Journal of Mathematics and Statistics

Research Article

Parameter-Expanded Data Augmentation for Analyzing Nominal Data With Missing Values Using Multinomial Probit Models

Suwash Silwal and Xiao Zhang

Department of Mathematical Sciences, Michigan Technological University, Houghton, MI, USA

Article history
Received: 09-01-2026
Revised: 09-04-2026
Accepted: 15-04-2026

Corresponding Author:
Zhang Xiao
Department of Mathematical Sciences, Michigan Technological University, Houghton, MI, USA
Email: xzhang35@mtu.edu

Abstract: Nominal data occur in many scientific fields, such as health-related studies, transportation, and econometrics. Due to the ubiquitousness of missing data, it is inevitable to analyze nominal data with missing values. As is well known, statistical methods for analyzing nominal variables are quite limited. The Multinomial Probit (MNP) model has been an essential tool for analyzing nominal categorical data, but the computational complexity of maximum likelihood functions and stringent model identification bring a rigorous task for both likelihood-based estimation and Markov Chain Monte Carlo (MCMC) sampling, hence confine the utilization of this important model. Advanced developments in Bayesian computation have shown the promising performance of Parameter-Expanded Data Augmentation (PX-DA) in MCMC sampling. Accordingly, in this investigation, we propose PX-DA to analyze nominal data with missing values using MNP models. We conduct our investigation through simulation studies and an application to the Mental Health Client-Level Data. Our investigation demonstrates that the proposed methods significantly improve the convergence and mixing of MCMC sampling components and can handle nominal data with substantial missingness, and thus enhance the practical usage of MNP models in the field of discrete data analysis.

Keywords: Multinomial Probit (MNP) Model, Markov Chain Monte Carlo (MCMC), Parameter-Expanded Data Augmentation (PX-DA), Missing Data

Introduction

The MNP model is a discrete choice model for individual decision-making and has been commonly explored to analyze nominal data due to its well-known advantage over the popular logit model, i.e., its relaxation of independence of irrelevant alternatives et al., 2018; Hausman and Wise, 1978; McCulloch and Rossi, 1994). However, it is challenging to be utilized due to its computational complexity of maximum likelihood functions and fragile model identification.

Frequentist methods have been proposed to address the computational challenges, such as a minimum distance method (Berndt et al., 1974) approximation methods (Clark, 1961; Horowitz et al., 1982; Kamakura and Russell, 1989); the method of simulated moments estimator (McFadden, 1989; Pakes and Pollard, 1989) and simulated maximum likelihoods (Geweke, 1991; Hajivassiliou and McFadden, 1998; Keane, 1992, 1994). Other related work can also be referred to Bunch (1991); Bunch and Kitamura (1991); Geweke et al. (1994); Hajivassiliou et al. (1996); Train (2009).

The rapid advancements in Bayesian computation have significantly rekindled the interest in the study of the MNP model. In particular, the emergence of Gibbs sampling (Gelfand and Smith, 1990) and the development of MCMC methods (Tanner and Wong, 1987; Gilks et al., 1995) have laid the foundation in developing MCMC sampling algorithms for the MNP model. Pioneering works include Albert and Chib (1993); McCulloch and Rossi (1994); Nobile (1998), McCulloch et al. (2000); Imai and van Dyk (2005); Zhang et al. (2008); McCulloch and Rossi (1994) conducted seminal work using Gibbs sampling to estimate the MNP model. Nobile (1998) proposed a hybrid Markov chain that includes a Metropolis step to improve the convergence and mixing of the MCMC sampling developed by McCulloch and Rossi (1994). Further, McCulloch et al. (2000) modified the approach of McCulloch and Rossi (1994) by addressing non-identification through prior specification.

SCIENCE Publications

© 2026 Suwash Silwal and Xiao Zhang, Vijaya Chandra Jadala and Y Sreeman. This open-access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

However, this method suffers slow convergence, commented by Nobile (2000). Albert and Chib (1993); Zhang et al. (2008) proposed rigorous Metropolis-Hastings (MH) algorithms to sample a restricted covariance matrix due to the model identification issue.

PX-DA has been proven to be prominent in the convergence and mixing of the MCMC sampling components (Liu and Wu, 1999; Meng and Van Dyk, 1999); Lawrence et al. (2008); Zhang (2020, 2022) investigated the matter using multivariate probit models for multivariate ordinal data, while Zhang (2025) extended to multivariate ordinal data with missing values. Imai and van Dyk (2005); Zhang (2026) implemented it to MNP models for nominal data without missingness. However, the omnipresence of missing data requires statistical methods that can handle missing nominal data.

