American Journal of Theoretical and Applied Statistics
2026; 15(5): 231-242
10.11648/j.ajtas
doi: 10.11648/j.ajtas.20261505.14

Bayesian Latent Class Models for Multiple Imputation of Multivariate Categorical Data Under Ignorable Missingness Mechanism

Daniel Fundi Murithi*, **Josephine Njeri Ngure**, **Victor Muthama Musau**

Department Pure and Applied Sciences, Kirinyaga University, Kerugoya, Kenya

Email address:
fundidaniel55@gmail.com (Daniel Fundi Murithi), jngure@kyu.ac.ke (Josephine Njeri Ngure), vmusau@kyu.ac.ke (Victor Muthama Musau)

To cite this article:
Daniel Fundi Murithi, Josephine Njeri Ngure, Victor Muthama Musau. (2026). Bayesian Latent Class Models for Multiple Imputation of Multivariate Categorical Data Under Ignorable Missingness Mechanism. *American Journal of Theoretical and Applied Statistics, 15*(5), 231-242. https://doi.org/10.11648/j.ajtas.20261505.14

Received: 10 August 2026; **Accepted:** 20 August 2026 **Published:** 20 September 2026

Abstract: Missing data are a pervasive challenge in the analysis of multivariate categorical data and can lead to biased estimation and reduced statistical efficiency when handled inappropriately. Bayesian latent class multiple imputation provides a flexible framework for addressing this problem by exploiting the underlying latent structure of the data. However, conventional Bayesian latent class imputation models generally assume that latent class membership is independent of observed covariates, potentially limiting estimation accuracy when covariates are informative of the latent class structure. This study proposes a covariate-dependent Bayesian latent class model for multiple imputation under the MAR mechanism. The proposed approach models latent class membership using multinomial logistic regression, allowing observed covariates to influence class allocation while preserving the latent class formulation for the response variables. Bayesian inference is performed using Gibbs sampling, with Pólya–Gamma data augmentation facilitating efficient posterior estimation of the regression coefficients. The performance of the proposed model was evaluated through a comprehensive simulation study under varying strengths of covariate effects and sample sizes, and compared with the conventional Bayesian latent class model. Performance was assessed using multiple imputation inference, latent class proportion recovery, regression coefficient recovery, posterior stability, and empirical coverage probabilities. The proposed model produced unbiased inference and demonstrated improved estimation accuracy when covariates were associated with latent class membership, while maintaining performance comparable to the conventional model when covariate effects were absent. Furthermore, RMSE decreased consistently as the sample size increased, providing empirical evidence of posterior concentration and improved estimation precision. These findings demonstrate that incorporating covariates into Bayesian latent class multiple imputation provides a practical and effective extension of the conventional latent class framework for multivariate categorical data with missing values under the MAR mechanism.

Keywords: Multiple Imputation, Latent Class Model, Missing at Random, Multivariate Categorical Data, Multinomial Logistic Regression, Bayesian Latent Class model, Covariate-dependent Latent Class Model

1. Introduction

Missing data are a common challenge in demographic, health, social science, and other survey-based studies, where many variables of interest are categorical and often exhibit complex dependence structures. Missing responses may arise from respondent refusal, nonresponse to specific survey items, or other data collection challenges, resulting in incomplete observations that can compromise statistical analysis if not appropriately addressed [1]. In particular, item nonresponse is common in survey research and may lead to biased parameter estimates, reduced statistical efficiency, and unreliable statistical inference when observations with missing values are ignored or handled using inappropriate methods.

Several approaches have been developed to address missing data, including complete-case analysis, likelihood-based methods, weighting techniques, and imputation procedures. Among the traditional methods, listwise deletion remains one of the simplest and most widely used approaches because of its ease of implementation. However, it discards all incomplete observations and may result in substantial information loss and biased inference when the proportion of missing data is non-negligible or when respondents with missing values differ systematically from complete respondents [2]. These limitations have motivated the development of more principled approaches that preserve sample size while appropriately accounting for uncertainty associated with missing observations.

MI has emerged as one of the most flexible and statistically rigorous frameworks for handling incomplete data [3–6]. Rather than replacing each missing value with a single estimate, MI generates multiple plausible values drawn from an appropriate predictive distribution, producing several completed datasets that are analysed separately and subsequently combined using standard pooling rules. By incorporating both within- and between-imputation variability,MI yields valid statistical inference while explicitly accounting for uncertainty arising from the imputation process [3, 7]. Owing to this flexibility, MI has become a preferred approach for analysing multivariate datasets, particularly when different substantive models are of interest.

The effectiveness of multiple imputation depends critically on the specification of the imputation model. For multivariate categorical data, conventional joint modelling approaches such as log-linear models become computationally impractical as the number of variables increases because of the rapid growth in higher-order interaction terms [7, 8]. Fully conditional specification methods, including MICE, provide greater modelling flexibility but require the specification of numerous conditional models and may fail to adequately preserve complex multivariate dependence structures in high-dimensional categorical data [4, 9]. These limitations have created interest in alternative modelling strategies capable of representing complex categorical relationships more efficiently.

BLCM provide a flexible framework for modelling multivariate categorical data by representing the joint distribution as a finite mixture of latent classes [10, 11]. Under the local independence assumption, dependence among observed variables is explained through an unobserved latent class variable, allowing complex joint distributions to be approximated without explicitly modelling numerous higher-order interactions [12, 13]. Within the MI framework, BLCM generate imputations from posterior predictive distributions, thereby preserving important multivariate relationships while naturally incorporating parameter uncertainty [3, 14, 15]. Consequently, latent class-based MI has demonstrated favourable performance for multivariate categorical data, particularly in terms of bias reduction, estimation efficiency, and preservation of complex dependence structures [10, 16].

Despite these advantages, existing Bayesian latent class multiple imputation methods remain limited in their ability to incorporate auxiliary covariate information into the latent class structure. Most existing approaches assume constant latent class proportions across individuals, implying that class membership is independent of observed covariates [11]. In many practical applications, however, demographic and socioeconomic characteristics explain substantial population heterogeneity and therefore contain valuable information for predicting latent class membership. Ignoring such information may reduce the flexibility of the latent class representation and consequently limit the quality of imputations and subsequent statistical inference.

This limitation reveals an important methodological gap in the current literature. While Bayesian latent class multiple imputation provides a powerful framework for analysing incomplete multivariate categorical data under the MAR assumption, comparatively little attention has been devoted to modelling latent class membership as a function of observed covariates within a unified Bayesian multiple imputation framework. Incorporating covariate information into the latent class allocation has the potential to produce a more informative representation of population heterogeneity while improving the efficiency and accuracy of multiple imputation.

