Bayesian Estimation of Latent Class Model for Survey Data Subject to Item Nonresponse

Samah Zakaria
Statistics Department, Faculty of Economics and Political Science, Cairo University, Egypt
samahzkaria@feps.edu.eg

Mai Sherif Hafez
Statistics Department, Faculty of Economics and Political Science, Cairo University, Egypt
mai.sherif@feps.edu.eg

Ahmed Mahmoud Gad
Statistics Department, Faculty of Economics and Political Science, Cairo University, Egypt
ahmed.gad@feps.edu.eg

Abstract

Latent variable models are widely used in social sciences for measuring constructs (latent variables) such as ability, attitude, behavior, and wellbeing. Those unobserved constructs are measured through a number of observed items (variables). The observed variables are often subject to item nonresponse that may be nonignorable. Incorporating a missingness mechanism within the model used to analyze data with nonresponse is crucial to obtain valid estimates for parameters, especially when the missingness is nonignorable. In this paper, we propose a latent class model (LCM) where a categorical latent variable is used to capture a latent phenomenon, and another categorical latent variable is used to summarize response propensity. The proposed model incorporates a missingness mechanism. Bayesian estimation using Markov Chain Monte Carlo (MCMC) methods are used for fitting this LCM. Real data with binary items from the 2014 Egyptian Demographic and Health Survey (EDHS14) are used. Different levels of missingness are artificially created in order to study results of the model under low, moderate and high levels of missingness.

Keywords: Bayesian estimation; Latent class model; Nonignorable item nonresponse; Response propensity.

1 Introduction

Latent variable modeling is an important tool in multivariate data analysis. One of the main reasons behind using such a technique is trying to measure constructs or concepts that cannot be directly measured, which are often met in social sciences (e.g. wellbeing, satisfaction, attitude, democracy, etc.…). These are referred to as latent (unobserved) factors or variables and can be measured via a number of manifest (observed) variables or items.

Latent variable models are classified according to nature of the observed variables (categorical or continuous), nature of the latent variables (categorical or continuous) and inclusion or not of covariates. We consider a latent class model (LCM) where both latent and manifest variables are categorical, while incorporating a missingness mechanism that accounts for item nonresponse which also involves a categorical latent variable measuring an individual’s propensity to respond

Kuha et al. (2018) propose models for nonresponse in survey questions where the response propensity is a categorical variable and the latent variable of attitude is assumed to be continuous, both depending on a respondent's individual characteristics. Bacci and Bartolucci (2014) define such models where both latent variable of interest and response propensity variable are categorical, and conditionally independent given a set of covariates, and the nonresponse model may depend on both latent variables. These two studies depend on Expectation-Maximization algorithm in estimating the models.

There are two main approaches for estimating latent variable models, the first of which depends on iterative techniques such as the EM algorithm, first introduced by Dempster et al. (1977). As models get more complex, so does the implementation of the EM. An alternative methodology for estimating parameters of a latent variable model is to adopt a Bayesian approach based on MCMC. Unlike the EM algorithm, MCMC does not require exact numerical calculation for the E-step, or precalculation of derivatives for the M-step, thus providing easier implementation (Hafez, 2015). Moustaki and Knott (2005) compare the EM and MCMC estimation methods for latent variable models, where they use real examples with categorical data to illustrate this comparison. They use binary or ordinal observed variables and assume that the latent variable is continuous (latent trait model). They also fit the model for binary responses with missing values. The comparison is made in terms of parameter estimates and standard errors. They show that MCMC methods have become popular in the area of latent variable modeling mainly because they allow estimation of complex models with much flexibility.

Various researchers have focused on studying and developing Bayesian estimation for latent class models. Galindo-Garre and Vermunt (2006) compare the quality of various types of posterior mode point and interval estimates for the parameters of latent class models with both the classical maximum likelihood estimates and the bootstrap estimates proposed by De Menezes (1999). Their simulation study shows that parameter estimates and standard errors obtained by the Bayesian approach are more reliable than the corresponding parameter estimates and standard errors obtained by maximum likelihood and parametric bootstrapping. Pan and Huang (2013) propose a Bayesian framework to perform the joint estimation of the number of latent classes and model parameters by applying the reversible jump Markov chain Monte Carlo to analyze finite mixtures of multivariate multinomial distributions. Latent class analysis is based on the assumption that within each class the observed class indicator variables are independent of each other. Asparouhov and Muthén (2011) explore a new Bayesian approach that relaxes this assumption to an assumption of approximate independence. Instead of using a correlation matrix with correlations fixed to zero, they use a correlation matrix where all correlations are estimated using an informative prior with mean zero but non-zero variance. White et al. (2016) propose a Bayesian approach for the analysis of LCMs. It is shown how simple marginalization of the parameters in a LCM leads to a form of the model for which MCMC sampling algorithms can be used to quantify precisely the uncertainty in the number of groups in the data, as well as which variables give the best clustering. Thanoon and Adnan (2016) use ordered categorical variables to compare between linear with covariate and nonlinear interactions of covariates and latent variables in Bayesian structural equation models. Gibbs sampling method is applied for estimation and model comparison.

The model presented in this paper builds upon the same model framework presented in O'Muircheartaigh and Moustaki (1999), Moustaki and Knott (2000), Hafez et al. (2015), Bacci and Bartolucci (2014), and Kuha et al. (2018). Unlike their work, the model proposed in this paper assumes both latent variables to be categorical and dependent by assuming a certain structure among them. Whereas the EM algorithm has been used for estimation of the above models, Bayesian estimation using MCMC is adopted in this paper.

