Simulation Study on Selection Latent Class Models with Missing Data

Jun Zhang* Mark Reiser†

Abstract

Longitudinal biomedical studies often encounter substantial missing data. An increasing number of articles introducing methods for handling missing data have discussed and used latent class models as a flexible way of modeling correlated multivariate categorical data. However, one key assumption of latent class modeling, the validity of the number of latent classes for missing data, has not been examined. The aim of this paper is to investigate the "correct" number of latent classes through simulation studies with missing values. We apply Monte Carlo simulation to generate a longitudinal study with 6 time points and two different missing mechanisms: missing completely at random and missing not at random. A linear mixed model with random intercept and slope is assumed for each latent class. We choose the most efficient approach to evaluate model performances with different latent classes: information criteria. Furthermore, we have investigated how the following factors influence the selection of latent classes for missing data: covariates effects, missing probabilities and the degree of associations among repeated measures. Due to the difficulties to identify the missing mechanism(s) in practice, missing patterns are also investigated in fitting latent class models.

Key Words: Missing mechanism, missing patterns, latent class model, information criteria

1. Introduction

Latent class modeling now is wildly used and frequently appearing in medical and statistical journals. A potential application of latent class models (LCM) is for exploring missing data (dropouts or intermittent missing) in longitudinal studies. In the intermittent missing cases, missing-data patterns could have many forms and the effects from missing patterns might be difficult to assess. For instance, in a series of depression studies described in J Roy's papers (2007), patients were randomly assigned to receive either drug plus psychotherapy or psychotherapy alone. Data were collected weekly during that period of 17 weeks including baseline. As mentioned in the paper, data at baseline were completely collected, but there was a large quantity of missing data afterwards. There were 379 unique missing-data patterns that were observed.

Latent class models with 3 latent classes were used by J Roy to assess whether subjects from different missing-data patterns had different responses on the changes in depression over time. However, one of difficulties, also a key condition for using latent class models, is deciding the number of latent classes. Garrett (2000) suggested using graphical methods for selecting the number of classes. Some researchers also proposed a Bayesian approach to select the number of latent classes by specifying a prior for the number of classes. One could select the model with the highest posterior probability for that number of classes. In this paper, we perform a systematic simulation study and investigate selection of the appropriate number of latent classes via different information criteria. In Section 2, we make a short review of latent class models. In Section 3, we first give a brief description of simulation studies, then elaborate methods and models that are used in longitudinal simulation studies. We present and analyze simulation outputs in Section 4, with conclusion and discussion in Section 5.

*First author's affiliation, Physical Sciences, A-Wing, P.O. Box 871804, Tempe, AZ 85287-1804

†Second author's affiliation, Physical Sciences, A-Wing, P.O. Box 871804, Tempe, AZ 85287-1804

2. Review of Latent Class Models

Lazarsfeld (1950a) first proposed latent class models in 1950 when they used this technique as a tool for building typologies based on observed dichotomous variables. The basic idea underlying LCM is some parameters of a postulated statistical model differ across unobserved subgroups. These subgroups form the categories of a categorical latent variable.

Let π_{ij} be the probability of a positive response on variable i for a person in category j ($i = 1, 2, \ldots, p;\ j = 0, 1, \ldots, c_i - 1$) and let η_j be the prior probability that a randomly chosen individual is in class j which satisfies $\sum_{j=1}^{K} \eta_j = 1$. For the case of K latent classes, the distribution of an individual responses becomes

$$f(\mathbf{x}) = \sum_{j=0}^{K} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i} \tag{1}$$

where $\mathbf{x}$ is the response vector of an individual. The posterior probability that an individual with response vector $\mathbf{x}$ belongs to category j is thus

$$h(j|\mathbf{x}) = \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i} / f(\mathbf{x}) \ \ (j = 1, 2, \ldots, K) \tag{2}$$

We can use (2) to construct an allocation rule according to which an individual is placed in the class for which the posterior probability is greatest. The principle statistical task is to estimate parameters and testing goodness of fit. On the substantive side the main problem is to identify the latent classes, i.e. to interpret them in terms which make practical sense.

The parameters estimations could be found by maximum likelihood approaches. The log-likelihood function derived from (1) is complicated, but it can be maximized using standard optimization routines. McHugh(1956) showed the standard Newton-Raphson technique to solve this optimization problem. However, an easier method which enables larger problems to be tackled is offered by the EM algorithm. The fundamental reference for EM is Dempster(1977) supplemented by Wu (1983), but the EM algorithm for latent class model was given by Goodman (1978). From (1) the log-likelihood with sample of size n is

$$l = \sum_{h=1}^{n} log\{\sum_{j=1}^{K} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_{ih}} (1 - \pi_{ij})^{1-x_{ih}}\} \tag{3}$$

This log-likelihood function has to be maximized subject to $\sum \eta_j = 1$. David(1987) found the parameter estimations in latent class model by taking partial derivatives:

$$\hat{\eta}_j = \sum_{h=1}^{n} h(j|\mathbf{x_h})/n \tag{4}$$

$$\hat{\pi}_{ij} = \sum_{h=1}^{n} x_{ih} h(j|\mathbf{x}_h)/n\hat{\eta}_j \tag{5}$$

where $i = 1, 2, \ldots, p;\ j = 1, 2, \ldots, K$.