This paper addresses this gap by developing a Bayesian latent class multiple imputation model for multivariate categorical data under the MAR assumption that incorporates covariate-dependent latent class membership through a multinomial logistic regression model. By allowing latent class probabilities to vary across individuals according to observed covariates, the proposed model provides a more flexible representation of the underlying population structure while retaining the advantages of Bayesian latent class multiple imputation. The performance of the proposed approach is evaluated through comprehensive simulation studies and compared with the conventional Bayesian latent class multiple imputation model.

2. Bayesian Latent Class Multiple Imputation Model

2.1. Latent Class Model

LCM provide a flexible framework for modelling multivariate categorical data by representing the dependence among observed variables through an unobserved categorical latent variable. Rather than modelling the potentially large number of interactions among observed variables directly, LCM partition the population into a finite number of homogeneous latent classes, within which the observed responses are assumed to be conditionally independent. This efficient representation enables complex dependence structures to be captured while maintaining computational tractability, making latent class models particularly suitable for high-dimensional categorical data. Consequently, LCM have been widely applied in diverse fields, including health research for disease and symptom classification and marketing for consumer segmentation [17–21].

Consider a sample of n independent individuals measured on p categorical response variables. Let

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip}), \qquad i = 1, \ldots, n,$$

denote the response vector for the i-th individual, and let $L_i \in \{1, \ldots, K\}$ denote the corresponding latent class membership, where K is the total number of latent classes. The probability of membership in latent class k is denoted by

$$\pi_k = P(L_i = k), \qquad k = 1, \ldots, K,$$

where

$$\pi_k \geq 0, \qquad \sum_{k=1}^{K} \pi_k = 1.$$

The LCM represents the observed response distribution as a finite mixture of latent class-specific distributions. Accordingly, the marginal probability of observing the response pattern $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})$ is given by

$$P(\mathbf{Y}_i = \mathbf{y}_i) = \sum_{k=1}^{K} P(\mathbf{Y}_i = \mathbf{y}_i \mid L_i = k)\, P(L_i = k). \quad (1)$$

A fundamental assumption of the LCM is *local independence*, whereby the observed response variables are assumed to be conditionally independent given latent class membership [17]. Consequently,

$$P(\mathbf{Y}_i = \mathbf{y}_i \mid L_i = k) = \prod_{j=1}^{p} P(Y_{ij} = y_{ij} \mid L_i = k). \quad (2)$$

Substituting Equation (2) into Equation (1) yields

$$P(\mathbf{Y}_i = \mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} P(Y_{ij} = y_{ij} \mid L_i = k). \quad (3)$$

Equation (3) shows that the latent class model is a finite mixture model in which the overall response distribution is expressed as a weighted average of class-specific multinomial distributions, with the mixing proportions determined by the latent class probabilities. Under the local independence assumption, the dependence among the observed variables is explained entirely through the latent class structure. Although this assumption may not hold exactly in every application, it provides a flexible and computationally efficient representation of multivariate categorical data that forms the basis for the covariate-dependent extension developed in the following section.

2.2. Covariate-Dependent Latent Class Membership

The LCM presented in Section 2.1 assumes that the class membership probabilities are identical for all individuals. In many applications, however, the probability of belonging to a latent class is influenced by observed individual characteristics such as demographic, socioeconomic, or behavioural variables. Incorporating covariates into the LCM allows the class membership probabilities to vary across individuals while preserving the latent class structure. This extension improves the model's ability to explain population heterogeneity and has been widely adopted in latent class regression and concomitant-variable models [22–26].

Let

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})^\top$$

denote the vector of observed covariates for individual i. Conditional on the covariates, the distribution of the response vector is given by

$$P(\mathbf{Y}_i = \mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} P(\mathbf{Y}_i = \mathbf{y}_i \mid L_i = k, \mathbf{x}_i)\, P(L_i = k \mid \mathbf{x}_i). \quad (4)$$

Consistent with the local independence assumption introduced in Section 2.1, the covariates influence the observed responses only through the latent class membership. Consequently,

$$P(\mathbf{Y}_i = \mathbf{y}_i \mid L_i = k, \mathbf{x}_i) = P(\mathbf{Y}_i = \mathbf{y}_i \mid L_i = k),$$

so that Equation (4) becomes

$$P(\mathbf{Y}_i = \mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_k(\mathbf{x}_i) \prod_{j=1}^{p} P(Y_{ij} = y_{ij} \mid L_i = k), \quad (5)$$

where

$$\pi_k(\mathbf{x}_i) = P(L_i = k \mid \mathbf{x}_i),$$

with

$$\pi_k(\mathbf{x}_i) \geq 0, \qquad \sum_{k=1}^{K} \pi_k(\mathbf{x}_i) = 1.$$

Following covariate-dependent latent class framework, the class membership probabilities are modelled using a baseline-category multinomial logistic regression [26–28]. Specifically,

$$\pi_k(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_k)}{\sum_{r=1}^{K} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_r)}, \qquad k = 1, \ldots, K, \quad (6)$$

where $\boldsymbol{\beta}_k$ denotes the regression coefficient vector associated with latent class k. For identifiability, one class is set as the baseline by fixing

$$\boldsymbol{\beta}_1 = \mathbf{0}.$$

This parameterization guarantees that the class membership probabilities remain positive and sum to one for every covariate pattern while preserving the multinomial logistic representation of the latent class probabilities [29]. Consequently, the covariates affect the joint distribution of the observed responses indirectly through their influence on latent class membership, whereas the class-specific response probabilities remain unchanged [26]. The resulting covariate-dependent LCM provides the likelihood formulation upon which the Bayesian estimation procedure presented in the following sections is developed.

2.3. Bayesian Model Specification

The covariate-dependent latent class model is formulated within a Bayesian framework by combining the observed-data likelihood with prior distributions for the unknown model parameters. Bayesian inference provides a coherent probabilistic framework for incorporating prior information while accounting for parameter uncertainty through posterior distributions [30, 31]. Let

$$\Theta = (\boldsymbol{\beta}, \mathbf{\Pi})$$

denote the collection of unknown model parameters, where $\boldsymbol{\beta}$ represents the regression coefficients governing covariate-dependent class membership and $\mathbf{\Pi}$ denotes the latent class-specific item-response probabilities.

2.3.1. Observed-Data Likelihood

Suppose that for a subject i, the response vector

$$\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip})$$

contains both observed and missing components. Under the MAR assumption, inference is based on the observed responses by integrating over the missing values rather than modelling the missing responses directly [6, 32]. Let $\mathbf{Y}_i^{obs}$ denote the observed portion of the response vector and $\mathbf{x}_i$ denote associated covariate vector.