In this paper, we cover cases where observed items are categorical (binary) subject to item nonresponse, and where the latent variables used to summarize both the phenomenon of interest and the response propensity are assumed to be categorical. A structure is assumed among the two latent variables in such a way that allows latent class membership to affect the probability of response, thus accounting for a possibly nonignorable missingness. This model has been estimated by the authors in previous work using the EM algorithm, where different types of missingness (missing completely at random, missing at random, and missing not at random) have been compared. In this paper, Bayesian estimation using MCMC is used to fit the outlined LCM subject to item nonresponse. Different levels of missingness are artificially created in order to study the performance of the model under low, moderate and high levels of missingness.

The rest of this article is organized as follows. The specification of the proposed LCM is described in Section 2. Section 3 outlines Bayesian estimation for LCM parameters using MCMC. In Section 4, Bayesian estimation method is employed to the proposed model to measure people’s access to knowledge sources using real data from Egypt’s 2014 Demographic and Health Survey. Finally, a conclusion is given in Section 5.

2 Latent Class Model for Binary Items Subject to Nonresponse

The model studied here considers the case where all observed variables are categorical, particularly binary. The latent variable of interest is assumed to be categorical too. A missingness mechanism to account for item nonresponse is incorporated. The latent variable of response propensity is also assumed to be categorical. The two latent variables are assumed to be linked allowing dependence of response propensity on the latent phenomenon of interest. Different sets of covariates are allowed to affect both latent variables. Figure 1 gives a path diagram that illustrates the proposed model.

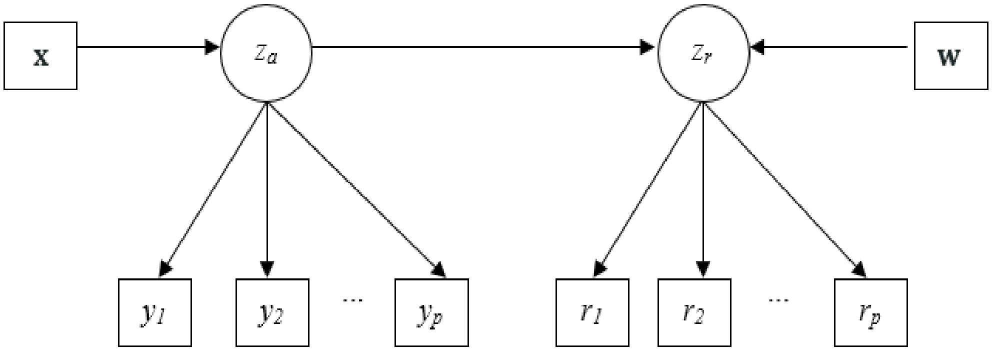

Figure 1: Path diagram for a model where the latent variable affects missingness mechanism

The latent variable of interest for attitude, ability or behavior is denoted by z_a while z_r is another latent variable for response propensity, y_i is an observed variable, p is the number of observed variables, **x** is a set of observed covariates affecting z_a, **w** is a set of observed covariates affecting z_r, that may be the same or different from those affecting z_a, and r_i is an indicator variable that takes value 1 when the manifest variable y_i is observed and takes value 0 when the manifest variable y_i is missing.

A LCM has two parts; a measurement part and a structural part. In our case, a third part is added to the model to incorporate the missingness mechanism.

2.1 Measurement model

The measurement part for a LCM describes the relationships between a set of categorical observed variables and a set of categorical latent variables. We assume the unidimensional case where one latent variable z_a is sufficient to explain relationships among the observed items y_i, where i =1, 2, …, p, and p is the number of manifest variables. In our case all observed variables y_i are binary, each having a Bernoulli distribution conditional on the latent variable z_a. A set of logistic regression equations are used to model these relationships. Thus, the probability of a positive response on manifest variable i can be presented as,

$$\text{logit}\, \pi_{ai}(z_a) = \alpha_{i0} + \alpha_{ia}\, z_a\ , \qquad (1)$$

Where $\pi_{ai}(z_a)$ is the probability of an individual’s positive response to a manifest variable i given their class membership of the latent variablez_a. The latent classes of the latent variable of interest z_a are mutually exclusive and exhaustive.

2.2 Missingness mechanism

The missingness mechanism can be incorporated in the model by introducing an indicator random variable r_i for missingness, corresponding to each manifest variable. For each individual, the indicator variable r_i takes value 1 if the manifest variable y_i is observed for this individual, and takes value 0 if it is missing, where i =1, 2, …, p, and p is the number of indicator variables.

Similar to the observed variablesy_i, each of the indicator variables r_i has a Bernoulli distribution. It is assumed that a categorical latent variable z_r that summarizes an individual’s response propensity, is responsible for explaining relationships among the p missingness indicators r_i. The probability that a variable y_i is observed ($r_i =$ 1) conditional on the latent class membership, can thus be modeled as,

$$\text{logit}\, \pi_{ri}(z_r) = \upsilon_{i0} + \upsilon_{ir}\, z_r\ , \qquad (2)$$

where $\pi_{ri}(z_r)$ is the probability that a manifest variable y_i is observed for an individual given each category of the latent variable z_r. As for the latent variable of interest z_a, the latent classes of the latent variable for response propensity z_r are mutually exclusive and exhaustive.