By realizing that $h(j|\mathbf{x}_h)$ is a complicated function of $\{\eta_j\}$ and $\{\pi_{ij}\}$, which is given by

$$h(j|\mathbf{x}_h) = \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_{ih}} (1 - \pi_{ij})^{1-x_{ih}} / \sum_{k=1}^{K} \eta_j \prod_{i=1}^{p} \pi_{ik}^{x_{ih}} (1 - \pi_{ik})^{1-x_{ih}} \tag{6}$$

However, if $h(j|\mathbf{x}_h)$ were known it would be easy to solve (4) and (5) for $\{\eta_j\}$ and $\{\pi_{ij}\}$. The EM algorithm could be applied on this fact by the following steps:

Step 1: choose an initial set of posterior probabilities $\{h(j|\mathbf{x}_h)\}$;
Step 2: update (4) and (5) to obtain a first approximation to $\{\hat{\eta}_j\}$ and $\{\hat{\pi}_{ij}\}$;
Step 3: substitute these estimates into (6) to obtain improved estimates of $\{h(j|\mathbf{x}_h)\}$;
Step 4: return to step 2 to obtain second approximations to the parameters and continue the iteration until convergence is attained.

With the feasible and efficient estimating techniques, latent class models have been proposed in areas such as contingency table (Rinaldo, Zhou, Fienberg 2007), longitudinal studies with dropout (J.Roy 2003) and intermittent missing data (Lin 2004). Also, a number of recent papers have established fundamental connections between the statistical properties of latent class models and their algebraic and geometric features (Smith 2003, 2005; Rusakov 2005). Though there are potentially benefits to implement latent class analysis in different discipline and fields, it is at the cost of making some strong assumptions. One of these assumptions is choosing the number of latent classes. As mentioned above, different methods are proposed to assess latent class models with different number of latent classes. However, no assessment has been investigated on latent class models for missing values. In the next section, we present the underlying methods and models of our simulation studies.

3. Methods and Models of Simulation Studies

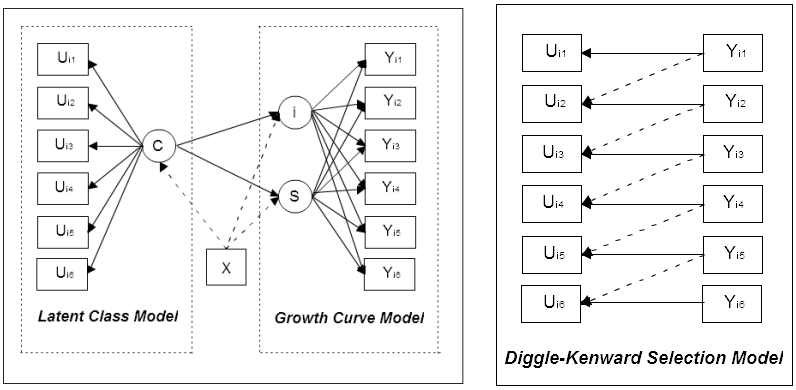
Rubin (1976) proposed three different missing mechanisms: missing completely at random (MCAR), missing at random(MAR) and not missing at random(NMAR). Data are said to be missing completely at random when the probability that responses are missing is unrelated to either the specific values that should have been obtained or the set of observed responses. For instance, in the longitudinal studies, let T be the total discrete time points, Y_{ij} be an observation for subject i at time j, and U_i be an $T \times 1$ vector of response indicators for subject i: $U_i = (U_{i1}, U_{i2}, \ldots, U_{iT})'$ with $U_{ij} = 1$ if the corresponding response Y_{ij} is observed and $U_{ij} = 0$ if Y_{ij} is missing. In addition, associated with Y_i is an $T \times p$ matrix of covariates, X_i. Given U_i, the complete set of responses Y_i can be partitioned into two components Y_i^o and Y_i^M, corresponding to those responses that are observed and missing, respectively. Longitudinal data are MCAR when U_i is independent of both Y_i^o and Y_i^M, i.e. (notations for i and j have different meanings from those in latent class model)

$$Pr(U_i|Y_i^o, Y_i^M, X_i) = Pr(U_i)$$

Data are said to be missing at random when the probability that responses are missing depends on the set of observed responses, but is unrelated to the specific missing values that should have been obtained. For instance, longitudinal data are MAR when U_i is conditionally independent of Y_i^M, given Y_i^o, i.e.

$$Pr(U_i|Y_i^o, Y_i^M, X_i) = Pr(U_i|Y_i^o, X_i)$$

The third type of missingness of data is referred to not missing at random. Missing data are said to be NMAR when the probability that responses are missing is related to the specific values that should have been obtained. That is, the conditional distribution of U_i is related to Y_i^M given Y_i^o, and $Pr(U_i|Y_i^o, Y_i^M, X_i)$ depends on at least some elements of Y_i^M. Our interests focus on two of three types of missingness (MCAR and NMAR) and corresponding mixture models. In the simulation studies that we have performed, datasets with different missing mechanisms are simulated and investigated by fitting latent class models. Three underlying assumptions of missingness in the datasets have been investigated : MCAR missing mechanism, NMAR missing mechanism and a mixture of both