Rubin (1976) introduced the concept of missing data mechanism, which can be categorized as Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). The MCAR means the missingness is irrelevant to the variables included in the model; the MAR means the missingness is relevant to the observed values of the variables included in the model; and the MNAR means the missingness is relevant to the unobserved values of the variables included in the model. Ignorability requires both MAR and independent priors of the parameters of the model and those of the missingness, i.e., those parameters are distinct. Under the assumption of ignorability, the likelihood-based and Bayesian inferences can be solely based on the observed quantities without taking into consideration the distribution of the missingness. However, without ignorability, the statistical inference must take account of the missing data mechanism, and the estimation can be sensitive to the assumption of the non-ignorable model, for which the data contains no information (Carpenter and Smuk, 2021; Harel and Zhou, 2007; Little and Rubin, 2019).

In this investigation, we develop PX-DA to analyze nominal data with missing values using MNP models and assume the ignorable missing data mechanism, which is the basis for Bayesian missing data analysis and multiple imputation (Buuren, 2018; Buuren and Groothuis-Oudshoorn, 2011; Schafer, 1997; Wijesuriya et al., 2025; Yucel, 2011).

Materials and Methods

The MNP Model

Let $Y_i = \left(Y_{i1}, \dots, Y_{i(k+1)}\right)^\top$ denote a multinomial vector, where $i = 1, \dots, n$ indexes subjects and $j = 1, \dots, k+1$ indexes categories. Define $Y_{ij} = 1$ if subject i chooses category j and $Y_{ij} = 0$ otherwise. To simplify, we introduce a vector $d = (d_1, \dots, d_n)^\top$, where each entry corresponds to the chosen category for subject i, i.e., $d_i = j$ is equivalent to $Y_{ij} = 1$. The MNP model assumes an underlying latent utility vector $U_i = \left(U_{i1}, \dots, U_{i(k+1)}\right)^\top$ associated with each observed Y_i. The observed category is determined by the maximum component of U_i, reflecting the notion that the individual selects the category with the maximum latent utility score. Formally, this can be stated as follows:

$$\mathrm{d_i = j \Leftrightarrow U_{ij} \geq \max_{1\leq l\leq (k+1)} U_{il}} \tag{1}$$

Further, the MNP assumes that the vector U_i is distributed as a multivariate normal distribution with mean $A_i\,\beta$ and covariance matrix S, where A_i denotes the $(k+1) \times p$ covariate matrix for subject i, and β represents the $p \times 1$ regression parameter vector. i.e:

$$U_i = A_i\beta + \delta_i, \tag{2}$$

Where $\delta_i \sim \mathcal{N}_{k+1}(0, S)$, a multivariate normal distribution. Each row of A_i represents either subject-specific covariates with identical elements across the alternatives, or outcome-specific covariates where elements may differ across the alternatives. It is noticeable that the model defined in Eq. 1 is not identifiable. The model remains unchanged if any constant is added or multiplied to both sides. The identifiable MNP model is commonly specified as follows:

$$d_i = \begin{cases} 0, & if \max_{1\leq l\leq k} Z_{il} < 0, \\ j, & if \max_{1\leq l\leq k} Z_{il} = Z_{ij} > 0 \end{cases} \tag{3}$$

Where $Z_i \sim \mathcal{N}_k(X_i\beta, \Sigma_r)$ with X_i being the covariate matrix, β being the regression parameter vector, and Σ_r being the covariance matrix with the first element $\{\Sigma_\mathrm{r}\}_{11} = 1$ (Geweke et al., 1994; Imai and van Dyk, 2005; McCulloch et al., 2000; McCulloch and Rossi, 1994; Nobile, 1998; Zhang, 2026).

However, Σ_r, being a restricted covariance matrix with $\{\Sigma_r\}_{11} = 1$, results in a demanding task to develop efficient MCMC sampling methods, and rigorous MH algorithms are usually entailed (Albert and Chib, 1993; Zhang et al., 2008). Even with the circumvention of the MH algorithm, using Gibbs sampling still shows deprived convergence in MCMC sampling (McCulloch et al., 2000). Based on the seminal work of Liu and Wu (1999); Meng and Van Dyk (1999); MacEachern (2007), the MCMC samplings based on non-identifiable models may improve the convergence and mixing of the MCMC components in comparison with those based on the identifiable models. On that account, we construct a non-identifiable MNP model by assuming $Z_i \sim \mathcal{N}_k\left(\sigma_{11}^{-1/2}\, X_i\beta, \Sigma_r\right)$ in Eq. 2 (Zhang, 2026), where σ_{11} mimics an artificial variance parameter. Evidently, the identifiable model in Eq. 2 becomes non-identifiable due to involving the artificial parameter σ_{11}. Define $W_i = \sigma_{11}^{1/2} Z_i$, then correspondingly, the model can be written as:

$$\boldsymbol{d}_i = \begin{cases} 0, & if \max_{1\le l\le k} W_{il} < 0 \\ j, & if \max_{1\le l\le k} W_{il} = W_{ij} > 0 \end{cases} \tag{4}$$

With $W_i \sim \mathcal{N}_k(X_i\beta, \sigma_{11}\Sigma_r)$. In particular, $\sigma_{11}\Sigma_r$ is a covariance matrix without restriction and we denote $\Sigma = \sigma_{11}\Sigma_r$ in the following discussion.