2.3 Structural model

The structural part describes relationships among the categorical latent variables in the model, and possibly relationships between the latent variables and covariates. These are observed variables, other than those used as measures of the latent variables, such as socio-economic characteristics that may affect the latent variables.

Both of the latent variable of interest z_a and the latent variable for response propensity z_r are assumed to be binary, each of them having a Bernoulli distribution. In this case, the structural model will be given by

$$\text{logit}\,\pi_{z_a}(\mathbf{x}) = \alpha_{a0} + \sum_{h=1}^{H} \beta_h x_h , \qquad (3)$$

$$\text{logit}\,\pi_{z_r}(z_a, \mathbf{w}) = \alpha_{r0} + \phi z_a + \sum_{l=1}^{L} \gamma_l w_l , \qquad (4)$$

where $\pi_{z_a}(\mathbf{x})$ is the probability of belonging to the first class of a latent variable of interest z_a given a set of observed covariates x_h, and $\pi_{z_r}(z_a, \mathbf{w})$ is the probability of belonging to the first class of a latent variable for response propensity z_r given the latent variable z_a and a set of observed covariates w_l. The observed covariates affecting z_r may be the same or different from those affecting z_a. If the coefficient ϕ turns out to be significant, this may be taken as evidence of nonrandom missingness. Since then the level/probability of missingness will be associated with certain levels of the latent variable of interest, and hence incorporating a missingness mechanism is inevitable.

3 Bayesian Estimation for Latent Variable Models

A LCM includes estimates for two sets of parameters; parameters involving the probability of membership in each latent class and parameters representing the conditional probabilities of each response (item-response probabilities) given class membership.

The loglikelihood for a random sample of size *n* is given by

$$L = \sum_{m=1}^{n} \log \{f(\mathbf{y}_m, \mathbf{r}_m)\}. \qquad (5)$$

Given the model specification presented by equations $(1), (2), (3)$ and (4), the joint distribution of the observed variables is given by

$$f(\mathbf{y}_m, \mathbf{r}_m) = \sum_{z_a} \sum_{z_r} g(\mathbf{y}_m \,|\, z_a)\; g(\mathbf{r}_m \,|\, z_r)\; h(z_a, z_r \,|\, \mathbf{x}, \mathbf{w}) \quad (6)$$

where $\mathbf{y}_m$ and $\mathbf{r}_m$ represent the $2p$ observed variables for the m^{th} individual.

The conditional distribution of $\mathbf{y}_m \,|\, z_a$ is Bernoulli, given by

$$g(\mathbf{y}_m \,|\, z_a) = \prod_{i=1}^{p} [\,\pi_{ai}(z_a)]^{y_{im}}\; [1 - \pi_{ai}(z_a)]^{1-y_{im}}, \qquad (7)$$

and that of $\mathbf{r}_m \,|\, z_r$ is also Bernoulli, given by

$$g(\mathbf{r}_m \,|\, z_r) = \prod_{i=1}^{p} [\,\pi_{ri}(z_r)]^{r_{im}}\; [1 - \pi_{ri}(z_r)]^{1-r_{im}}. \qquad (8)$$

The joint distribution of z_a and z_r can be written as

$$h(z_a, z_r \,|\, \mathbf{x}, \mathbf{w}) = h(z_r | z_a, \mathbf{w})\; h(z_a | \mathbf{x}), \qquad (9)$$

where both the conditional distribution of the latent variable of interest given covariates $h(z_a|\mathbf{x})$, and that of the response propensity latent variable given the latent variable z_a

and covariates $h(z_r|z_a, \mathbf{w})$ are assumed to be Bernoulli. In the estimation of the model, a given response to a manifest variable is weighted by the probability of responding to this variable, which is a function of response propensity and class membership of the latent variable of interest.

The model presented in this paper will be estimated via a Bayesian approach using MCMC. Inference about unobserved parameters is based on the posterior distribution of the unobserved quantities (including parameters and latent variables) conditional on the observed data. MCMC is used to make draws from this posterior distribution.

Let $\mathbf{v}$ denote a vector with all the unknown quantities including parameters $\boldsymbol{\theta}$ and latent variables; such that $\mathbf{v}' = (\boldsymbol{\theta}, z_a, z_r)$. The loglikelihood given by (5) can be written as

$$\log L(\mathbf{v} \,\big|\, \mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w}) = \sum_{\mathrm{m}=1}^{\mathrm{n}} \log \int \ldots \int g(\mathbf{y_m}, \mathbf{r_m}|\mathbf{v}, \mathbf{x}, \mathbf{w})\, h(\mathbf{v})\, d\mathbf{v}. \qquad (10)$$

The joint posterior distribution of the parameter vector $\mathbf{v}$ is

$$h(\mathbf{v}|\mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w}) = \frac{g(\mathbf{y}, \mathbf{r}|\mathbf{v}, \mathbf{x}, \mathbf{w})\, h(\mathbf{v})}{f(\mathbf{y}, \mathbf{r})} \;\; g(\mathbf{y}, \mathbf{r}|\mathbf{v}, \mathbf{x}, \mathbf{w})\, h(\mathbf{v}). \qquad (11)$$

The above expression has no closed form, and even if it did, we would have to perform multiple integration to obtain the marginal distribution for each coefficient. So, as is usual for Bayesian analysis, we will use the Gibbs sampler.

