</th>
<th>Age 84-89</th>
<th>Age 66-71</th>
<th>Age 72-77</th>
<th>Age 78-83</th>
<th>Age 84-89</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>149.83</td>
<td>148.52</td>
<td>147.61</td>
<td>147.08</td>
<td>50.81</td>
<td>48.42</td>
<td>46.05</td>
</tr>
<tr>
<td>minimum</td>
<td>130</td>
<td>120</td>
<td>130</td>
<td>121</td>
<td>30</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>maximum</td>
<td>182</td>
<td>178</td>
<td>175</td>
<td>175</td>
<td>105</td>
<td>93</td>
<td>78</td>
</tr>
<tr>
<td>skewness</td>
<td>-0.09</td>
<td>-0.42</td>
<td>-0.29</td>
<td>-0.95</td>
<td>0.32</td>
<td>0.34</td>
<td>0.29</td>
</tr>
<tr>
<td>kurtosis</td>
<td>0.64</td>
<td>1.63</td>
<td>0.62</td>
<td>2.13</td>
<td>-0.13</td>
<td>-0.18</td>
<td>-0.31</td>
</tr>
<tr>
<td>sample size</td>
<td>851</td>
<td>646</td>
<td>332</td>
<td>86</td>
<td>874</td>
<td>683</td>
<td>370</td>
</tr>
</tbody>
</table>

### Table 5(a): The values of each information criterion from the model with C=1 to C=3 (Male)

<table>
<thead>
<tr>
<th>C = 1</th>
<th>AIC</th>
<th>BIC</th>
<th>CAIC</th>
<th>aBIC</th>
<th>Entropy</th>
<th>NEC</th>
<th>CLC</th>
<th>ICLBIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C = 2</td>
<td>19843.43</td>
<td>20057.13</td>
<td>19920.56</td>
<td>0.272</td>
<td>0.836</td>
<td>20831.42</td>
<td>21131.12</td>
<td></td>
</tr>
<tr>
<td>C = 3</td>
<td>19783.03</td>
<td>20106.07</td>
<td>19899.62</td>
<td>0.390</td>
<td>0.838</td>
<td>21078.77</td>
<td>21531.80</td>
<td></td>
</tr>
</tbody>
</table>

### Table 5(b): The values of each information criterion from the model with C=1 to C=3 (Female)

<table>
<thead>
<tr>
<th>C = 1</th>
<th>AIC</th>
<th>BIC</th>
<th>CAIC</th>
<th>aBIC</th>
<th>Entropy</th>
<th>NEC</th>
<th>CLC</th>
<th>ICLBIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C = 2</td>
<td>23951.12</td>
<td>24170.81</td>
<td>24170.84</td>
<td>24034.22</td>
<td>0.308</td>
<td>1.075</td>
<td>25037.68</td>
<td>25354.37</td>
</tr>
<tr>
<td>C = 3</td>
<td>23810.38</td>
<td>24142.47</td>
<td>24142.52</td>
<td>23936.00</td>
<td>0.389</td>
<td>0.897</td>
<td>25321.94</td>
<td>25784.03</td>
</tr>
</tbody>
</table>

*The class indicated in the bold type is favored by each information criterion.*
and the second class in the model with $C = 3$, although these values are still not large enough.

When recalling the fact that the information criteria such as BIC, CAIC, ICL.BIC tend to underestimate the number of classes when degree of separation is small, these low degrees of separation in the data seemed to lead these information criteria to favor the model with $C = 2$, especially in data for males. Additionally, the reasons that CLC and ICL.BIC favored the models with $C = 2$ seems to be attributed to the fact that small degree of separation caused relatively unclear classification results and so the value $EN$ which relates to entropy became large. Furthermore, from the theoretical points of view in gerontology, the model with $C = 3$ is more reasonable in interpreting changes for weight and height of the old. Akiyama et al (2008) utilized mixture model and showed that there are old
people whose activities of daily living keep high in all age categories. Similar trajectories of weight and height, which were not extracted in the model with $C = 2$ but with $C = 3$ for height and weight data should be observed among such old people, since rapid changes of weight and height sometimes indicate one or more of several malfunctions such as osteoporosis and diabetes in the old.