(a) Diagram of simulated models

(b) Models for simulating missing data

Figure 1: Models studied in the simulations: latent class model and growth curve model (left); Diggle-Kenward selection model (right)

missing mechanisms, MCAR and NMAR. We considered a longitudinal study for 6 time points with mixed effects (or growth curve model):

$$y_{ij} = g_{0i} + g_{1i}t_j + \beta_2 x_{1ij} + \beta_3 x_{2ij} + \varepsilon_{ij} \tag{7}$$

where

$$g_{0i} = \beta_0 + b_{0i}$$
$$g_{1i} = \beta_1 + b_{1i}$$

In this model, y_{ij} is the observation for subject i and time j, x_{1ij}, x_{2ij} are two covariates, b_{0i} is the random intercept for subject i with mean μ_{b_0} and variance $\sigma^2_{b_0}$, b_{1i} is the random slope for subject i with mean μ_{b_1} and variance $\sigma^2_{b_1}$. (In the simulated growth curve model, we assume the following parameters: random intercept b_0 and random slope b_1 are normally distributed with mean vector $[1, 2]$, and variance covariance structure $\begin{bmatrix} 1 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}$.) In this model, two time-invariant covariates x_1 and x_2 were also include for the analysis purpose. To represent missing values, we used the following Diggle-Kenward selection model to indicate missingness of a value at time j:

$$log[\frac{P(U_{ij} = 1|y_{ij}, y_{i,j-1})}{P(U_{ij} = 0|y_{ij}, y_{i,j-1})}] = \alpha_j + \gamma_1 y_{ij} + \gamma_2 y_{i,j-1} \tag{8}$$

where α_j is a const intercept in the above logit expression, γ_1 and γ_2 are the coefficients of the observations y_{ij} and $y_{i,j-1}$, respectively.

3.1 Simulation Model of MCAR Missing Mechanism

To illustrate the simulation methods, we started from a simple case: fitting latent class models in simulated data that contains one missing mechanism. To simulate datasets followed by the assumed model in equation (7), Monte Carlo technique is applied. In the simulation with MCAR missing mechanism, we set the coefficients of covariates in equation (8) to be zeros, that is: $\gamma_1 = \gamma_2 = 0$ and set the intercept in the logit expression $\alpha_j = 1$, which corresponds to a probability of 0.27 of having missing data on the dependent variables (observations), i.e.

$$P(U_{ij} = 1|y_{ij}, y_{i,j-1}) = \frac{1}{1 + exp(-1)} \tag{9}$$

In this case, the missing probability is not related to either current or previous observations. This would reflect missing completely at random. A total of 1000 samples of MCAR were created using Monte Carlo method and each sample has 1000 observations. There are 64 different missing patterns in the simulated data, including the complete case. Latent class models with different number of classes have been applied for this data, in order to evaluate how the responses change through 6 time points from a grouping perspective. Covariates, as potential factors for explaining responses, were also investigated for whether they have effects on determing the number of latent classes.

3.2 Simulation Model of NMAR Missing Mechanism

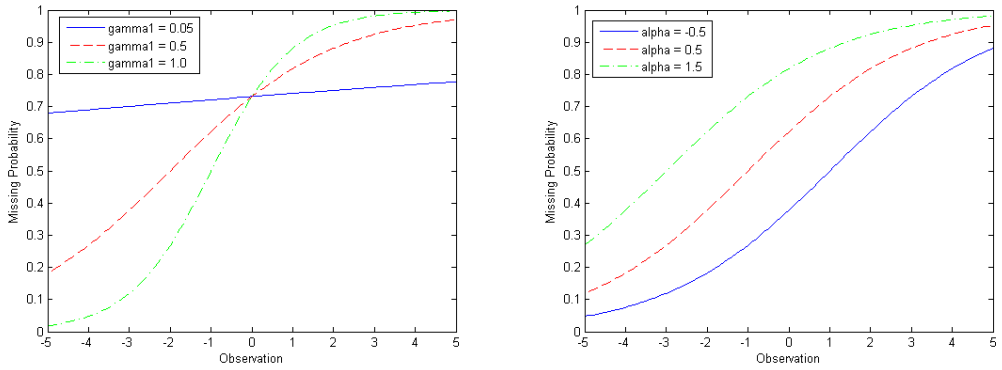
Another type of simulations we are interested in was comparing latent class models for missing values under NMAR. In some cases, even accounting for all the available observed information, the reason for observations being missing still depends on the unseen observations themselves. This motivates us to fit latent class models for this type of missingness, and the conditional probability is defined as follows: considering the current observation of y_{ij} for subject i at time j, missingness of y_{ij} could partially or fully depends on the unobserved values of y_{ij}, the conditional probability has the same expression with equation (8), i.e.

$$P(U_{ij} = 1|y_{ij}, y_{i,j-1}) = \frac{1}{1 + exp\{-(\alpha_j + \gamma_1 y_{ij} + \gamma_2 y_{i,j-1})\}}$$

where coefficients $\alpha_j, \gamma_2 \in R$ could take arbitrary values. In the above expression of conditional probability, changing the value of γ_1 or γ_2 will change the association between responses and missing values. For instance, we assume equation (8) only involves parameters α_j and γ_1, which also means that the missingness for current observation is only related with current observation. Figure2(a) shows that the parameter γ_1 determines the steepness of the curve over the middle of the range. This means that a given change in the value of y_{ij} will produce a larger change in the missing probability of a positive response when this parameter is large than when it is small. Figure 2(b) demonstrates the missing probability curves by changing the values of α_j. With the increase of α_j, there is a larger chance for an observation to be missing, compared with a lower α_j. Therefore, changing parameter values in equation (8) should alter the association among the missing value indicators and might have an influence on deciding the number of latent classes. The related simulation studies and corresponding results will be given in the next section. For simulations in this part, each simulation generated 1000 replicates and each replicate had 1000 observations, followed by NMAR.