Conditioning on the latent class membership, the observed-data likelihood for subject i is obtained by marginalizing over the latent classes,

$$P(\mathbf{Y}_i^{obs} \mid \mathbf{x}_i, \Theta) = \sum_{k=1}^{K} \pi_k(\mathbf{x}_i) \prod_{j \in O_i} P(Y_{ij} \mid L_i = k),$$

where O_i denotes the set of observed response variables for subject i, K is the number of latent classes, $\pi_k(\mathbf{x}_i)$ is the covariate-dependent probability of belonging to latent class k, and $P(Y_{ij} \mid L_i = k)$ is the latent class-specific response probability for variable j.

Assuming independence across subjects, the observed-data likelihood for the complete sample is given by,

$$L(\Theta) = \prod_{i=1}^{n} P(\mathbf{Y}_i^{obs} \mid \mathbf{x}_i, \Theta).$$

This observed-data likelihood forms the basis for Bayesian estimation by utilizing only the observed responses while appropriately accounting for uncertainty arising from missing data through marginalization [6].

2.3.2. Prior Distributions

Prior distributions are assigned to all unknown model parameters to complete the Bayesian specification of the proposed model. Bayesian prior distributions provide a principled mechanism for incorporating existing information while regularizing estimation and quantifying uncertainty in posterior inference [30, 31].

For the latent class-specific item-response probabilities,

$$\boldsymbol{\pi}_{kj} = (\pi_{kj1}, \ldots, \pi_{kjC_j}),$$

where C_j denotes the number of categories for response variable j, independent Dirichlet priors are assigned,

$$\boldsymbol{\pi}_{kj} \sim \text{Dirichlet}(\alpha_{kj1}, \ldots, \alpha_{kjC_j}),$$

where $\alpha_{kjc} > 0$ are fixed hyperparameters reflecting prior beliefs regarding the response probabilities within each latent class.

The covariate effects governing latent class membership are represented by the multinomial logistic regression coefficients,

$$\boldsymbol{\beta}_k, \qquad k = 2, \ldots, K,$$

with the first latent class serving as the baseline for identifiability. Independent multivariate normal priors are assigned,

$$\boldsymbol{\beta}_k \sim N(\mathbf{0}, \Sigma_\beta),$$

where Σ_β is a prespecified covariance matrix controlling the degree of prior uncertainty. Weakly informative normal priors are adopted to allow the observed data to dominate posterior inference while providing regularization of the regression coefficients.

Since the latent class probabilities $\pi_k(\mathbf{x}_i)$ are determined through the multinomial logistic regression model introduced in Section 2.2, no separate prior distribution is specified for these probabilities. Instead, their posterior distributions are induced through the prior assigned to the regression coefficients.

Combining the observed-data likelihood with the prior distributions through Bayes' theorem yields the joint posterior distribution [30],

$$P(\Theta \mid \mathbf{Y}^{obs}, \mathbf{X}) \propto L(\Theta) P(\Theta),$$

where

$$P(\Theta) = P(\boldsymbol{\beta}) P(\mathbf{\Pi}).$$

The resulting posterior distribution forms the basis for Bayesian estimation and posterior predictive multiple imputation described in the following sections.

2.4. Posterior Estimation

Posterior inference for the proposed covariate-dependent Bayesian latent class model is based on Bayes' theorem, whereby information from the observed data is combined with the prior distributions to obtain the posterior distribution of the unknown parameters [30, 31]. Specifically,

$$P(\Theta \mid \mathbf{Y}^{obs}, \mathbf{X}) = \frac{L(\Theta)P(\Theta)}{P(\mathbf{Y}^{obs} \mid \mathbf{X})},$$

where $L(\Theta)$ denotes the observed-data likelihood, $P(\Theta)$ represents the joint prior distribution of the unknown model parameters, and $P(\mathbf{Y}^{obs} \mid \mathbf{X})$ is the marginal likelihood. Since the marginal likelihood involves high-dimensional integration over the parameter space, direct analytical evaluation of the posterior distribution is generally intractable.

Posterior inference is therefore performed using MCMC methods, specifically a Gibbs sampling algorithm that iteratively samples from the full conditional distributions of the unknown quantities [33]. At each iteration, the latent class memberships, latent class-specific response probabilities, and regression coefficients are sequentially updated using the most recent values of the remaining parameters. Repeated sampling from these conditional distributions generates a Markov chain whose stationary distribution converges to the joint posterior distribution of the model parameters.

Following an initial burn-in period, posterior samples are retained to approximate the posterior distribution. Posterior summaries, including posterior means, posterior standard deviations and credible intervals, are subsequently computed from the retained samples for statistical inference [30].

The multinomial logistic regression governing the covariate-dependent latent class probabilities does not admit conjugate full conditional distributions for the regression coefficients. Consequently, posterior estimation is performed using the Pólya–Gamma data augmentation algorithm, which enables efficient Gibbs sampling while preserving the Bayesian formulation of the proposed model [34].

2.5. Multiple Imputation Procedure

Multiple imputation is performed within the Bayesian framework by generating plausible values for the missing responses from their posterior predictive distribution. This approach propagates uncertainty associated with both the missing data and the model parameters into subsequent statistical inference [3, 35]. Let $\mathbf{Y}^{obs}$ and $\mathbf{Y}^{mis}$ denote the observed and missing components of the response matrix, respectively. The posterior predictive distribution of the missing responses is

$$P(\mathbf{Y}^{mis} \mid \mathbf{Y}^{obs}, \mathbf{X}) = \int P(\mathbf{Y}^{mis} \mid \Theta, \mathbf{Y}^{obs}, \mathbf{X}) \times P(\Theta \mid \mathbf{Y}^{obs}, \mathbf{X})\, d\Theta. \quad (7)$$

Since the posterior predictive distribution generally does not admit a closed-form expression, imputations are generated using posterior samples obtained from the Gibbs sampler [30]. For each retained posterior draw $\Theta^{(t)}$, missing responses are sampled from

$$\mathbf{Y}^{mis(t)} \sim P(\mathbf{Y}^{mis} \mid \Theta^{(t)}, \mathbf{Y}^{obs}, \mathbf{X}),$$

thereby producing one completed dataset. Repeating this procedure for m retained posterior draws generates m multiply imputed datasets.

Parameter estimation is subsequently performed on each completed dataset. Let $\hat{Q}_l$ and U_l denote the estimate and associated variance obtained from the l-th completed dataset. The pooled estimate is given by,

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_l,$$

with within-imputation variance

$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m} U_l,$$

between-imputation variance

$$B = \frac{1}{m-1}\sum_{l=1}^{m} (\hat{Q}_l - \bar{Q})^2,$$

and total variance given by Rubin's combining rules [3],

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B.$$

The pooled estimates and their corresponding variances incorporate uncertainty arising from both missing data and parameter estimation, thereby providing statistically valid inference under the MAR assumption [3, 6]. The resulting multiply imputed datasets are subsequently used for parameter estimation, interval estimation and model evaluation in the simulation studies and real-data application.