Parameter-Expanded Data Augmentation for MNP Models With Missing Values

We then extend the work of Zhang (2025) for complete nominal data to missing nominal data. Denote $d = (d_{obs}, d_{mis})$, where $d_{obs} = \left(d_{obs,1}, \dots, d_{obs,n_1}\right)$ indicates the observed n_1 chosen categories and $d_{mis} = \left(d_{mis,n_1+1}, \dots, d_{mis,n}\right)$ indicates the unobserved $(n - n_1)$ chosen categories. Correspondingly, we denote $W = (W_{obs}, W_{mis})$ with $W_{obs} = \left(W_{obs,1}, \dots, W_{obs,n_1}\right)$ corresponding to d_{obs} and $W_{mis} = \left(W_{mis,n_1+1}, \dots, W_{mis,n}\right)$ corresponding to d_{mis}. Then the joint posterior density of β, Σ, W_{obs} and W_{mis} given d_{obs} can be characterized as:

$$p(\beta, \Sigma, W_{obs}, W_{mis} \mid d_{obs}) \propto p(\beta) \times p(\Sigma) \times \\ \prod_{i=1}^{n_1} I_{obs,i}\phi\left(W_{obs,i}; X_{obs,i}\beta, \Sigma\right) \times \\ \prod_{i=(n_1+1)}^{n} \phi\left(W_{mis,i}; X_{mis,i}\beta, \Sigma\right)$$

Where ϕ is the standard normal density function, and:

$$I_{obs,i} = 1_{(d_{obs,i}=0, W_{obs,il}<0, l=1,..,k)} + \sum_{l=1}^{k} 1_{\left(d_{obs,l}=1, W_{obs,il}=\max_{1\le h\le k}(W_{obs,ih},0)\right)}$$

Being an indicator function indicating the compatibility of $d_{obs,i}$ and $W_{\mathrm{obs},i}$.

Then, each full conditional distribution can be derived as follows:

- $\beta \mid (\Sigma, W_{obs}, W_{mis}, d_{obs}) \sim \mathcal{N}_p(\hat{\beta}, V_\beta)$, a multivaratie normal distribution, assuming the prior of β follows $\mathcal{N}_p(b, C)$, where $V_\beta = (\sum_{i=1}^{n} X_i^T \Sigma^{-1} X_i + C^{-1})^{-1}$ and $\hat{\beta} = V_\beta(\sum_{i=1}^{n} X_i^T \Sigma^{-1} W_i + C^{-1} b)$.
- $W_{\mathrm{obs},ij} \mid \left(\beta, \Sigma, d_{\mathrm{obs},i}, W_{\mathrm{obs},iq,q\neq j}\right)$ is a truncated normal distribution, i.e.,

$$p\left(W_{obs,ij} \mid \beta, \Sigma, d_{obs,i}, W_{obs,iq}, q \neq j\right) \\ = I_{obs,i} \times p\left(W_{obs,ij} \mid \beta, \Sigma, W_{obs,iq}, q \neq j\right) \\ = I_{obs,i} \times \phi\left(W_{obs,ij}; m_{obs,ij}, \tau_{obs,ij}\right)$$

Where, $m_{obs,ij}$ and $\tau_{obs,ij}$ are conditional mean and variance of $W_{obs,ij}$ given $W_{obs,i(-j)}$:

- $W_{mis,ij} \mid \left(\beta, \Sigma, d_{obs,i}, W_{mis,iq,q\neq j}\right) \propto W_{mis,ij} \mid \left(\beta, \Sigma, W_{mis,iq,q\neq j}\right)$ is a normal distribution without truncation, since d_{obs} is only related to W_{obs}, and contains no information about W_{mis}, thus:

$$p\left(W_{mis,ij} \mid \beta, \Sigma, d_{obs,i}, W_{mis,iq}, q \neq j\right) \\ = p\left(W_{mis,ij} \mid \beta, \Sigma, W_{mis,iq}, q \neq j\right) \\ = \phi\left(W_{mis,ij}; m_{mis,ij}, \tau_{mis,ij}\right)$$

Where $m_{\mathrm{mis},ij}$ and $\tau_{mis,ij}$ are conditional mean and variance of $W_{\mathrm{mis},ij}$ given $W_{\mathrm{mis},i(-j)}$:

- $\Sigma \mid (\beta, W_{obs}, W_{mis}, d_{obs}) \sim \mathcal{W}^{-1}(V^*, m^*)$, an inverse-Wishart distribution, assuming a conjugate prior for $\Sigma \sim \mathcal{W}^{-1}(V, m)$, where, $V^* = \sum_{i=1}^{n} (W_i - X_i\beta)(W_i - X_i\beta)^\top + V$ and $m^* = n + m + k + 1$