3.3 Simulation Mixture Model of MCAR and NMAR

In a longitudinal study, data are collected from baseline to the end of the study. The presence of a big amount of missing values is common, accompanying with complicate missing mechanisms. Though it's often difficult to distinguish what missing mechanisms are involved in the collected data, ideally a combination with MCAR and NMAR is a possible case. This motivates us to investigate a mixture model of combining these two different types of missing mechanisms. For simulations in this part, we have generated 1000 samples and each sample is consisted of different proportions of MCAR and NMAR, either 500 observations for each of missing mechanism or 800 observations for MCAR and 200 observations of NMAR, depending on the research goals. We will announce this proportion in the simulation results. The conditional probabilities for MCAR and NMAR are defined in previous two formulas.



(a) Missing probability curves for different values of γ_1 and $\alpha_j = 1, \gamma_2 = 0$

(b) Missing probability curves for different values of α_j and $\gamma_1 = 0.5, \gamma_2 = 0$

Figure 2: Missing probability curves

Besides exploring the method to choose the optimal number of latent classes, covarites in the growth curve model, different settings of missing probabilities, and the associations among the y's may be of interested and investigated on selection number of latent classes. To generate different associations among the observations, one could change the parameters of random slope in growth curve model (7). For instance, with a higher value of μ_{b_1}, samples with highly associated observations would be generated. All these factors of interests should be explored by fitting latent class models on samples with different settings.

4. Analysis of Simulation Results

To compare performances of latent class models with different number of latent classes, Clogg (1995) and Aitkin (1981, 1985) indicated that chi-squared likelihood ratio statistics were not theoretically correct for LCM selections. A $K-1$ classes LCM is obtained by putting one parameter value at the boundary of a K-classes model. The likelihood ratio between the two LCMs may not follow a single χ^2 distribution if the constrained model ($K-1$ classes) is obtained from the full model (K classes) by placing parameters at their boundary values. Several alternative methods, including information criteria, parametric resampling, etc. were suggested to solve the problem. Information criteria are probably one of the most convenient methods than other methods such as parametric resampling. We apply as the efficient approaches and compare the performances of several information criteria to evaluate latent class models, including AIC, BIC, CAIC, DBIC, and other four information criteria.

4.1 Information Criteria

Yang (2004) discussed many information criteria that can be used to compare LCMs. Akaike information criterion (AIC) was one of the earliest propositions of information criteria. AIC has the following form

$$AIC_k = -2logL(\theta_k) + 2p_k$$

where $log\ L(\theta_k)$ is log-likelihood from MLE, p_k is the total number of free parameters in model k. However, Woodruffe (1982) showed that AIC is not theoretically consistent; consequently, AIC will not select the correct model when sample size (N) is near infinity.

Schwarz (1978) proposed Bayesian information criterion (BIC) which has the following form

$$BIC_k = -2logL(\theta_k) + p_k\ log\ N$$

Haughton (1988) showed BIC is consistent when sample size goes large and hence can lead to a correct choice of model when N goes infinity.

Bozdogan (1987) derived a consistent version of AIC, called CAIC from the Kullback-Leibler information measure with the form

$$CAIC_k = -2logL(\theta_k) + p_k\ (log\ N + 1)$$

Since CAIC puts severe penalty on over-parameterization than BIC or AIC, it tends to favor a model with fewer parameters.

Draper (1995) modified the penalty part of BIC, and DBIC is defined as follows

$$DBIC_k = -2logL(\theta_k) + p_k(log\ N - log\ 2\pi)$$

When sample size N goes infinity, the added term is asymptotically insignificant, but it has a notable effect on the log-likelihood for small to moderate sample sizes.

We also included HQ information criterion which was invented by Hannan (1979), HT-AIC information criterion discovered by Hurvich (1989), sample size adjusted BIC (BICa) and CAIC (CAICa) to compare the performance among latent class models with different latent classes. For each simulation that we investigated, 1000 samples were simulated on different latent class models with latent classes either from 1 to 5 or from 2 to 5, depending one which simulation is processed. When we performed simulations of MCAR or NMAR alone, LCMs with latent classes from 1 to 5 were compared. For simulations of mixture of MCAR or NMAR, we compared LCMs with latent classes from 2 to 5. One can check in the latter case, LCMs with one latent class won't be suggested by all of the information criteria among 1000 samples. For each information criterion, a smaller value indicates a better model of fit on the simulated data. After fitting latent class models on 1000 samples, tallies were made for the numbers latent classes indicated by each criterion, with number of latent classes, ranging either from 1 to 5 latent classes, or from 2 to 5 latent classes. To illustrate directly, we summarize the tallies and corresponding proportions for each information criterion in tables and marked the favored LCM in red.

4.2 Model selection for LCMs

We first consider the selection of LCMs with initial parameters in equation (7) and (8), which are used to simulate the samples. Three different underlying missing types are investigated: MCAR, NMAR and a mixture of MCAR and NMAR. To simulate a growth curve model with MCAR type of missingness, we assume the random intercept is normally distributed with mean 1 and variance 1; the random slope is also normally distributed with mean 2 and variance 0.2; for the MCAR missingness, we choose the default intercept term $\alpha_j = 1$ in the logit expression (9). To simulate a growth curve model with NMAR type of missing, we use the same model parameters in (7) as former one and assume the missing status for current observation is only related with current observation, not previous one, i.e. $\alpha_j = 1$, $\gamma_1 = 0.2$ and $\gamma_2 = 0$. To simulate a growth curve model with a mixture of two types of missing mechanisms, 500 observations are generated from each missing mechanism using the same model parameters. The voting results are shown in Table1-3.