3. Simulation Study

3.1. Simulation Design

A simulation study was conducted to evaluate the finite-sample performance of the proposed covariate-dependent Bayesian latent class multiple imputation model under MAR mechanism. The proposed model, denoted by $\mathrm{BLCM}_{\mathrm{MAR}}$, was compared with the conventional Bayesian latent class model without covariates, denoted by BLCM_0. Complete multivariate categorical datasets were generated from a LCM, after which missing values were introduced under a MAR mechanism.

The simulation design considered three scenarios representing increasing levels of covariate influence on latent class membership: no covariate effects (Scenario A), moderate covariate effects (Scenario B), and strong covariate effects (Scenario C). Each scenario was evaluated across two sample sizes and two missingness levels. For every simulation condition, 500 datasets were generated and analysed using BLCM_{MAR} and BLCM_0. Posterior estimation was performed using a Gibbs sampling algorithm with 1,500 iterations, and model performance was assessed using bias, RMSE, posterior stability, empirical coverage probabilities, and latent class occupancy. The simulation scenarios and implementation settings are summarized in Tables 1 and 2.

Table 1. *Simulation scenarios.*

Scenario	Covariate effects on latent class membership
A	None
B	Moderate
C	Strong

Table 2. *Simulation settings.*

Component	Specification
Latent classes (K)	3
Response variables	6 categorical variables
Response categories	3 per variable
Covariates	$X_1, \ldots, X_6$
Sample sizes	500, 1000
Variables with missing values	$Y_2,\ Y_3$
Missingness mechanism	MAR
Missingness rates	10%, 20%
Simulation replications	500
Posterior iterations	1,500
Multiply imputed datasets	100
Models compared	BLCM_0, BLCM_{MAR}
Software	R version 4.5.2

3.2. Data-Generating Process

The simulation data were generated to reflect the structure of the proposed covariate-dependent Bayesian latent class model. For each simulation replication, covariates were first generated for all individuals, followed by latent class membership through a multinomial logistic regression model. Conditional on the latent class, multivariate categorical responses were generated under the local independence assumption. Finally, missing values were introduced according to MAR mechanism that depended only on observed covariates and responses.

3.2.1. Covariate Generation

For each individual $i = 1, \ldots, n$, six trichotomous covariates, $X_1, \ldots, X_6$, were independently generated from a categorical distribution with equal category probabilities,

$$X_{ij} \sim \text{Categorical}\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), \qquad j = 1, \ldots, 6.$$

The resulting covariate vector for individual i is denoted by

$$\mathbf{x}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})^\top.$$

3.2.2. Latent Class Generation

Given the covariate vector $\mathbf{x}_i$, latent class membership was generated from a multinomial logistic regression model with the first latent class taken as the reference category,

$$\log \frac{\Pr(L_i = k \mid \mathbf{x}_i)}{\Pr(L_i = 1 \mid \mathbf{x}_i)} = \beta_{k0} + \mathbf{x}_i^\top \boldsymbol{\beta}_k, \qquad k = 2, \ldots, K. \quad (8)$$

The corresponding latent class probabilities were obtained through the softmax transformation,

$$\pi_k(\mathbf{x}_i) = \frac{\exp\left(\beta_{k0} + \mathbf{x}_i^\top \boldsymbol{\beta}_k\right)}{\sum_{r=1}^{K} \exp\left(\beta_{r0} + \mathbf{x}_i^\top \boldsymbol{\beta}_r\right)}, \qquad k = 1, \ldots, K. \quad (9)$$

To investigate the influence of covariates on latent class recovery and multiple imputation performance, three simulation scenarios were considered. Scenario A assumed no covariate effects on latent class membership, Scenario B represented moderate covariate effects, while Scenario C represented strong covariate effects obtained by increasing the magnitude of the regression coefficients while preserving their directional pattern. The regression coefficients used to generate the latent class memberships are presented in Table 3.

Table 3. *Regression coefficients used for generating latent class membership.*

Scenario	Class 2 coefficients	Class 3 coefficients
A	(0.30,0,0,0,0,0,0)	(0.00,0,0,0,0,0,0)
B	(0.30,0.3,0.2,0.3,0.2,0.3,0.2)	(0.00,-0.3,0.3,-0.2,0.2,0.3,-0.3)
C	(0.30,0.6,0.4,0.5,0.4,0.5,0.4)	(0.00,-0.6,0.6,-0.4,0.4,0.5,-0.5)

3.2.3. Response Generation

Conditional on latent class membership, six trichotomous response variables were generated independently according to the local independence assumption,

$$P(\mathbf{Y}_i = \mathbf{y}_i \mid L_i = k) = \prod_{j=1}^{6} \prod_{s=1}^{3} (\pi_{kjs})^{I_{ijs}}, \quad (10)$$

where

$$I_{ijs} = \mathbf{1}(Y_{ij} = s).$$

The class-specific response probabilities used in the simulations are given in Table 4.

Table 4. *Class-specific response probabilities used in the simulation study.*

Item	Class 1	Class 2	Class 3
Y_1	$(.65, .25, .10)$	$(.25, .55, .20)$	$(.15, .25, .60)$
Y_2	$(.62, .28, .10)$	$(.28, .52, .20)$	$(.18, .25, .57)$
Y_3	$(.60, .30, .10)$	$(.25, .55, .20)$	$(.15, .30, .55)$
Y_4	$(.64, .26, .10)$	$(.26, .54, .20)$	$(.16, .30, .54)$
Y_5	$(.66, .24, .10)$	$(.28, .52, .20)$	$(.18, .30, .52)$
Y_6	$(.63, .27, .10)$	$(.27, .53, .20)$	$(.17, .30, .53)$

3.2.4. Missing-at-Random Mechanism

After generating the complete datasets, missing values were imposed on response variables Y_2 and Y_3 according to a missing-at-random mechanism. Let $R_{ij} = 1$ if Y_{ij} was observed and $R_{ij} = 0$ otherwise. Observation probabilities were generated using logistic regression models that depended exclusively on observed covariates and observed responses,

$$\begin{aligned}\operatorname{logit}\{\Pr(R_{i2}=1)\} = \gamma_{20} &+ \gamma_{21}X_{i1}^{c} + \gamma_{22}X_{i3}^{c} \\ &+ \gamma_{23}\mathbf{1}(Y_{i1}=1) \\ &+ \gamma_{24}\mathbf{1}(Y_{i4}=2).\end{aligned} \tag{11}$$

$$\begin{aligned}\operatorname{logit}\{\Pr(R_{i3}=1)\} = \gamma_{30} &+ \gamma_{31}X_{i2}^{c} \\ &+ \gamma_{32}X_{i5}^{c} + \gamma_{33}\mathbf{1}(Y_{i5}=1) \\ &+ \gamma_{34}\mathbf{1}(Y_{i6}=3).\end{aligned} \tag{12}$$

where $X_{ij}^{c} = X_{ij} - 2$ denotes the centred covariate. The intercept parameters were calibrated to produce average missingness levels of approximately 10% and 20% for Y_2 and Y_3, respectively. The performance of $\mathrm{BLCM_{MAR}}$ and $\mathrm{BLCM_0}$ across the simulation scenarios is presented and discussed in the following section.