Then, the MCMC sampling scheme can be implemented by four Gibbs steps, i.e., a multivariate normal distribution for $\beta \mid (\Sigma, W_{obs}, W_{mis}, d_{obs})$, a univariate truncated normal distribution for each $W_{\mathrm{obs},ij} \mid \left(\beta, \Sigma, d_{\mathrm{obs},i}, W_{\mathrm{obs},iq,q\neq j}\right)$, a univariate normal distribution without truncation for each $W_{mis,ij} \mid \left(\beta, \Sigma, d_{obs,i}, W_{mis,iq,q\neq j}\right)$ and an inverse-Wishart distribution for $\Sigma \mid (\beta, W_{obs}, W_{mis}, d_{obs})$. One key point is that based on sampled $W_{\mathrm{mis},i}$, the missing value $d_{\mathrm{mis},i}$ can be imputed, i.e., if $W_{mis,il} < 0, l = 1, \dots, k$, then $d_{mis,i} = 0$; otherwise if $W_{mis,ij} = \max_{1\le h\le k}\left(W_{mis,ih}, 0\right)$, then $d_{mis,i} = j$. Another important feature is that sampling covariance matrix $\boldsymbol{\Sigma}$ is a Gibbs sampling from the inverse-Wishart distribution stemmed from the non-identifiable model constructed in Eq. 3 instead of a demanding MH algorithm required by the identifiable model in Eq. 2. We refer to this algorithm as the parameter-expanded Gibbs sampling algorithm (PX-GS).

Based on Liu and Wu (1999), σ_{11}, the artificial variance parameter, can be marginalized out through two steps:

- Sampling $(\mathrm{W}_{\mathrm{obs}}, \mathrm{W}_{\mathrm{mis}}, \sigma_{11}) \mid (\beta, \Sigma_r, \mathrm{d}_{\mathrm{obs}})$ by sampling $\sigma_{11} \mid (\beta, \Sigma_r, \mathrm{d}_{\mathrm{obs}}) \propto \sigma_{11} | \Sigma_r$ followed by sampling $(W_{obs}, W_{mis}) \mid (\beta, \Sigma_r, \sigma_{11}, d_{obs}) \propto (W_{obs}, W_{mis}) \mid (\beta, \Sigma, d_{obs})$. Here:

$$\sigma_{11} \mid \boldsymbol{\Sigma}_r \sim \textit{Scaled-inv-}\,\chi^2_{[mk, tr(\boldsymbol{V}\boldsymbol{\Sigma}_r^{-1})]}$$

- Sampling $(\beta, \Sigma_r, \sigma_{11}) \mid (W_{obs}, W_{mis}, d_{obs})$ by sampling $\beta \mid (\Sigma, W_{obs}, W_{mis}, d_{obs})$ followed by sampling $\Sigma | (\beta, W_{obs}, W_{mis}, d_{obs})$

We denote this algorithm as the parameter-expanded Gibbs sampling with marginalization (PX-GSM).

Simulation Studies

We conduct simulation studies to investigate the performance of the proposed PX-GS and PX-GSM based on the nonidentifiable model and compare them with PX-MH based on the identifiable model, which entails the MH algorithm for sampling the restricted covariance matrix (Zhang et al., 2008). We generate six-choice categorical outcome with two covariates drawn from *Bernoulli (0.5)* distribution. The regression parameters $\beta = (\beta_1, \beta_2)$ are set at $(1, 2)$, and the 5-dimensional covariance matrix $(k = 5), \Sigma$, for the latent variable being a compound symmetry with equal correlation being 0.5. We assume relatively weak priors: A prior for β being $\mathcal{N}(0_2, 100\mathrm{I}_2)$ and an inverse-Wishart prior for Σ being $\mathcal{W}^{-1}(V, m)$ with $m = 15$ and $V = (m - k - 1)\mathrm{I}_\mathrm{k}$. Here I_2 and I_k are identity matrices.

We consider two missing data mechanisms: MCAR and MAR. For the MCAR mechanism, three levels of missing data are considered: $20\%, 40\%$, and 60%. For each observation of Y_i, a random number is drawn from a $Uniform\ distribution(0,1)$. If the generated number is less than the specified missing rate (e.g., $0.20, 0.40$, or 0.60), the corresponding Y_i is assigned as missingness. This procedure ensures that the missingness is completely random and is unrelated to any observed or unobserved information, thereby satisfying the assumption of MCAR.

For the MAR mechanism, missingness in Y is determined by the joint pattern of two fully observed covariates, X_1 and X_2. Each pattern - $(1, 1), (0, 1), (1, 0)$, and $(0, 0)$ is assigned a base probability of missingness to achieve overall rates of $20\%, 40\%$, or 60%. For example, in the 60% case, the base probabilities are $0.20, 0.15, 0.15$, and 0.10 for the respective patterns, which are then scaled to produce an overall missingness of approximately 60% across the dataset.