From these results, although inconsistent estimation results were observed by information criteria in both data of males and females, specification of the model with $C = 3$ seems to be more appropriate in considering the features of information criteria as is shown in the former simulation studies and the clear-cut interpretation to these data.
7. Discussion

In the present research, novel simulation studies were performed to investigate the performances of likelihood-based and entropy-based information criteria for estimating the number of latent classes of LGCMMs, considering the effect of true model complexity and model misspecification. In the simulation, two-variate latent change score models, which recently have attracted considerable attention, are used for data generation. A major contribution of the first simulation is to compare the performance of likelihood-based and entropy-based information criteria to detect classes under different model complexities (the number of true classes). The new results can be summarized as:

1. Increased model complexity worsens the performance of both likelihood-based and entropy-based information criteria, and this was salient in BIC and CAIC.

2. The influence of degree of separation is dominant, and CLC and ICL.BIC also frequently underestimate the number of classes when it is not sufficiently large. Therefore, as Henson et al. (2007) themselves implied in their study, the superior results of CLC and ICL.BIC shown in Figure 2 of the present study and Henson et al. (2007, p.215) are seriously biased since true model is too simple.

3. On average entropy-based information criteria correctly estimated models more frequently than did likelihood-based information criteria, and this was salient when degree of separation was large.

In the condition (a) of the first simulation, the overall results were consistent with Henson et al. (2007). Therefore it was again indicated that different information criteria have quite different properties that require different assumptions and that no single information criterion is always best even when statistical assumptions are completely satisfied. Additionally, it was also again indicated that these obtained results do not heavily depend on data generation models if they share the same degree of model complexity.

In the second simulation the two factors were considered that cause model misspecification. Namely, normal-mixture models were inappropriately fit to non-normal data that included outliers and upper bounds of measurement. The new results can be summarized as:

4. When a normal-mixture is inappropriately fit to non-normal data that include outliers within classes, although this seriously worsens the performance of many information criteria, BIC, CAIC, and ICL.BIC are relatively robust. Additionally, overextracted classes with trivially small mixture proportions can be detected when the sample size is large.

5. When a normal-mixture is inappropriately fit to non-normal data that include upper-bounded measurement values (i.e., right-censored data) within classes, although this worsens the performance of almost all information criteria, entropy-based information criteria are relatively robust.

6. Although no single information criterion is always best, ICL.BIC is relatively robust to model misspecification and shows better performance than any others on average, especially when degree of separation is sufficiently large.

The present simulations are useful in evaluating the performance of information criteria, but there remains a need for additional simulation studies under other conditions such as the different types of model misspecification, different model complexity, different numbers of occasions and
variables, and different estimation methods. As for model misspecification in mixtures in the latent change score model, it should be noted that assumptions such as homoscedasticity of error and residual variances ($\psi^2$ and $\omega^2$) and identical restrictions of autoregression parameters ($\beta$) and coupling parameters ($\gamma$) over occasion $t$ may be strong and are not fulfilled in actual data. Such an influence of misspecification regarding restrictions of parameters should be investigated in future research.

In the present study, to investigate the performance of information criteria under various conditions, mean and variance-covariance matrices calculated from the specified model parameters (i.e., $\mu^2(\Theta)$ and $\Sigma^2(\Theta)$) were used in the simulations. On the other hand, there are several studies where population model parameters were specified by referring to the actual analysis results (e.g., Henson et al., 2007), thus parameter settings conducted in the present research may seem to be ad hoc. However, at least in the present research, various kinds of variance-covariance matrices could be specified that reflect plausible or justifiable conditions often encountered in actual research, and what is more the similarity of the results with Henson (2007) et al seems to be an additional support to show the strength of the evidence obtained through the present simulations. Importantly, whatever approach is used, the robustness of obtained results should be investigated through the accumulation of both additional simulation research and empirical knowledge, because in actual longitudinal data analysis observed mean vectors and variance-covariance matrices significantly differ for respective data.

As described in this research, real data do not necessarily include developmental trajectories explained from normal mixtures with a finite (and often smaller) number of classes. Actual behavior and corresponding developmental trajectories may be much more complex than expected, in which case latent classes may be overextracted as a result of a combination of factors regarding misspecification of statistical models. We should therefore recognize that evaluating different developing trajectories through LGCMM is a challenging part of the research process, since our tool appears to be fragile in real world scenarios. However, as Vrieze(2012) noted, applied researchers and methodologists alike will benefit from improved understanding of the asymptotic and finite-sample behavior of various information criteria according to features of specified models and obtained data, and therefore additional future studies are strongly desired.

REFERENCES


(Received: September 19, 2013, Accepted: August 5, 2014)