The main steps of the Bayesian approach for such a latent variable model are as outlined by Bartholomew et al. (2011) and Moustaki and Knott (2005)

1. Inference is based on the posterior distribution $h(\mathbf{v}|\mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w})$, of the unknown parameters $\mathbf{v}$ conditional on the observed data $\mathbf{y}, \mathbf{r}$ and covariates $\mathbf{x}, \mathbf{w}$. Depending on the model fitted, the form of the distribution can be very complex.
2. The mean vector of the posterior distribution $h(\mathbf{v}|\mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w})$ can be used as an estimator of $\mathbf{v}$.
3. Standard deviation of the posterior distribution $h(\mathbf{v}|\mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w})$ can be used to compute standard errors of parameter estimates.
4. In general, the posterior mean $E(\psi(\mathbf{v})|\mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w})$ can be used as a point estimate of a function of the parameter $\psi(\mathbf{v})$, where $E(\psi(\mathbf{v})|\mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w}) = \int \ldots \int \psi(\mathbf{v})\, h(\mathbf{v}|\mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w})\, d\mathbf{v}$.
5. Analytic evaluation of the above expectation is impossible. Alternatives include numerical evaluation, analytic approximations and Monte Carlo Integration.

To avoid the integration required in the posterior expectation, Monte Carlo integration is used in which the integrals are approximated by an average of quantities calculated from sampling. Samples are drawn from the posterior distribution of all the unknown parameters $h\left(\mathbf{v}^{(r)}\middle|\mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w}\right)$, and the expectation over the posterior is approximated by the average over N samples:

$$E\big(h(\mathbf{v}|\mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w})\big) = \frac{1}{N} \sum_{r=1}^{N} h\left(\mathbf{v}^{(r)}\middle|\mathbf{y}, \mathbf{r}, \mathbf{x}, \mathbf{w}\right). \qquad (12)$$

The samples drawn from the posterior distribution do not have to be independent. Samples are drawn from the posterior distribution through a Markov chain with $h(\mathbf{v}|\mathbf{y},\mathbf{r},\mathbf{x},\mathbf{w})$ as its stationary distribution. Algorithms such as the Gibbs sampler and Metropolis-Hastings are used in Bayesian inference. Gibbs sampling is used to produce a sequence of iterations $\mathbf{v}^0,\mathbf{v}^1,\dots,\mathbf{v}^k$ that form a Markov chain, which eventually converges to its stationary distribution, taken to be the posterior distribution. For Bayesian estimation, we use WinBUGS (Bayesian inference Using Gibbs Sampling) (Lunn et al., 2000).

3.1 Choosing prior distributions

The posterior distribution $h(\mathbf{v}|\mathbf{y},\mathbf{r},\mathbf{x},\mathbf{w})$ of the unknown parameters given the data, is obtained by multiplying the likelihood by a prior distribution as shown in equation (11). Thus, a prior distribution needs to be assumed for each parameter of interest of the vector $\mathbf{v}$. We assume vague or noninformative priors to emphasize the likelihood of the data rather than the prior. A normal distribution with mean 0 and a large variance taken to be 10000 is assumed for all parameters of interest defining the outlined model.

3.2 Assessing convergence in MCMC

One of the main issues with MCMC estimation is when to decide that the produced Markov chain has converged to its stationary distribution, which is the posterior distribution of the parameters given the data. Convergence is checked graphically by looking at trace plots showing the full history of estimated values plotted against iteration number for each parameter. A chain is said to have converged when trace plots for parameters depict random patterns that move around the parameter space quickly indicating that the chain is mixing well. It is common practice to run more than one chain simultaneously. In that case, one can be reasonably confident about convergence if all the chains are overlapping one another.

A more formal approach to assess convergence is via convergence diagnostics. These are statistics that have been developed by researchers to facilitate making the decision of convergence. An extensive review of convergence assessment techniques for MCMC is given in Brooks and Roberts (1998). Several convergence diagnostics including those proposed by Raftery and Lewis (1992), Geweke (1992), Heidelberger and Welch (1983), Gelman and Rubin (1992) and Brooks and Gelman (1998) can be produced by CODA (Plummer et al., 2006); an R package that we use for analyzing output obtained from WinBUGS.

4 Application

In this paper, we apply the LCM outlined in Section 2 to study the access of people to knowledge sources using data from the 2014 Egyptian Demographic and Health Survey (EDHS14). The EDHS14 involved two questionnaires: a household questionnaire and an individual questionnaire. The EDHS14 household questionnaire was used to collect information on the socioeconomic status of the households as well as on the nutritional status and anemia levels among women and children. During the main fieldwork and callback phases of the survey, out of $29{,}471$ households selected for the EDHS14, $28{,}630$ households were found. Among those households, $28{,}175$ were successfully interviewed, which represents a response rate of 98.4 percent. Among $28{,}175$ of households who were successfully interviewed, we base our analysis on $27{,}850$ who provided complete answers.

For choosing the items that measure the latent variable of interest, and for determining the suitable number of classes for each categorical latent variable separately, Bayesian estimation is implemented in M*plus* (Muthén and Muthén, 2011). However, M*plus* does not support Bayesian estimation for models with more than one categorical latent variable. WinBUGS (Spiegelhalter et al., 1996) is thus used for estimation of the overall model.