We investigate sample sizes of 500, 2000, and 4000, respectively. For each scenario, we generate 100 simulated datasets and run 20,000 iterations with 5,000 burn-in. MCMC convergence diagnostics are performed using the R package coda (Plummer et al., 2006). Our investigation shows that the results for MCAR and MAR are similar. Basically, increasing missing data percentages exacerbates the absolute relative bias (ARB) and root mean square error (RMSE) values, i.e., the data with 20% missingness have the smallest ARB and RMSE values while those with 60% missingness have the largest of those values. For illustration purposes, we present those of MAR with 60% missingness in Tables 1-3 and Figures 1-3. The tables exhibit in general PX-GSM has the smallest ARB and RMSE values with the best 95% credible interval coverage probabilities (CP) among the three algorithms. However, PX-MH produces the largest ARB and RMSE values for the regression parameters and the variances while PX-GS shows the largest ARB and RMSE values for correlations. The values of ARB and RMSE become smaller and the CP are improving with increasing sample sizes for all algorithms.

Figures 1-3 display the autocorrelation function (ACF) plots. It is evident that for regression parameters β_1 and β_2, with increasing sample sizes, PX-GSM manifests supremacy compared to PX-MH and PX-GS. For correlations, Figures 1 and 2 show that for sample sizes 500 and 2000, PX-GS has a little edge compared to PX-GSM, while Figure 3 shows that for sample size 4000, PX-GSM and PX-GS produce similar ACF values. For variances, PX-GS shows a slight advantage in comparison with PX-GSM for all sample sizes. In general, PX-MH displays slowest decreased ACF values among these three algorithms.

In summary, PX-GSM outperforms PX-MH and PX-GS in estimated parameters, and exhibits exceptional in the convergence of the regression parameters; PX-GS shows a slight advantage in the convergence of the correlations and variances in comparison with PX-GSM; PX-MH indicates weakness in estimation and convergence compared with PX-GS and PX-GSM.

Table 1: Averaged posterior means (Mean), standard deviations (s.d.), ARBs, RMSEs and CPs for sample size 500 based on 100 generated datasets under MAR with 60% missing values

Parameter	True	PX-MH				PX-GS				PX-GSM			
		Mean (s.d.)	ARB	RMSE	CP	Mean (s.d.)	ARB	RMSE	CP	Mean (s.d.)	ARB	RMSE	CP
$\beta_1/\sqrt{\sigma_{11}}$	1.0	1.45 (0.28)	0.45	0.55	61	1.37 (0.26)	0.38	0.47	73	1.34 (0.31)	0.35	0.49	91
$\beta_2/\sqrt{\sigma_{11}}$	2.0	3.02 (0.52)	0.51	1.17	43	2.85 (0.49)	0.43	1.00	57	2.80 (0.61)	0.41	1.04	84
ρ_{12}	0.5	0.08 (0.21)	0.85	0.48	41	0.07 (0.21)	0.87	0.48	42	0.11 (0.24)	0.79	0.47	68
ρ_{13}	0.5	0.07 (0.21)	0.86	0.49	40	0.06 (0.21)	0.88	0.49	32	0.09 (0.24)	0.81	0.48	66
ρ_{14}	0.5	0.09 (0.20)	0.82	0.47	42	0.07 (0.21)	0.85	0.48	35	0.11 (0.24)	0.79	0.47	63
ρ_{15}	0.5	0.09 (0.21)	0.83	0.47	47	0.07 (0.21)	0.85	0.48	43	0.11 (0.24)	0.79	0.47	65
ρ_{23}	0.5	0.11 (0.21)	0.79	0.46	45	0.08 (0.21)	0.84	0.47	42	0.12 (0.24)	0.76	0.46	72
ρ_{24}	0.5	0.11 (0.22)	0.80	0.46	49	0.08 (0.21)	0.84	0.48	35	0.11 (0.24)	0.79	0.47	64
ρ_{25}	0.5	0.11 (0.21)	0.79	0.46	52	0.06 (0.21)	0.88	0.49	37	0.08 (0.24)	0.84	0.49	65
ρ_{34}	0.5	0.09 (0.21)	0.84	0.48	46	0.07 (0.21)	0.86	0.48	37	0.10 (0.24)	0.81	0.48	61
ρ_{35}	0.5	0.10 (0.20)	0.81	0.46	43	0.07 (0.21)	0.87	0.49	38	0.09 (0.24)	0.82	0.49	63
ρ_{45}	0.5	0.11 (0.21)	0.79	0.46	47	0.07 (0.21)	0.85	0.48	40	0.09 (0.24)	0.82	0.49	58
σ_{22}/σ_{11}	1.0	1.40 (0.64)	0.46	0.81	97	1.13 (0.50)	0.20	0.55	100	1.17 (0.64)	0.26	0.71	100
σ_{33}/σ_{11}	1.0	1.36 (0.60)	0.44	0.78	97	1.14 (0.51)	0.21	0.56	100	1.16 (0.63)	0.26	0.70	99
σ_{44}/σ_{11}	1.0	1.39 (0.62)	0.46	0.81	98	1.14 (0.50)	0.21	0.55	100	1.15 (0.62)	0.25	0.68	100
σ_{55}/σ_{11}	1.0	1.45 (0.64)	0.54	0.87	93	1.15 (0.50)	0.22	0.57	100	1.13 (0.60)	0.26	0.67	100