Table 1 describes the voting results of LCMs for MCAR missing mechanism. There are 10 replicates that are failed in convergence when fitting the models. All the information criteria support the LCM with one latent class, with spreading trends in both AIC and HT.

Table 1: Number of Latent Class Tallies on MCAR simulation*

Information Criteria	Latent class model LC1	LC2	LC3	LC4	LC5
AIC	810 (0.82)	155 (0.16)	20 (0.02)	2 (0.002)	3 (0.003)
BIC	990 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
CAIC	990 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
DBIC	990 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
HQ	990 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
HT	823 (0.83)	149 (0.15)	14 (0.01)	1 (0.001)	3 (0.003)
BICa	988 (0.998)	2 (0.002)	0 (0.00)	0 (0.00)	0 (0.00)
CAICa	989 (0.999)	1 (0.001)	0 (0.00)	0 (0.00)	0 (0.00)

*Latent class models are fitted without incorporating covariates, $\alpha_j = 1$, $\gamma_1 = 0$, $\gamma_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

Table 2: Number of Latent Class Tallies on NMAR simulation*

Information Criteria	Latent class model LC1	LC2	LC3	LC4	LC5
AIC	229 (0.23)	306 (0.31)	197 (0.20)	141 (0.14)	117 (0.12)
BIC	987 (0.997)	3 (0.003)	0 (0.00)	0 (0.00)	0 (0.00)
CAIC	990 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
DBIC	978 (0.99)	12 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)
HQ	916 (0.925)	72 (0.073)	2 (0.002)	0 (0.00)	0 (0.00)
HT	253 (0.25)	343 (0.35)	197 (0.20)	120 (0.12)	77 (0.08)
BICa	899 (0.908)	88 (0.089)	3 (0.003)	0 (0.00)	0 (0.00)
CAICa	902 (0.911)	85 (0.086)	3 (0.003)	0 (0.00)	0 (0.00)

*Latent class models are fitted without incorporating covariates, $\alpha_j = 1$, $\gamma_1 = 0.2$, $\gamma_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

Table 2 summarizes the results for NMAR missing mechanism, most information criteria suggest the model with one latent class, except AIC and HT. Both AIC and HT present significant spreading trends in the voting results, and reverse the results to LCM with two latent classes. As discussed before, AIC tends to give a inaccurate suggestion due to its inconsistency when sample size gets large. HT information criteria is derived from AIC and it inherits inconsistency property as well. Simulation results demonstrate that a LCM with a homogeneous group is favored for single missing mechanism.

The results of selection of LCMs for a mixture of two missing mechanisms are summarized in Table 3. All information criteria support LCM with two latent classes, while there are large dispersion of tallies over AIC and HT. By reviewing the way we simulate data for a mixture of two missing mechanism, two datasets with the single missing mechanism are merged. Simulation results indicate this mixing and suggest that LCM with two heterogeneous groups has a better of model of fit. Without loss of generality, we choose the results in Table 3 and the corresponding models as the reference results and models, to investigated the following factors of interests.

Table 3: Number of Latent Class Tallies on Mixture of MCAR and NMAR*

Information Criteria	Latent class model				
	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	387 (0.39)	282 (0.28)	201 (0.20)	125 (0.13)
BIC	0 (0.00)	992 (0.997)	3 (0.003)	0 (0.00)	0 (0.00)
CAIC	0 (0.00)	992 (0.997)	3 (0.003)	0 (0.00)	0 (0.00)
DBIC	0 (0.00)	987 (0.992)	8 (0.008)	0 (0.00)	0 (0.00)
HQ	0 (0.00)	964 (0.969)	29 (0.029)	2 (0.002)	0 (0.00)
HT	0 (0.00)	438 (0.44)	286 (0.29)	177 (0.18)	94 (0.09)
BICa	0 (0.00)	952 (0.957)	41 (0.041)	2 (0.002)	0 (0.00)
CAICa	0 (0.00)	954 (0.959)	39 (0.039)	2 (0.002)	0 (0.00)

*Latent class models are fitted without incorporating covariates, $\alpha_j = 1$, $\gamma_1 = 0(MCAR), = 0.2(NMAR)$, $\gamma_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

4.2.1 Covariates Effect

In general, covariates potentially affects the relationship between the dependent variable and other independent variables of primary interest. Two covariates are included in our simulation studies, namely, X_1 and X_2, and both covariates are generated from standard normal distribution in Monte Carlo simulations. In equation (7), covariates provide extra information on observations y_{ij} and those observations are potentially influenced on missing indicators U_{ij}, as expressed in equation (8). The covariates effect on selection of LCMs may be of interests. To investigate this effect, we evaluate LCMs for the mixture of the two missing mechanisms, with or without incorporating covariates in LCMs. One could do the same study on LCMs for single missing mechanism. While fitting LCMs without covariates for 1000 replicates, 995 are successfully converged; 992 among 1000 samples are fitted for LCMs with covariates.