4.1 Selecting items and number of classes for measuring access to knowledge sources

We assume access to knowledge sources to be a latent variable measured by a number of items. For choosing the items, we depend on the Bristol definition of information deprivation. The Bristol indicators were originally developed by a team at the University of Bristol based on the "deprivation approach" to poverty (Gordon et al., 2003) defining children between $2 - 18$ years old who are with no access to radio, television, telephone (land line or mobile phone), computer or newspapers at home as information (knowledge sources) deprived children. We propose a LCM to study the access of people to knowledge sources in Egypt, assuming it to be a latent variable that is explained by a number of items which gives more flexibility to the definition. We apply this definition to the household head since the availability of such devices in a household will facilitate access to all household members.

Collins and Lanza (2010) introduce two criteria that define a strong relation between each observed variable and a latent variable. A distribution of the item-response probabilities for each observed variable across the latent classes is the first criterion. The probability of a response to any observed variable does not depend on the latent variable if that observed variable and the latent variable are independent. An array of the item-response probabilities corresponding to each observed variable that are close to 1 and 0 is the second criterion.

We begin the analysis with eight binary items which are radio, television, land line telephone, mobile phone, computer, video, smart phone and satellite dish. By applying the previous item selection criteria for measuring the latent class variable, four out of eight binary items are selected as measures of the latent variable that we label as “Access to Knowledge Sources”. These four binary items are access to radio, telephone (land line), computer and smart phone. The other four items (television, mobile phone, video and satellite dish) were excluded as they do not contribute much to measuring the latent variable.

Goodman (1974) discusses the identifiability condition of the model ($2^p > (\mathrm{p}+1) \times \mathrm{K}$), where p is the number of binary items and K is the number of classes. According to this condition, the suitable number of classes for the latent variable may be two or three classes. By running the measurement model with two versus three classes for the latent variable of interest, it is found that the most appropriate number of classes in terms of model fit depending on entropy-based criterion (see Celeux and Soromenho, 1996), and interpretability is two classes. A two-class latent variable is thus assumed for the latent variable of interest, in accordance with the model specification outlined in Section 2.

4.2 Fitting LCM for data subject to different levels of missingness

The EDHS14 data have a negligible percentage of missingness. We therefore create artificial missingness within our selected items to illustrate the proposed model. The missingness is created in such a way that makes the probability of an individual not responding to one of the items depend on the value of covariates. The covariates chosen here are wealth index (x_1) and educational level of household head (x_2). An indicator variable r_{mi} is created for each item y_i, that takes value 1 if item y_i is observed for individual m and takes value 0 if it is made to be missing. The probability of a missing response is thus modeled by

$$P(missing) = \text{logit}\,(\alpha_0 + \alpha_1\, x_1 + \alpha_2\, x_2) \tag{13}$$

Four uniform random variables $[0,1]$ corresponding to each item in our study are created. The criterion is, if the $P(missing) > U_i\,[0,1]$, then the corresponding observation will be deleted and treated as missing. It is worth mentioning that both the choice of covariates and values for parameters α_0, α_1 and α_2 in equation (13) are arbitrary. The model is fitted at different levels of missingness: all cases fully observed, $3\% - 52\%$ missingness in each item resulting in $7\% - 80\%$ overall missingness.

Similar to the latent variable of interest, the suitable number of classes for the missingness latent variable “Response Propensity” may be two or three classes according to the identifiability condition. It is found that the most appropriate number of classes in terms of model fit depending on entropy-based criterion and interpretability is two classes, which coincides with our assumption of a two-class latent variable of missingness. This is satisfied at different levels of missingness.

Table 1 summarizes results for a model that analyzes datasets with different levels of missingness. The first three columns show the values for parameters α_0, α_1 and α_2 that are used to create different levels of missingness in the data. The next two columns show the percentage of missingness in each of the four items of the study (radio, telephone, computer and smart phone) and the resulting percentage of overall missingness in the data, respectively. For example, for 3% missingness in each item, a 7% of overall missingness is created in the data. That is 7% of observations have at least one item missing. The last two columns show the estimated regression coefficient ϕ of “Response Propensity” on “Access to Knowledge Sources” and the corresponding p-value, respectively.

Table 1: Effect of “Access to Knowledge Sources” on “Response Propensity” at different levels of missingness, “Access to Knowledge Sources” data

Values for parameters	Percent of missingness in each item	Percent of overall missingness	Estimated parameter ϕ	P-value

α_0	α_1	α_2				
0	-4	1	3%	7%	25.737	999.000
0	-1	0.1	10%	27.7%	4.084	0.000
0	-1	0.5	20%	46.2%	2.717	0.000
0	-1	0.8	32%	62.5%	1.338	0.000
0	-1	1	41%	70.7%	-51.243	999.000
0	-0.8	1	52%	80%	-2118.079	999.000

It is noted that there is an insignificant effect of “Access to Knowledge Sources” on “Response Propensity” at very low and very high levels of overall missingness. However, researchers facing very low percentage of missing values usually exclude those cases, and those facing very high percentages such as 70% and 80% would not usually consider the data as reliable. On the other hand, “Access to Knowledge Sources” has a significant positive effect on “Response Propensity” at moderate levels of overall missingness. Despite the missingness being created at random based on covariates, an individual’s response propensity is still related to their level of access to knowledge sources, depicting nonrandom missingness at most realistic levels of missingness. We will thus focus on analyzing the complete dataset and those with moderate levels of missingness.