Table 2: Averaged posterior means (Mean), standard deviations (s.d.), ARBs, RMSEs and CPs for sample size 2000 based on 100 generated datasets under MAR with 60% missing values

Parameter	True	PX-MH				PX-GS				PX-GSM			
		Mean (s.d.)	ARB	RMSE	CP	Mean (s.d.)	ARB	RMSE	CP	Mean (s.d.)	ARB	RMSE	CP
$\beta_1/\sqrt{\sigma_{11}}$	1.0	1.20(0.13)	0.23	0.28	59	1.19(0.17)	0.21	0.28	83	1.09(0.21)	0.15	0.27	98
$\beta_2/\sqrt{\sigma_{11}}$	2.0	2.45(0.25)	0.24	0.56	53	2.44(0.34)	0.23	0.59	81	2.22(0.41)	0.15	0.53	99
ρ_{12}	0.5	0.28(0.13)	0.47	0.28	54	0.25(0.16)	0.52	0.31	66	0.37(0.17)	0.34	0.25	91
ρ_{13}	0.5	0.27(0.13)	0.48	0.28	55	0.25(0.16)	0.51	0.31	67	0.35(0.17)	0.36	0.26	90
ρ_{14}	0.5	0.28(0.14)	0.48	0.28	56	0.26(0.16)	0.49	0.30	74	0.36(0.17)	0.33	0.25	91
ρ_{15}	0.5	0.28(0.13)	0.47	0.28	57	0.27(0.16)	0.47	0.30	70	0.34(0.17)	0.36	0.26	88
ρ_{23}	0.5	0.28(0.13)	0.47	0.29	56	0.23(0.16)	0.54	0.32	65	0.33(0.18)	0.38	0.27	91
ρ_{24}	0.5	0.28(0.13)	0.48	0.28	48	0.24(0.16)	0.54	0.32	65	0.33(0.18)	0.37	0.27	90
ρ_{25}	0.5	0.29(0.14)	0.47	0.28	54	0.25(0.16)	0.51	0.31	70	0.31(0.18)	0.40	0.28	87
ρ_{34}	0.5	0.27(0.14)	0.49	0.29	57	0.24(0.16)	0.52	0.32	66	0.31(0.18)	0.41	0.29	90
ρ_{35}	0.5	0.27(0.14)	0.48	0.29	56	0.24(0.16)	0.51	0.31	69	0.29(0.18)	0.42	0.29	84
ρ_{45}	0.5	0.29(0.13)	0.46	0.28	57	0.25(0.16)	0.51	0.31	69	0.29(0.18)	0.44	0.30	85
σ_{22}/σ_{11}	1.0	1.15(0.33)	0.33	0.50	88	1.05(0.41)	0.17	0.46	98	1.05(0.47)	0.20	0.53	100
σ_{33}/σ_{11}	1.0	1.13(0.34)	0.36	0.52	83	1.06(0.42)	0.20	0.48	100	0.99(0.46)	0.23	0.53	98
σ_{44}/σ_{11}	1.0	1.16(0.33)	0.39	0.54	76	1.07(0.42)	0.24	0.50	99	0.97(0.45)	0.26	0.55	94
σ_{55}/σ_{11}	1.0	1.16(0.33)	0.36	0.52	74	1.08(0.42)	0.21	0.48	100	0.90(0.42)	0.23	0.50	96

Table 3: Averaged posterior means (Mean), standard deviations (s.d.), ARBs, RMSEs and CPs for sample size 4000 based on 100 generated datasets under MAR with 60% missing values