Table 3 describes the results of LCMs without incorporating covariates. All information criteria support a LCM with two latent classes, i.e. a LCM with two heterogeneous groups has a better model of fit. Table 4 lists the tallies of LCMs with covariates and most information criteria suggests the same number of latent classes as the case of without covariates, except AIC and HT. Due to the inconsistency of AIC and HT, they don't correctly identify a model, in particular, they select the model with more latent classes than it actually had. Simulations have shown the covariates don't alter the choice of number of latent classes of LCMs which the models are applied for data with two missing mechanisms, MCAR and NMAR. However, the auxiliary information provided by covariates "un-stabilizes" the selection of LCMs by information criteria. For instance, one of the best performing information craiteria, BIC supports for a two latent class model in most cases (with probability $p \approx 0.997$) when there is no covariates considered; and it loses this certainty while covariates are incorporated (with probability $p \approx 0.991$). Other information criteria have more significant loss on this certainty while incorporating covariates into models. AIC and HT are severely sensitive to the covariates effects. AIC drops this probability from 0.39 to 0.14 for supporting a LCM with two latent classes.

4.2.2 Association Effect among Responses

As we discussed before, model parameters in (7) and (8) are initialized at the beginning of data simulations. In this part we consider the changes on parameters in equation(7), more

Table 4: Number of Latent Class Tallies on Mixture of MCAR and NMAR with covariates

Information Criteria	Latent class model LC1	LC2	LC3	LC4	LC5
AIC	144 (0.14)	214 (0.22)	259 (0.26)	375 (0.38)	
BIC	0 (0.00)	983 (0.991)	1 (0.001)	1 (0.001)	7 (0.007)
CAIC	0 (0.00)	792 (0.80)	157 (0.16)	30 (0.03)	13 (0.01)
DBIC	0 (0.00)	974 (0.982)	10 (0.010)	1 (0.001)	7 (0.007)
HQ	0 (0.00)	928 (0.936)	52 (0.052)	5 (0.005)	7 (0.007)
HT	0 (0.00)	286 (0.288)	290 (0.292)	222 (0.224)	194 (0.196)
BICa	0 (0.00)	775 (0.78)	168 (0.17)	35 (0.04)	14 (0.01)

*With covariates,low missing probabilities, high association among responses.

Table 5: Number of Latent Class Tallies on Mixture of MCAR and NMAR with low associations among responses(without covariates)

Information Criteria	Latent class model LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	701 (0.706)	232 (0.234)	41 (0.041)	19 (0.019)
BIC	0 (0.00)	993 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
CAIC	0 (0.00)	993 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
DBIC	0 (0.00)	993 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
HQ	0 (0.00)	986 (0.993)	7 (0.007)	0 (0.00)	0 (0.00)
HT	0 (0.00)	737 (0.742)	215 (0.217)	29 (0.029)	12 (0.012)
BICa	0 (0.00)	984 (0.991)	9 (0.009)	0 (0.00)	0 (0.00)
CAICa	0 (0.00)	985 (0.992)	8 (0.008)	0 (0.00)	0 (0.00)

*$\alpha_j = 1, \gamma_1 = 0(MCAR), = 0.2(NMAR), \gamma_2 = 0, \mu_{b_0} = 1, \mu_{b_1} = 1, \sigma^2_{b_0} = 1, \sigma^2_{b_1} = 0.2, cov(b_0, b_1) = 0.1$.

specifically, we simulate growth curve models with missingness by altering the parameters in the random slope term to different values, i.e. the mean and variance of b_{1i}. To avoid the redundant tables, we provide one of the simulations with two different initialized mean values of b_{1i}: while $\mu_{b_{1i}} = 1$ represents a lower association among observations, $\mu_{b_{1i}} = 2$ indicates a higher association.

Table 5 displays the results for the lower association. All information criteria agree a LCM with two heterogeneous groups will fit the missing values better. By comparison, the results for the higher association case are shown in Table 3. It is indicated that with increasing the degree of associations among responses, the choice of number of latent classes won't change. However, the problem of changes in the "selection certainty" draws our attention again. One of the worst behaviored information criteria AIC losses its choice certainty from 0.706 to 0.39.

4.2.3 Missing Probability Effect

To investigate the selection of LCMs for missing values, Diggle-Kenward selection model are intensively used in our simulation studies, as described in equation (8). In this expression, the missing probability for the current observation y_{ij} is determined by the value of

Table 6: Number of Latent Class tallies on Mixture of MCAR and NMAR with high missing probability(without covariates)

Information Criteria	Latent class model				
	LC1	LC2	LC3	LC4	LC5
AIC	0 (0.00)	0 (0.00)	5 (0.005)	546 (0.553)	437 (0.442)
BIC	0 (0.00)	15 (0.015)	849 (0.859)	124 (0.126)	0 (0.00)
CAIC	0 (0.00)	15 (0.015)	849 (0.859)	124 (0.126)	0 (0.00)
DBIC	0 (0.00)	0 (0.00)	470 (0.476)	511 (0.517)	7 (0.007)
HQ	0 (0.00)	0 (0.00)	186 (0.19)	766 (0.77)	36 (0.04)
HT	0 (0.00)	0 (0.00)	8 (0.008)	609 (0.616)	371 (0.376)
BICa	0 (0.00)	0 (0.00)	167 (0.169)	777 (0.786)	44 (0.045)
CAICa	0 (0.00)	0 (0.00)	169 (0.169)	776 (0.786)	43 (0.045)

*$\alpha_j = 1$, $\gamma_1 = 0(MCAR), = 0.6(NMAR)$, $\gamma_2 = 0$, $\mu_{b_0} = 1$, $\mu_{b_1} = 2$, $\sigma^2_{b_0} = 1$, $\sigma^2_{b_1} = 0.2$, $cov(b_0, b_1) = 0.1$.

previous observation $y_{i,j-1}$, current observation y_{ij} and initialized parameter values α_j, γ_1, and γ_2. Changing any one of these values will lead to a change in missing probabilities and potentially affect the structure of LCMs. For instance, increasing the coefficient γ_1 will lead to a higher missing probability for the current observation y_{ij}, while holding other parameters fixed. Table 3 and 6 present the model selection results for a paired values of γ_1 (0.2, 0.4) which are set to simulate the missingness. $\gamma_1 = 0.2$ corresponds to a lower missing probability, when $\gamma_1 = 0.4$ corresponds to a higher missing probability. $\alpha_j = 1$ and $\gamma_2 = 0$ are fixed in this comparison.