Table 2 gives parameter estimates and standard errors for the complete data with covariates, and for the overall model at different levels of missingness estimated using the Bayesian MCMC method. For complete data with covariates, the first 2000 iterations have been discarded as a burn-in period and two chains have been run for 11000 iterations when convergence has been attained. A Multivariate Potential Scale Reduction Factor (MPSRF) is estimated by 1.03, and each univariate Potential Scale Reduction Factor (PSRF) is ≤ 1.06 for each parameter individually, which is taken as an indication of convergence. In case of 27.7% missingness, the first 4000 iterations have been discarded as a burn-in period and two chains have been run for 10000 iterations when convergence has been attained. MPSRF is estimated by 1.11, and each univariate PSRF is ≤ 1.08. In case of 46.2% missingness, the first 6000 iterations have been discarded as a burn-in period and two chains have been run for 10000 iterations when convergence has been attained, MPSRF is estimated by 1.04, and each univariate PSRF is ≤ 1.02. In case of 62.5% missingness, the first 2000 iterations have been discarded as a burn-in period and two chains have been run for 10000 iterations when convergence has been attained. MPSRF is estimated by 1.1, and each univariate PSRF is ≤ 1.09. For all the previous cases, we also looked at trace plots and the Heidelberger and Welch (1983) stationary and interval width tests. All parameters of the model passed that test. The Geweke (1992) criterion showed that all parameters have converged. Convergence diagnostics were obtained from CODA package in R

Table 2: Parameter estimates and standard errors from MCMC for complete data with covariates for measurement model, and for the overall model at different levels of missingness, “Access to Knowledge Sources” data

Item	Parameters	Complete data with covariates	27.7%	46.2%	62.5%
Measurement Model					
Radio	α_{10}	-0.104*** (0.022)	-0.136*** (0.022)	-0.113*** (0.023)	-0.084*** (0.026)
	α_{1a}	-1.143*** (0.030)	-1.123*** (0.032)	-1.119*** (0.034)	-1.164*** (0.037)
Telephone	α_{20}	-0.2191*** (0.023)	-0.280*** (0.023)	-0.257*** (0.024)	-0.248*** (0.027)
	α_{2a}	-2.196*** (0.040)	-2.185*** (0.043)	-2.166*** (0.045)	-2.169*** (0.050)
Computer	α_{30}	1.394*** (0.036)	1.170*** (0.033)	1.219*** (0.035)	1.229*** (0.039)
	α_{3a}	-3.366*** (0.046)	-3.158*** (0.045)	-3.245*** (0.049)	-3.395*** (0.056)
Smart phone	α_{40}	0.2831*** (0.026)	0.163*** (0.025)	0.236*** (0.027)	0.270*** (0.029)
	α_{4a}	-3.144*** (0.046)	-3.079*** (0.050)	-3.214*** (0.055)	-3.314*** (0.061)
Missingness Model					
r (Radio)	v_{10}		4.515*** (0.121)	2.755*** (0.043)	1.758*** (0.026)
	$v_{1\mathrm{r}}$		-3.269*** (0.126)	-2.533*** (0.051)	-2.575*** (0.040)
r (Telephone)	v_{20}		4.642*** (0.130)	2.690*** (0.042)	1.751*** (0.025)
	$v_{2\mathrm{r}}$		-3.375*** (0.134)	-2.455*** (0.050)	-2.573*** (0.040)
r (Computer)	v_{30}		4.414*** (0.120)	2.696*** (0.041)	1.692*** (0.025)
	$v_{3\mathrm{r}}$		-3.190*** (0.124)	-2.518*** (0.049)	-2.517*** (0.040)
r (Smart phone)	v_{40}		4.578*** (0.123)	2.677*** (0.041)	1.722*** (0.025)
	$v_{4\mathrm{r}}$		-3.362*** (0.127)	-2.473*** (0.049)	-2.528*** (0.040)
Structural Model					
z_a on z_r	ϕ		4.084*** (0.233)	2.717*** (0.126)	1.338*** (0.088)

Notes: *** indicates a p-value < 0.01
The MCMC s.d are reported between brackets

Table 3 shows the calculated conditional probabilities for the measurement model “Access to Knowledge Sources” and the missingness model “Response Propensity” assuming a two-class latent variable for each of them. These probabilities are computed from the estimated parameters given in Table 2. The estimated parameters, standard errors, and thus calculated conditional probabilities for each item given class membership seem to be robust with respect to level of missingness. The model accounts for missingness thus retaining the same structure for classes of the latent variable of interest, even at high levels of missingness. The estimated conditional probabilities are consistently higher given membership of the first class, compared to those given membership of the second, for both latent variables. The first latent class of "Access to Knowledge Sources" may thus be labeled as "High access to knowledge sources" and the second latent class as "Low access to knowledge sources". The first latent class of response propensity may be labeled as "High response propensity" and the second latent class as "Low response propensity". The computation time for Bayesian estimation until convergence is attained is approximately 10, 17, 19, and 13 hours for complete data with covariates, data with 27.7% missingness, data with 46.2% missingness, and data with 62.5% missingness, respectively.