Parameter	True	PX-MH				PX-GS				PX-GSM			
		Mean (s.d.)	ARB	RMSE	CP	Mean (s.d.)	ARB	RMSE	CP	Mean (s.d.)	ARB	RMSE	CP
$\beta_1/\sqrt{\sigma_{11}}$	1.0	1.13(0.09)	0.16	0.19	54	1.12(0.14)	0.15	0.21	94	0.97(0.16)	0.08	0.19	100
$\beta_2/\sqrt{\sigma_{11}}$	2.0	2.29(0.16)	0.17	0.40	49	2.26(0.28)	0.15	0.43	94	1.97(0.31)	0.08	0.37	99
ρ_{12}	0.5	0.34(0.09)	0.37	0.21	46	0.35(0.13)	0.32	0.21	81	0.50(0.12)	0.18	0.15	95
ρ_{13}	0.5	0.36(0.09)	0.33	0.20	52	0.36(0.13)	0.30	0.20	90	0.48(0.12)	0.16	0.15	93
ρ_{14}	0.5	0.35(0.09)	0.34	0.20	51	0.36(0.13)	0.30	0.21	84	0.46(0.13)	0.20	0.17	96
ρ_{15}	0.5	0.34(0.10)	0.36	0.21	51	0.35(0.13)	0.33	0.22	76	0.42(0.13)	0.25	0.19	94
ρ_{23}	0.5	0.36(0.10)	0.33	0.20	62	0.36(0.13)	0.31	0.21	86	0.46(0.13)	0.18	0.17	96
ρ_{24}	0.5	0.35(0.09)	0.35	0.21	51	0.35(0.13)	0.33	0.22	84	0.42(0.14)	0.24	0.19	96
ρ_{25}	0.5	0.35(0.09)	0.34	0.20	53	0.34(0.13)	0.34	0.22	82	0.39(0.14)	0.26	0.20	94
ρ_{34}	0.5	0.35(0.09)	0.34	0.21	63	0.34(0.13)	0.33	0.22	81	0.40(0.14)	0.26	0.20	92
ρ_{35}	0.5	0.34(0.09)	0.37	0.22	55	0.34(0.13)	0.34	0.22	84	0.36(0.14)	0.31	0.23	86
ρ_{45}	0.5	0.35(0.09)	0.34	0.21	57	0.35(0.13)	0.33	0.22	81	0.34(0.14)	0.35	0.24	83
σ_{22}/σ_{11}	1.0	1.06(0.21)	0.30	0.39	65	1.04(0.37)	0.21	0.45	96	0.98(0.37)	0.21	0.45	95
σ_{33}/σ_{11}	1.0	1.09(0.22)	0.32	0.41	61	1.05(0.37)	0.22	0.44	99	0.90(0.35)	0.21	0.43	94
σ_{44}/σ_{11}	1.0	1.10(0.20)	0.34	0.42	55	1.06(0.38)	0.21	0.46	99	0.82(0.32)	0.25	0.45	87
σ_{55}/σ_{11}	1.0	1.11(0.22)	0.33	0.42	62	1.07(0.38)	0.21	0.45	98	0.74(0.31)	0.31	0.47	80

Real Data Application

The Mental Health Client-Level Data collects information on individuals receiving mental health by the Substance Abuse and Mental Health Services Administration in collaboration with State Mental Health Agencies. This data includes demographic, clinical, and National Outcome Measures information, which is valuable to behavioral health service providers, researchers, public, federal, state, and local governments (Samhsa, 2022).

We select clients who received mental health services within institutions under the justice system (e.g., jails, prisons, juvenile detention centers) in Michigan during the 2022 reporting period. In this analysis, the first mental health diagnosis is used as the response variable, and two category-specific covariates with intercepts are considered: Age (midpoint of age groups) and Gender (Female = 0, Male = 1). The dataset contains a total of 4383 observations, of which 2350 cases (53.6%) have missing values for the response variable.

The original mental health diagnosis consists of 13 distinct disorders, which we recoded into six main categories: Trauma and stressor-related disorders (Category 1), Anxiety disorders (Category 2), Bipolar disorders (Category 3), Depressive disorders (Category 4), Schizophrenia or other psychotic disorders (Category 5), and all other conditions grouped into other disorders/conditions (the reference group). We consider relatively weak priors as

$\beta \sim \mathcal{N}(0,100\mathrm{I}_{15})$ and $\Sigma \sim \mathcal{W}^{-1}(\mathrm{V} = (m-k-1)\mathrm{I}_5, m = 15)$ and run PX-MH, PX-GS and PX-GSM, each with 120,000 iterations and 20,000 burn-in.

Table 4 shows that age is not significantly related to Trauma and stressor-related disorders (β_{11}), but has significantly positive relationship with Anxiety disorders (β_{21} based on PX-MH), Bipolar disorders (β_{31} based on PX-MH and PX-GS), Depressive disorders (β_{41}), Schizophrenia or other psychotic disorders (β_{51}), suggesting older individuals tend to have symptoms of those disorders than younger ones. It is noticeable that the effect of gender is not significant across all disorder categories, but Anxiety disorders (β_{22}) are negatively associated with gender, showing that females are more likely to have Anxiety disorders compared to males.

However, in this data, the estimated variances (between 0.20 and 0.60) for correlations are in general large. This large variance phenomenon in correlations has been realized and commented on by Keane (1992) and Zhang (2025b) due to the well-known model identification issue.