4. Results and Discussions

This section presents the simulation results comparing the proposed covariate-dependent Bayesian latent class multiple imputation model ($\mathrm{BLCM_{MAR}}$) with the conventional Bayesian latent class model without covariates ($\mathrm{BLCM_0}$) under MAR mechanism. The results are discussed in terms of multiple imputation performance, latent class recovery, and regression coefficient recovery across the three simulation scenarios.

4.1. Multiple Imputation Inference Performance

The MI inference performance was evaluated using bias, RMSE, posterior stability, and empirical 95% coverage under the three simulation scenarios.

Table 5. *Multiple imputation performance under MAR.*

Scenario	n	Method	Bias	RMSE	Stability	Cov. 95%
A	500	$\mathrm{BLCM_0}$	0.0000	0.0089	0.0185	0.9000
A	500	$\mathrm{BLCM_{MAR}}$	0.0000	0.0089	0.0188	0.8889
A	1000	$\mathrm{BLCM_0}$	0.0000	0.0053	0.0152	0.9778
A	1000	$\mathrm{BLCM_{MAR}}$	0.0000	*0.0052*	0.0153	0.9889
B	500	$\mathrm{BLCM_0}$	0.0000	0.0085	0.0174	0.9667
B	500	$\mathrm{BLCM_{MAR}}$	0.0000	*0.0079*	0.0172	0.9667
B	1000	$\mathrm{BLCM_0}$	0.0000	*0.0064*	0.0114	0.9556
B	1000	$\mathrm{BLCM_{MAR}}$	0.0000	0.0068	0.0114	0.9333
C	500	$\mathrm{BLCM_0}$	0.0000	*0.0086*	0.0230	0.9444
C	500	$\mathrm{BLCM_{MAR}}$	0.0000	0.0087	0.0236	0.9333
C	1000	$\mathrm{BLCM_0}$	0.0000	*0.0056*	0.0146	0.9222
C	1000	$\mathrm{BLCM_{MAR}}$	0.0000	0.0059	0.0144	0.9444

Note. Bias, RMSE, posterior stability, and empirical 95% coverage are averaged across the target category-probability parameters. The lowest RMSE for each scenario and sample size is shown in bold.

Across all simulation scenarios, both models produced negligible bias, indicating that the corresponding estimators were essentially unbiased. Consequently, differences in inferential performance are primarily reflected in the RMSE, posterior stability, and empirical coverage.

Under Scenario A, where covariates do not influence latent class membership, both models exhibited nearly identical performance. At the smaller sample size ($n = 500$), identical RMSE values were obtained, while at $n = 1000$ the proposed $\mathrm{BLCM_{MAR}}$ achieved a marginally lower RMSE than $\mathrm{BLCM_0}$. These findings indicate that incorporating covariates does not adversely affect inference when the covariates carry no information about latent class membership.

Under Scenario B, representing moderate covariate effects, $\mathrm{BLCM_{MAR}}$ achieved the lowest RMSE for the smaller sample size, demonstrating the benefit of incorporating covariate information when latent class membership depends on observed characteristics. At $n = 1000$, both models produced comparable estimates, with only minor differences in RMSE and posterior stability.

For Scenario C, involving strong covariate effects, the two models again exhibited similar inferential performance. Although $\mathrm{BLCM_0}$ attained slightly lower RMSE values, the differences were small, and both models maintained satisfactory posterior stability and empirical 95% coverage. These results suggest that incorporating covariates neither degrades estimation accuracy nor introduces substantial additional uncertainty, even under strong covariate effects.

Across all scenarios, posterior stability improved as the sample size increased from 500 to 1000, reflecting the increased precision associated with larger samples. Similarly, empirical 95% coverage probabilities remained close to their nominal level for both models, indicating reliable interval estimation under the MAR mechanism.

Overall, the proposed $\mathrm{BLCM_{MAR}}$ performs competitively with $\mathrm{BLCM_0}$ while preserving unbiased estimation, stable posterior inference, and appropriate interval coverage. More importantly, the model provides measurable gains in inferential accuracy when covariates contribute to latent class membership, while maintaining performance comparable to $\mathrm{BLCM_0}$ when covariate effects are absent. These findings demonstrate that incorporating covariates into Bayesian latent class multiple imputation offers a flexible and effective extension of the conventional latent class model under the MAR assumption without compromising inferential reliability.

To further examine the behaviour of the proposed multiple imputation procedure, the RMSE was evaluated over increasing sample sizes. Figure 1 presents the RMSE for sample sizes $n = 500, 1000, 1500, 2000$, and 2500 under the three simulation scenarios. As the sample size increases, posterior distributions increasingly concentrated around the true parameter values, leading to improved estimation precision. Consequently, a systematic reduction in RMSE provides empirical evidence of posterior concentration under the proposed Bayesian latent class multiple imputation model.