Table 3: Item-response conditional probabilities from the MCMC for complete data with covariates for measurement model, and for the overall model at different levels of missingness, "Access to Knowledge Sources" data

	Complete data with covariates		27.7% missingness		46.2% Missingness		62.5% missingness	
	1st class	2nd class	1st class	2nd class	1st class	2nd class	1st class	2nd class
"Access to Knowledge Sources"	Probability of a "Yes"							
Radio	0.474	0.223	0.466	0.221	0.472	0.226	0.479	0.223
Telephone	0.445	0.082	0.431	0.078	0.436	0.081	0.438	0.082
Computer	0.802	0.122	0.763	0.120	0.772	0.117	0.774	0.103
Smart phone	0.570	0.054	0.541	0.051	0.559	0.048	0.567	0.045
"Response Propensity"	Probability of a "Not missing"							
$r_{(Radio)}$			0.989	0.777	0.940	0.555	0.853	0.306
$r_{(Telephone)}$			0.990	0.780	0.936	0.558	0.852	0.305
$r_{(Computer)}$			0.988	0.773	0.937	0.544	0.844	0.305
$r_{(Smart\ phone)}$			0.990	0.771	0.936	0.551	0.848	0.309

Notes: The probability of a "No" response can be calculated by subtracting the item-response probabilities shown above from 1.
The probability of a "Missing" response can be calculated by subtracting the item-response probabilities shown above from 1.

Covariates affecting the latent variable "Access to Knowledge Sources" are wealth index, educational level of household head, sex of household head (male/ female), age in years of household head, and place of residence (urban/ rural). Covariates affecting the missingness latent variable "Response propensity" are sex of household head (male/ female), age in years of household head, and place of residence (urban/ rural). We do not study the effect of wealth index and educational level of household head on "Response Propensity" as they are used in creating the missingness.

From Table 4, it is noted that wealth index, educational level of household head, age of household head, and place of residence have significant negative effects on people's "Access to Knowledge", while sex has a significant positive effect. Considering the definition of the latent variable and its two classes, this indicates that the probability of having high access to knowledge sources is higher for older, richer males with higher levels of education for those living in rural areas. These all seem to be expected results except for the area of residence covariate where people living in rural areas are usually expected to have lower access to knowledge sources. One possible explanation of this unexpected result may be that the devices (radio, telephone, computer and smartphone) are available in rural areas, which facilitates access to knowledge sources, but no information is available on whether these devices are used as sources for knowledge or mainly for entertainment and communication. The same effects for covariates are depicted with the complete data and at different levels of missingness, indicating robustness of the model.

Table 4: Parameter estimates and standard errors from the MCMC of the covariates effects for complete data with covariates for measurement model, and for the overall model at different levels of missingness, "Access to Knowledge Sources" data

Item	Parameters	Complete data with covariates	27.7% missingness	46.2% missingness	62.5% missingness
Covariates effects on z_a					
Intercept	α_{a0}	18.48*** (0.484)	21.660*** (0.583)	20.970*** (0.608)	19.950*** (0.522)
Place of residence (Rural)	β_1	-2.806*** (0.11)	-3.595*** (0.132)	-3.394*** (0.138)	-3.167*** (0.124)
Wealth index	β_2	-2.702*** (0.072)	-3.246*** (0.081)	-3.356*** (0.085)	-3.293*** (0.076)
Sex (Female)	β_3	0.383*** (0.099)	0.394*** (0.104)	0.529*** (0.106)	0.552*** (0.111)
Age	β_4	-0.043*** (0.003)	-0.047*** (0.003)	-0.035*** (0.003)	-0.028*** (0.003)
Educational level	β_5	-0.629*** (0.020)	-0.614*** (0.023)	-0.492*** (0.024)	-0.424*** (0.024)
Covariates effects on z_r					
Intercept	α_{r0}		-8.150***	-4.575***	-1.570

		(0.308)	(0.174)	(0.138)
Place of residence (Rural)	γ_1	2.559*** (0.078)	2.452*** (0.059)	2.010*** (0.049)
Sex (Female)	γ_2	-0.476*** (0.092)	-0.995*** (0.081)	-1.472*** (0.086)
Age	γ_3	0.018*** (0.002)	-0.026*** (0.002)	-0.044*** (0.002)

Notes: *** indicates a p-value < 0.01
The MCMC s.d are reported between brackets

The negative coefficients of age and sex of household head on “Response Propensity” indicate that older people and females have higher probability of responding. The positive coefficient of place of residence on “Response Propensity” indicates that people living in urban areas have higher propensity to respond.

Given the definition of the latent variables and interpretation of their classes, the significant positive effect ϕ of “Access to Knowledge Sources” on “Response Propensity” reported at the end of Table 2, at all levels of missingness under consideration, indicates that the probability of having high response propensity increases with high access to knowledge sources, which is taken as evidence of nonrandom missingness since higher levels of response are associated with higher levels of access to knowledge sources even after controlling for covariates.

5 Conclusion

In this paper, we have studied a LCM with two categorical latent variables; one for the phenomenon of interest, and the other for response propensity. Bayesian estimation has been adopted to fit the proposed model. Non informative priors have been assumed for all model parameters. We have used some of the diagnostics available in CODA to check the convergence of our models.