Table 3 illustrates the voting results for the lower missing probability, a LCM with two latent classes are suggested by all information criteria. Clearly it is suggested that LCM from Table 3 is changed in the higher missing probability case, based on the cell values in Table 6. While both BIC and CAIC support for a LCM with three heterogeneous groups, all the other information criteria vote for four latent classes. This change shows an evidence of the influence of missing probability on the LCM selection, i.e. with a higher missing probability, LCMs with more heterogeneous groups are preferred.

To investigate the selection of LCMs, we have checked the missing mechanisms and related factors that derived from changing parameters in either model equation (7) or missing values generating mechanism (8), and through simulation studies we conclude their influences on deciding the number of latent classes. To fit the datasets which consist of two assumed missing mechanisms groups, the cases where a LCM with three heterogeneous groups is suggested are worthy to be researched further. However, the assumed missing mechanisms usually cannot be identified in practice. In particular, there is no statistical methods or tests on NMAR and the mixture of MCAR and NMAR. By contrast, missing patterns could be directly observed and it may provide another perspective to understand LCMs. In the last part of this section, we focus on exploring the behaviors of missing patterns on LCMs with three latent classes.

4.3 Missing Patterns in LCMs

In the above simulations, longitudinal studies with 6 time points are considered. We define U_{ij} as the missing indicator for subject i at time j. The possible missing patterns are $2^6 = 64$. For a large sample size, many of the missing patterns will be repeated. In our simulations, each sample has 1000 observations and a list of the observed missing patterns

Table 7: Posterior probability for LCMs with three classes (first 10 frequent missing patterns)

Missing Pattern	Frequency f	$h(1\|\mathbf{x})$	$h(2\|\mathbf{x})$	$h(3\|\mathbf{x})$	$C_{1\|\mathbf{x}}$	$C_{2\|\mathbf{x}}$	$C_{3\|\mathbf{x}}$
011111	223 (3,220)	0.056	0.934	0.009	12.488	208.282	2.007
001111	96 (2,94)	0.948	0	0.052	91.008	0	4.992
101111	54 (1,53)	0.955	0	0.045	51.57	0	2.43
000011	50 (35,15)	0.115	0	0.885	5.75	0	44.25
000111	44 (16,28)	0.707	0	0.293	31.108	0	12.892
000001	39 (36,3)	0	0	1	0	0	39
000000	33 (32,1)	0	0	1	0	0	33
001011	33 (21,12)	0.494	0	0.506	16.302	0	16.698
000010	30 (29,1)	0	0	1	0	0	30
010111	27 (5,22)	0.167	0.625	0.207	4.509	16.875	5.589

*Missing data are simulated using Diggle-Kenward model ($\alpha_j = 1,\ \gamma_1 = 0.4,\ \gamma_2 = 0.4$). 0 is observed response, 1 is missing response.

together with their associated frequencies is given in Appendix. The posterior probability $h(j|\mathbf{x})$ of an individual with missing pattern $\mathbf{x}$ belonging to jth groups could be obtained when the corresponding LCM is fitted, based on the definition in equation (2). In our case, three posterior probabilities for each latent class would be calculated for each missing patterns and these results are given in the Appendix as well. A missing pattern $\mathbf{x}$ is allocated in the class for which the posterior probability is greatest.

Let $C_{j|\mathbf{x}}$ be the posterior count for jth latent class given missing pattern $\mathbf{x}$, and can be calculated as the product of observed frequency f and posterior probability $h(j|\mathbf{x})$. Based on the posterior counts we could explore the missing patterns in deciding allocation of latent classes. For instance, LCMs with three heterogeneous groups in our simulation studies are of interests to investigate further. For instance, Table 7 lists the posterior probabilities and counts for the first 10 missing patterns in one of our simulation studies. 0 in missing patterns represents for observed responses, 1 means missing in responses. Tow numbers in the parenthese for frequency item are frequencies counted from MCAR and NMAR, respectively. The total frequency for the 8th missing pattern $\mathbf{x}_8$ is 223, where responses are only observed at the first time point. And 220 out of 223 come from NMAR mechanism, only 3 come from MCAR mechanism. Among on the posterior counts $C_{j|\mathbf{x}_8}$ $(j = 1, 2, 3)$ for this pattern, the second latent class has the most counts 208.282. Therefore, this pattern counts for the second latent class. In fact, the second latent class consists of three missing patterns: $\mathbf{x}_6$, $\mathbf{x}_7$ and $\mathbf{x}_8$. $\mathbf{x}_8$ is the majority in this group, i.e. observations with this type of missing pattern will be allocated in the second latent class.

From the inspection on all missing patterns in each simulation, one could find that the first two latent classes mainly consist of missing patterns from NMAR mechanism, and missing patterns from MCAR forms the third class. Compared with cases where LCMs with two classes are preferred, we find that there is a seperation in the NMAR mechanism, which lead to an additional class. Further, we could observe that in LCMs, latent classes are represented by homogeneous responses, i.e. homogeneous missing patterns fall into one class.