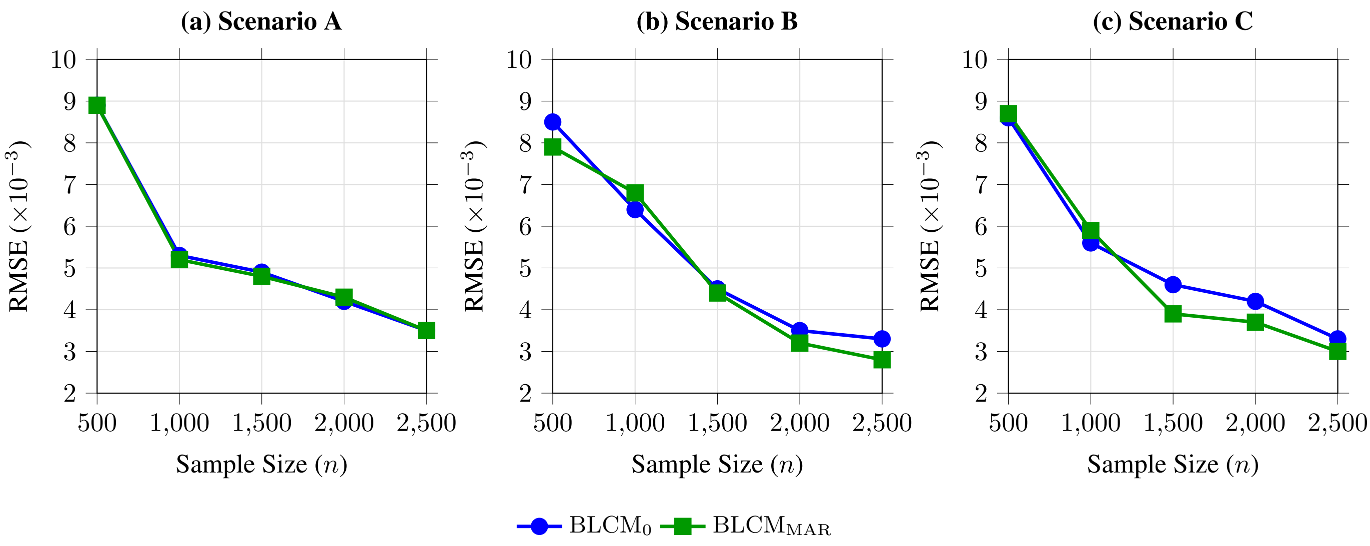

Figure 1. *RMSE as a function of sample size for BLCM$_0$ and BLCM$_{\mathrm{MAR}}$ under the three simulation scenarios.*

Figure 1(a) shows that, under Scenario A, the RMSE decreases steadily as the sample size increases for both competing models. Since no covariate effects are present in this scenario, BLCM$_0$ and BLCM$_{\mathrm{MAR}}$ exhibit nearly identical estimation accuracy across the range of sample sizes considered.

Under Scenario B (Figure 1(b)), both models again exhibit a consistent reduction in RMSE with increasing sample size. However, BLCM$_{\mathrm{MAR}}$ achieves lower RMSE than BLCM$_0$ at the larger sample sizes, demonstrating the benefit of incorporating covariate information when latent class membership depends on observed covariates.

A similar pattern is observed under Scenario C (Figure 1(c)), where RMSE declines steadily as the sample size increases for both models. Although the differences between the models are modest, BLCM$_{\mathrm{MAR}}$ maintains estimation accuracy comparable to BLCM$_0$ while accounting for covariate-dependent class membership.

Across all simulation scenarios, the progressive reduction in RMSE with increasing sample size demonstrates improving estimation precision under both models. This empirical behaviour is consistent with increasing posterior concentration, whereby larger samples provide more information about the model parameters, resulting in progressively more precise posterior inference. Together with the negligible bias, stable interval coverage, and improved estimation accuracy observed in the preceding analyses, these results provide strong empirical support for the proposed BLCM$_{\mathrm{MAR}}$.

4.2. Latent Class Proportion Recovery

Class proportion recovery was evaluated using the bias and RMSE of the estimated latent class proportions under the three simulation scenarios. The corresponding results are presented in Table 6.

Table 6. *Class proportion recovery under the MAR mechanism.*

Scenario	n	Method	Bias	RMSE
A	500	BLCM$_0$	0.0000	*0.0410*
A	500	BLCM$_{\mathrm{MAR}}$	0.0000	0.0920
A	1000	BLCM$_0$	0.0000	*0.0450*
A	1000	BLCM$_{\mathrm{MAR}}$	0.0000	0.0983
B	500	BLCM$_0$	0.0000	*0.0543*
B	500	BLCM$_{\mathrm{MAR}}$	0.0000	0.0827
B	1000	BLCM$_0$	0.0000	0.0444
B	1000	BLCM$_{\mathrm{MAR}}$	0.0000	*0.0368*
C	500	BLCM$_0$	0.0000	0.0399
C	500	BLCM$_{\mathrm{MAR}}$	0.0000	*0.0324*
C	1000	BLCM$_0$	0.0000	0.0277
C	1000	BLCM$_{\mathrm{MAR}}$	0.0000	*0.0231*

Note. Bias and RMSE are averaged across the three latent class proportions. Bold RMSE denotes the better-performing model for each simulation setting.

Across all simulation scenarios, both models produced negligible bias, indicating accurate recovery of the latent class proportions. Consequently, differences in performance are primarily reflected in the RMSE values.

Under Scenario A, where covariates do not influence latent class membership, BLCM$_0$ consistently achieved lower RMSE than BLCM$_{\mathrm{MAR}}$. This result is expected because the data-generating mechanism does not involve covariate effects, and the additional regression parameters in BLCM$_{\mathrm{MAR}}$ provide little benefit for estimating the class proportions.

Under Scenario B, representing moderate covariate effects, BLCM$_0$ achieved lower RMSE at the smaller sample size, whereas BLCM$_{\mathrm{MAR}}$ produced more accurate class proportion estimates at $n = 1000$. This finding suggests that the benefits of incorporating covariates become more evident as the amount of information available for estimating the regression parameters increases.

For Scenario C, corresponding to strong covariate effects, $\mathrm{BLCM}_{\mathrm{MAR}}$ consistently achieved the lowest RMSE across both sample sizes. The improved recovery of the latent class proportions demonstrates the advantage of explicitly modelling covariate-dependent class membership when covariates are strongly associated with the latent classes. Accounting for these covariate effects reduces model misspecification and improves the accuracy of the estimated class proportions [36, 37].

Taken together, these results demonstrate that the relative performance of the competing models depends on the strength of the covariate effects. When covariates are absent, the simpler BLCM_0 provides the most accurate recovery of latent class proportions. However, as covariate effects become increasingly pronounced, the proposed $\mathrm{BLCM}_{\mathrm{MAR}}$ consistently provides more accurate estimates. These findings confirm that incorporating covariates into the Bayesian latent class model improves latent class recovery when covariates contain meaningful information while preserving reliable estimation across all simulation scenarios.

4.3. Recovery of Regression Coefficients

The performance of the proposed $\mathrm{BLCM}_{\mathrm{MAR}}$ in recovering the regression coefficients governing covariate-dependent latent class membership was evaluated using bias, RMSE, posterior stability, empirical coverage probabilities, and credible interval widths. Because regression coefficients are unique to the proposed covariate-dependent model, this evaluation focuses exclusively on $\mathrm{BLCM}_{\mathrm{MAR}}$. The corresponding results are presented in Tables 7 and 8.