The model has been applied to data from Egypt’s Demographic and Health Survey 2014. In the application, artificial missingness has been created to study the model under different levels of missingness. Parameter estimates obtained for the model were very close at different levels of missingness. One of the main findings of the model was that even with high levels of missingness, the proposed model retains the structure for the latent classes as for the complete data. It also succeeds to capture the same covariates effects at high levels of missingness. Another important result is that even after controlling for covariates, the probabilities of belonging to classes of the “Response Propensity” latent variable still depend on classes of the “Access to Knowledge Sources” latent variable making the missingness nonignorable. Higher levels of response were found to be associated with higher levels of “Access to Knowledge Sources” which may be due to higher levels of awareness. This result confirms the importance of accommodating the missingness mechanism within the modeling of the data.

Acknowledgement

The authors are very grateful to the Editor for helpful comments and to a referee

for important comments and suggestions on the manuscript.

References

1. Asparouhov, T., and Muthén, B. O. (2011). Using Bayesian priors for more flexible latent class analysis. In Proceedings of the 2011 Joint Statistical Meeting, Miami, FL. Retrieved from http://www.statmodel.com/download/BayesLCA.pdf
2. Bacci, S. and Bartolucci, F. (2014). A Multidimensional Latent Class IRT Model for Non-Ignorable Missing Responses. Structural Equation Modeling: A Multidisciplinary Journal, https://arxiv.org/abs/1410.4856
3. Bartholomew, D. J., Knott, M. and Moustaki, I. (2011). Latent Variable Models and Factor Analysis, Third edition. Wiley series in probability and statistics.
4. Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics, 7 (4), 434-455.
5. Brooks, S. P. and Roberts, G. O. (1998). Convergence assessment techniques for Markov Chain Monte Carlo. Statistics and Computing, 8, 319-335.
6. Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification, 13(2), 195–212.
7. Collins, L. M. and Lanza, S. T. (2010). Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences, Wiley series in probability and statistics.
8. De Menezes, L. M. (1999). On fitting latent class models for binary data: the estimation of standard errors. British Journal of Mathematical and Statistical Psychology, 52, 149-168.
9. 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, 39, 1-38.
10. Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical science, 7 (4), 457-472.
11. Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Bayesian Statistics, 4, 169-193.
12. Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 61, 215-231.
13. Gordon, D., Nandy, S., Pantazis, C., Pemberton, S., and Townsend, P. (2003). The distribution of poverty in the developing world. Report to UNICEF, University of Bristol, UK, Centre for International Poverty Research.
14. Galindo-Garre, F., and Vermunt, J. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. Behaviormetrika, 33(1), 43–59.
15. Hafez, M. S., Moustaki, I. and Kuha, J. (2015). Analysis of Multivariate Longitudinal Data Subject to Nonrandom Dropout. Structural Equation Modeling: A Multidisciplinary Journal, 22(2), 193-201.
16. Hafez, M. S. (2015). Analysis of Multivariate Longitudinal Categorical Data Subject to Nonrandom Missingness: A Latent Variable Approach. A thesis submitted to the

Department of Statistics of the London School of Economics for the degree of Doctor of Philosophy.

17. Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. Operations Research, 31 (6), 1109-1144.

18. Kuha, J., Katsikatsou, M., and Moustaki, I. (2018). Latent variable modelling with non-ignorable item nonresponse: multigroup response propensity models for cross-national analysis. Journal of the Royal Statistical Society, Series A, 181, in the press.

19. Lunn, D., A. Thomas, N. Best, and Spiegelhalter, D. (2000). Winbugs - a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10, 325-337.

20. Ministry of Health and Population [Egypt], El-Zanaty and Associates [Egypt], and ICF International (2014). Egypt demographic and health survey 2014. [Data file and code book]. Retrieved from https://www.dhsprogram.com/data/available-datasets.cfm

21. Moustaki, I. and Knott, M. (2000). Weighting for item nonresponse in attitude scales by using latent variable models with covariates. Journal of the Royal Statistical Society, Series A, 163(3), 445-459.

22. Moustaki, I. and Knott, M. (2005). Computational aspects of the EM and Bayesian estimation in Latent Variable Models. Chapter New Developments in Categorical Data Analysis for the Social and Behavioral Sciences, Psychology Press,103-124.

23. Muthén, L. K., and Muthén, B. O. (1998–2011). Mplus user's guide (6th ed.), Los Angeles, CA: Muthén & Muthén.

24. O'Muircheartaigh, C. and Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item nonresponse in attitude scale. Journal of the Royal Statistical Society, Series A, 162(2), 177-194.

25. Pan, J. C. and Huang, G. H. (2013). Bayesian inferences of latent class models with an unknown number of classes. Psychometrika, 1–26.

26. Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for MCMC. R News 6 (1), 7–11.

27. Raftery, A. L. and Lewis, S. (1992). How many iterations in the Gibbs Sampler? Bayesian Statistics, 4 (2), 763–773.

28. Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). Bugs: Bayesian inference using gibbs sampling (Tech. Rep.). MRC Biostatistics Unit, Institute of Public Health, Cambridge.

29. Thanoon, T. Y. and Adnan, R. (2016). Bayesian Analysis of Linear and Nonlinear Latent Variable Models with Fixed Covariate and Ordered Categorical Data. Pakistan Journal of Statistics and Operation Research, 12(1), 125–140.

30. White, A., Wyse, J. and Murphy, T. B. (2016). Bayesian variable selection for latent class analysis using a collapsed Gibbs sampler. Stat. Comput. 26(1–2), 511–527.