5. Discussion

This paper described simulation studies on selection number of latent classes for missing values and comparison results based on eight information criteria. The Bayesian informa-

tion criteria, consistency version of AIC (CAIC) and sample adjusted BICa are noteworthy information criteria to choose correct latent classes. AIC presents its inconsistency property in the simulation studies. HT has less consistent performance as well. These inconsistent information criteria are not suggested for real case studies.

Covariates and degree of association among responses do not account for deciding how many latent classes are best for fitting the data with different missing mechanisms. However, changing these parameters will influence on "selection certainty" of all inforamtion criteria. Increasing the degree of associations among responses or incorporating covariates in the simulation model will lead to the loss of "selection certainty". We also find that the selection by AIC and HT are more sensitive to these changes. Compared with those less-influential factors, missing probabilities directly have effects on deciding number of latent classes. A higher missing probability tends to make the number of latent classes larger. Bayesian Information Criterion (BIC) and consistent version of AIC (CAIC) suggest conservative LCMs with three classes, while other information criteria indicate that four classes are preferred. One would like to choose the smallest number of classes that allows the assumption of conditional independence to hold. A latent class model with too many classes can be a problem. One is it's difficult to interpret these classes due to the small size of classes.

Missing patterns are also investigated for the chosen latent classes. Posterior counts for each pattern are calculated and compared. The allocation for each pattern is based on the largest posterior probability, i.e. assgin a pattern to the class where the posterior probaility is the greatest. Studies indicate that latent classes in LCMs are represented by homogeneous missing patterns. And the underlying missing mechanism could account for the classes. For the two classes LCMs, one class mainly comes from missing patterns generated by MCAR, when the other is consisted of missing patterns from NMAR. LCMs with three classes in the simulations could be illustrated as a separation of missing patterns in NMAR.

If one wants to apply LCMs to capture the group characteristics for missing values, a simulation on deciding the number of latent classes is recommended before fitting the model. A further research on latent variables for missing indicators may be of interests. As shown in Figure 1, the assumed latent class C is related with latent variables i, s which are used as random intercept and random slope terms in the growth curve model. If the observations $\mathbf{Y}$ are continuous, both random terms could be continuous and the linked latent variables for missing indicator could be continuous as well.

REFERENCES

Aitkin, M., Anderson, D., and Hinde, J. (1981), "Statistical Modeling of Data on Teaching Styles," *Journal of the Royal Statistical Society. Series A (General)*, 144, 419-461.

Aitkin, M. and Rubin, Donald B. (1985), "Estimation and Hypothesis Testing in Finite Mixture Models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 47, 67-75.

Bartholomew, David J. (1987), *Latent Variable Models and Factor Analysis*, New York: Oxford University Press.

Bozdogan, H. (1987), "Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions," *Psychometrika*, 52, 345-370.

Clogg, C. C., Sobel, M. E., and Arminger, G. (1995), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum Press.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1-38.

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 45-97.

Fienberg, S. E., Hersh, P., Rinaldo, A., and Zhou, Y. (2007), "Maximum Likelihood Estimation in Latent Class Models for Contingency Table Data," in *arVix:0709.3535v1*.

Garrett, Elizabeth S., and Zeger, Scott L. (2000), "Latent class model diagnosis," *Biometrics*, 56, 1055-1067.

Hannan, E. J., and Quinn, B. G. (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 190-195.

Haughton, Dominique M. A. (1988), "On the Choice of a Model to Fit Data from an Exponential Family," *The Annals of Statistics*, 16, 342-355.

Hurvich, Clifford M., and Tsai, C. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297-307.

Lazarsfeld P. F. (1950), *The Logical and Mathematical Foundation of Latent Structure Analysis*, Princeton University Press.

Lin, H., McCulloch, C. E., and Rosenheck, R. A. (2004), "Latent Pattern Mixture Models for Informative Intermittent Missing Data in Longitudinal Studies," *Biometrics*, 60, 295-305.

Magidson, J., and Goodman, Leo A. (1978), *Analyzing Qualitative/Categorical Data: Log-linear Models and Latent Structure Analysis*, Cambring, Mass.

McHugh, R. B. (1956), "Efficient Estimation and Local Identification in Latent Class Analysis," *Psychometrika*, 21, 331-347.

Roy, Jason (2003), "Modeling Longitudinal Data with Non-ignorable Dropouts Using a Latent Dropout Class Model," *Biometrics*, 59, 829-836.

Roy, Jason (2007), "Latent class models and their application to missing-data patterns in longitudinal studies," *Statistical Methods in Medical Research*, 16, 441-456.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581-592.

Rusakov D., and Geigerm, D. (2005), "Asympototic Model Selection for Naive Bayesian Networks," *Journal of Machine Learning Research*, 6, 1-35.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461-464.

Settimi, R. and Smith, J. Q. (2000), "Geometry, Moments and Conditional Independence Trees with Hidden Variables," *The Annals of Statistics*, 28, 1179-1205.

Smith, J. Q. and Croft, J. (2003), "Bayesian Networks for Discrete Multivariate Data: an Algebraic Approach to Inference," *Journal of Multivariate Analysis*, 84, 387-402.

Woodruffe, M. (1982), "On Model Selection and the Arcsine Laws," *The Annals of Statistics*, 10, 1182-1194.

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95-103.

Yang, C. C. (2006), "Evaluating Latent Class Analysis Models in Qualitative Phenotype Identification," *Computational Statistics and Data Analysis*, 50, 1090-1104.