Table 7. *Recovery of regression coefficients under the MAR mechanism.*

Scenario	n	Bias	RMSE	Stability
A	500	-0.0988	0.4611	0.4781
A	1000	-0.0625	0.4457	0.4399
B	500	-0.1139	0.5696	0.5455
B	1000	-0.1651	0.4725	0.3682
C	500	-0.1262	0.8713	0.5589
C	1000	-0.1136	0.7699	0.5657

Note. Bias, RMSE, and stability are averaged across the regression coefficients governing covariate-dependent latent class membership.

Across all simulation scenarios, the proposed $\mathrm{BLCM}_{\mathrm{MAR}}$ recovered the regression coefficients with relatively small bias, indicating that the estimated covariate effects remained close to their true values. Compared with the recovery of latent class proportions, estimation of regression coefficients exhibited larger RMSE and posterior variability, reflecting the greater complexity of estimating multinomial logistic regression parameters within the LCM.

Under Scenario A, where covariates do not influence latent class membership, the regression coefficients are centred close to zero and are therefore inherently more difficult to estimate precisely. Nevertheless, both RMSE and posterior variability decreased slightly as the sample size increased from $n = 500$ to $n = 1000$, indicating improved estimation precision.

For Scenarios B and C, representing moderate and strong covariate effects, respectively, recovery of the regression coefficients likewise improved with increasing sample size. Although stronger covariate effects resulted in larger estimation error than observed under Scenario A, the reduction in RMSE under the larger sample size demonstrates that additional information improves estimation of the regression parameters. These findings are consistent with the expected behaviour of Bayesian estimators, whereby posterior uncertainty decreases as more data become available [38, 39].

Empirical coverage probabilities and average credible interval widths are summarized in Table 8. Coverage was evaluated using 90%, 95%, and 99% credible intervals, while the corresponding interval widths provide an indication of posterior uncertainty.

Table 8. *Coverage of regression coefficients under the MAR mechanism.*

Scenario	n	Cov_{90}	Cov_{95}	Cov_{99}	Width_{90}	Width_{95}	Width_{99}
A	500	0.8762	0.9143	0.9571	1.7567	2.0939	2.7538
A	1000	0.8905	0.9524	0.9857	0.9081	1.0824	1.4235
B	500	0.7619	0.8143	0.8762	1.3112	1.5628	2.0554
B	1000	0.4429	0.4762	0.5952	0.5930	0.7068	0.9295
C	500	0.4714	0.5524	0.6286	1.0820	1.2897	1.6961
C	1000	0.4524	0.4952	0.5714	0.6169	0.7353	0.9671

Note. Cov_{90}, Cov_{95} and Cov_{99} denote empirical coverage probabilities for the corresponding credible intervals. Width_{90}, Width_{95} and Width_{99} denote the associated average credible interval widths. Results are averaged across all regression coefficients.

Coverage probabilities were generally close to their nominal levels under Scenario A, indicating satisfactory uncertainty quantification when covariate effects were absent. As expected, interval widths decreased with increasing sample size, reflecting improved posterior precision.

For Scenarios B and C, empirical coverage declined below the nominal levels, particularly for the 95% credible intervals. This behaviour reflects the increased complexity of estimating multinomial logistic regression coefficients when covariate effects are present. Although interval widths became narrower with increasing sample size, the reduction in coverage indicates that finite-sample uncertainty remains challenging to quantify accurately in more complex latent class structures. Similar behaviour has been reported for Bayesian latent variable models in finite samples, where posterior uncertainty converges more slowly for regression parameters than for probability parameters [38–40].

Overall, the proposed $\mathrm{BLCM}_{\mathrm{MAR}}$ provides reliable estimation of the regression coefficients governing covariate-dependent latent class membership. Estimation accuracy and posterior precision improve with increasing sample size, while the finite-sample reduction in interval coverage under moderate and strong covariate effects highlights the greater difficulty of uncertainty quantification for regression parameters compared with latent class proportions. These findings complement the earlier results on latent class recovery and further demonstrate the effectiveness of the proposed covariate-dependent Bayesian latent class multiple imputation model under the MAR mechanism.

5. Conclusion

This study developed a covariate-dependent Bayesian latent class model for multiple imputation under MAR mechanism. By incorporating covariates into the latent class membership model, the proposed approach extends the conventional Bayesian latent class model ($BLCM_0$) while preserving the flexibility of latent class modelling for multivariate categorical data with missing values. The simulation study demonstrated that the proposed model yields unbiased inference, reliable recovery of latent class proportions and regression coefficients, and improved estimation accuracy when covariates are informative of latent class membership. Furthermore, the observed reduction in RMSE and improvement in posterior precision with increasing sample size provide empirical support for the desirable large-sample behaviour of the proposed Bayesian multiple imputation framework.

The proposed methodology therefore provides an effective extension of Bayesian latent class multiple imputation under the MAR assumption, particularly in applications where auxiliary covariates contribute to explaining latent class membership. Future research may extend the proposed framework to accommodate MNAR mechanisms through explicit modelling of the missingness process. It may also investigate the asymptotic properties of the proposed methodology, including posterior consistency and other large-sample theoretical properties. Such developments could lead to a more comprehensive Bayesian latent class multiple imputation framework for handling both ignorable and non-ignorable missing data mechanisms.

ORCID

0009-0001-7215-3239 (Daniel Fundi Murithi)

Abbreviations

BLCM	Bayesian Latent Class Models
LCM	Latent Class Models
MAR	Missing at Random
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MICE	Multiple Imputation Chained Equations
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MNAR	Missing Not at Random
RMSE	Root Mean Squared Error

Acknowledgments

The author acknowledges the research supervisors': Dr. Josephine Njeri Ngure and Dr. Victor Muthama Musau for their invaluable guidance and scholarly insights throughout the development of this study. Their contributions greatly enhanced the methodology.

Author Contributions

Daniel Fundi Murithi: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing

Josephine Njeri Ngure: Supervision, Validation, Writing – review & editing

Victor Muthama Musau: Supervision, Validation, Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] O. F. Ayilara, L. Zhang, T. T. Sajobi, R. Sawatzky, E. Bohm, and L. M. Lix, "Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry," *Health and Quality of Life Outcomes*, vol. 17, no. 1, p. 106, 2019, https://doi.org/10.1186/s12955-019-1181-2

[2] T. B. Pepinsky, "A note on listwise deletion versus multiple imputation," *Political Analysis*, vol. 26, no. 4, pp. 480–488, 2018, https://doi.org/10.1017/pan.2018.18

[3] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, 1987, https://doi.org/10.1002/9780470316696

[4] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2018, https://doi.org/10.1201/9780429492259

[5] C. K. Enders, *Applied Missing Data Analysis*. New York, NY: Guilford Publications, 2022.

[6] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. John Wiley & Sons, 2019, https://doi.org/10.1002/9781119482260

[7] J. L. Schafer and R. M. Yucel, "Computational strategies for multivariate linear mixed-effects models with missing values," *Journal of Computational and Graphical Statistics*, vol. 11, no. 2, pp. 437–457, 2002, https://doi.org/10.1198/106186002760180608

[8] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC, 1997, https://doi.org/10.1201/9781439821862

[9] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011, https://doi.org/10.18637/jss.v045.i03

[10] J. K. Vermunt, J. R. Van Ginkel, L. A. Van der Ark, and K. Sijtsma, "Multiple imputation of incomplete categorical data using latent class analysis," *Sociological Methodology*, vol. 38, no. 1, pp. 369–397, 2008, https://doi.org/10.1111/j.1467-9531.2008.00202.x

[11] D. Vidotto, J. K. Vermunt, and K. Van Deun, "Bayesian latent class models for the multiple imputation of categorical data," *Methodology*, vol. 14, no. 2, pp. 56–68, 2018, https://doi.org/10.1027/1614-2241/a000146

[12] G. J. McLachlan and D. Peel, *Finite Mixture Models*. John Wiley & Sons, 2000, https://doi.org/10.1002/0471721182

[13] D. W. van der Palm, L. A. van der Ark, and J. K. Vermunt, "Divisive latent class modeling as a density estimation method for categorical data," *Journal of Classification*, vol. 33, no. 1, pp. 52–72, 2016, https://doi.org/10.1007/s00357-016-9195-5

[14] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley & Sons, 2002, https://doi.org/10.1002/9781119013563

[15] P. D. Allison, "Missing data," in *The SAGE Handbook of Quantitative Methods in Psychology*, R. E. Millsap and A. Maydeu-Olivares, Eds. Thousand Oaks, CA: SAGE Publications, 2009, pp. 72–89, https://doi.org/10.4135/9780857020994.n4

[16] M. Gebregziabher and S. M. DeSantis, "Latent class based multiple imputation approach for missing categorical data," *Journal of Statistical Planning and Inference*, vol. 140, no. 11, pp. 3252–3262, 2010, https://doi.org/10.1016/j.jspi.2010.04.020

[17] P. F. Lazarsfeld and N. W. Henry, *Latent Structure Analysis*. Boston: Houghton Mifflin, 1968.

[18] L. A. Goodman, "The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach," *American Journal of Sociology*, vol. 79, no. 5, pp. 1179–1259, 1974, https://doi.org/10.1086/225676

[19] M. Janković, E. Masthoff, M. Spreen, P. De Looff, and S. Bogaerts, "A latent class analysis of forensic psychiatric patients in relation to risk and protective factors," *Frontiers in Psychology*, vol. 12, p. 695354, 2021, https://doi.org/10.3389/fpsyg.2021.695354

[20] B. Hoogerheide, E. Visser, E. T. Maas, S. H. Booij, and H. Riese, "Latent class analysis in a transdiagnostic psychiatric population to identify classes of psychosocial functioning: Using data from routine clinical practice," *Journal of Affective Disorders*, vol. 379, pp. 710–720, 2025, https://doi.org/10.1016/j.jad.2025.02.066

[21] T. H. Bijmolt, L. J. Paas, and J. K. Vermunt, "Country and consumer segmentation: Multi-level latent class analysis of financial product ownership," *International Journal of Research in Marketing*, vol. 21, no. 4, pp. 323–340, 2004, https://doi.org/10.1016/j.ijresmar.2004.06.002

[22] C. C. Clogg, "New developments in latent structure analysis," in *Factor Analysis and Measurement in Sociological Research*, S. Leinhardt, Ed. Beverly Hills, CA: Sage Publications, 1981, pp. 215–246.

[23] K. Bandeen-Roche, D. L. Miglioretti, S. L. Zeger, and P. J. Rathouz, "Latent variable regression for multiple discrete outcomes," *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1375–1386, 1997, https://doi.org/10.1080/01621459.1997.10473658

[24] C. C. Clogg and L. A. Goodman, "Latent structure analysis of a set of multidimensional contingency tables," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 762–771, 1984, https://doi.org/10.1080/01621459.1984.10477093

[25] J. K. Vermunt and J. Magidson, "Latent class cluster analysis," in *Applied Latent Class Analysis*, J. A. Hagenaars and A. L. McCutcheon, Eds. Cambridge: Cambridge University Press, 2002, pp. 89–106, https://doi.org/10.1017/CBO9780511499531.004

[26] J. K. Vermunt, "Latent class modeling with covariates: Two improved three-step approaches," *Political Analysis*, vol. 18, no. 4, pp. 450–469, 2010, https://doi.org/10.1093/pan/mpq025

[27] C. M. Dayton and G. B. Macready, "Concomitant-variable latent-class models," *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 173–178, 1988, https://doi.org/10.1080/01621459.1988.10478584

[28] K. Nylund-Gibson and A. Y. Choi, "Ten frequently asked questions about latent class analysis," *Translational Issues in Psychological Science*, vol. 4, no. 4, pp. 440–461, 2018, https://doi.org/10.1037/tps0000176

[29] A. Agresti, *Categorical Data Analysis*, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2013, https://doi.org/10.1002/9781119405269

[30] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. Boca Raton, FL: Chapman & Hall/CRC, 2013, https://doi.org/10.1201/b16018

[31] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. New York, NY: Springer, 2007, https://doi.org/10.1007/0-387-71599-1

[32] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976, https://doi.org/10.1093/biomet/63.3.581

[33] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, 1984, https://doi.org/10.1109/TPAMI.1984.4767596

[34] N. G. Polson, J. G. Scott, and J. Windle, "Bayesian inference for logistic models using Pólya–Gamma latent variables," *Journal of the American Statistical Association*, vol. 108, no. 504, pp. 1339–1349, 2013, https://doi.org/10.1080/01621459.2013.829001

[35] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC, 1997, https://doi.org/10.1201/9781439821862

[36] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press, 2022.

[37] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, 1st ed. New York, NY: Springer, 2004, https://doi.org/10.1007/978-0-387-21736-9

[38] A. W. van der Vaart, *Asymptotic Statistics*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK: Cambridge University Press, 1998, https://doi.org/10.1017/CBO9780511802256

[39] S. Ghosal and A. W. van der Vaart, *Fundamentals of Nonparametric Bayesian Inference*. Cambridge, UK: Cambridge University Press, 2017, https://doi.org/10.1017/9781139029834

[40] L. S. T. Ho, B. T. Nguyen, V. Dinh, and D. Nguyen, "Posterior concentration and fast convergence rates for generalized Bayesian learning," *arXiv preprint arXiv:2111.10243*, 2021. [Online]. Available: https://arxiv.org/abs/2111.10243