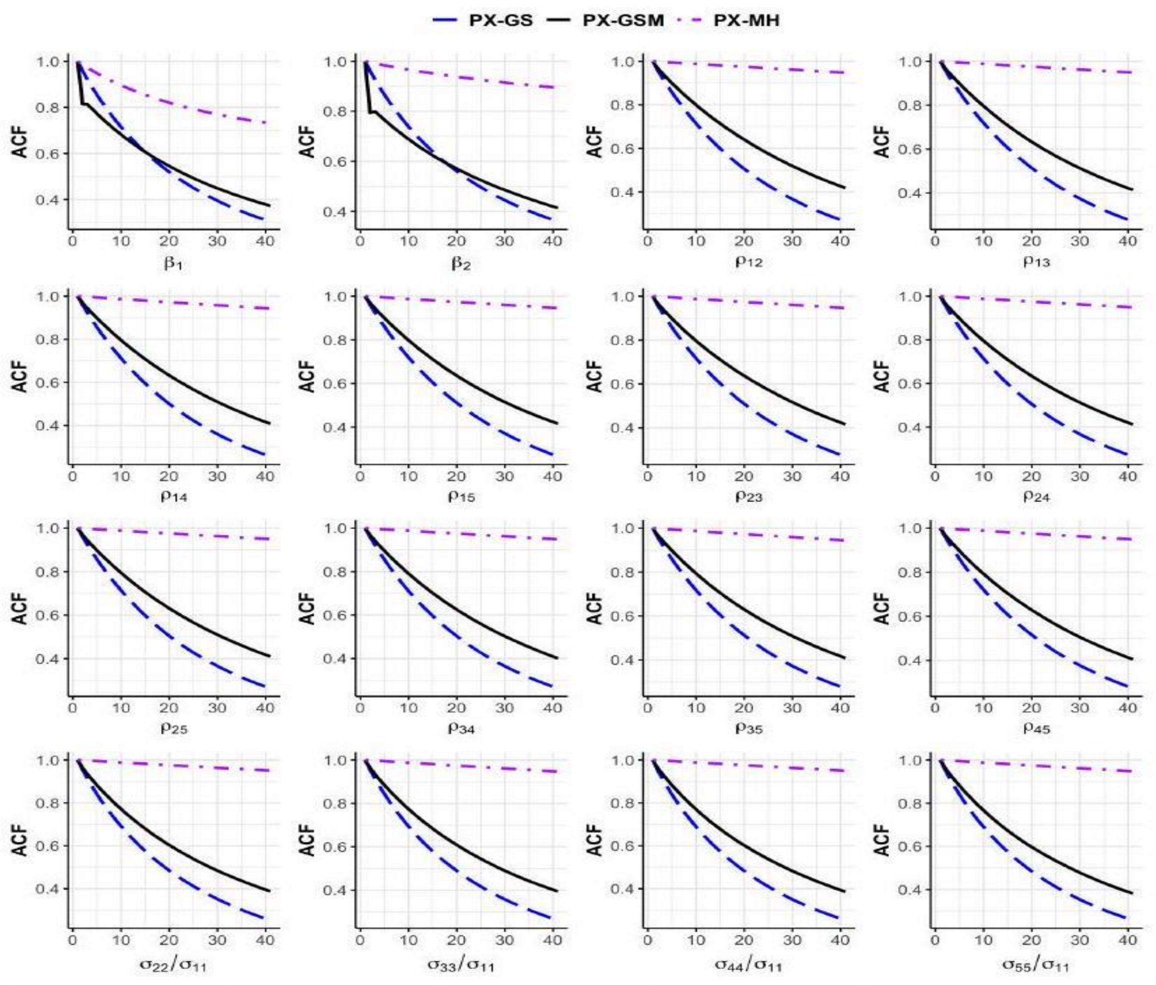
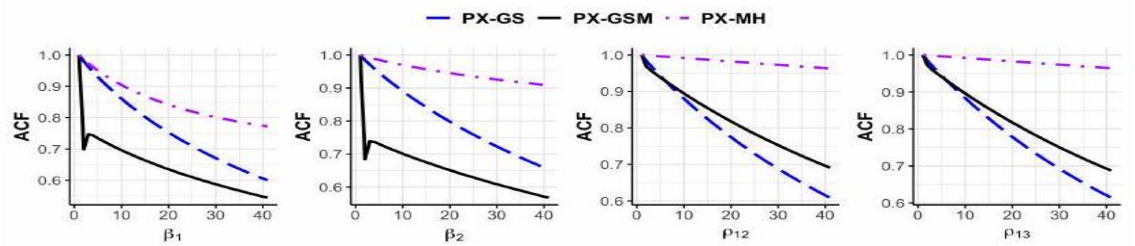


Fig. 1: The ACF plots of the regression parameters, correlations and variances for sample size 500 based on 100 generated datasets under MAR with 60% missing values



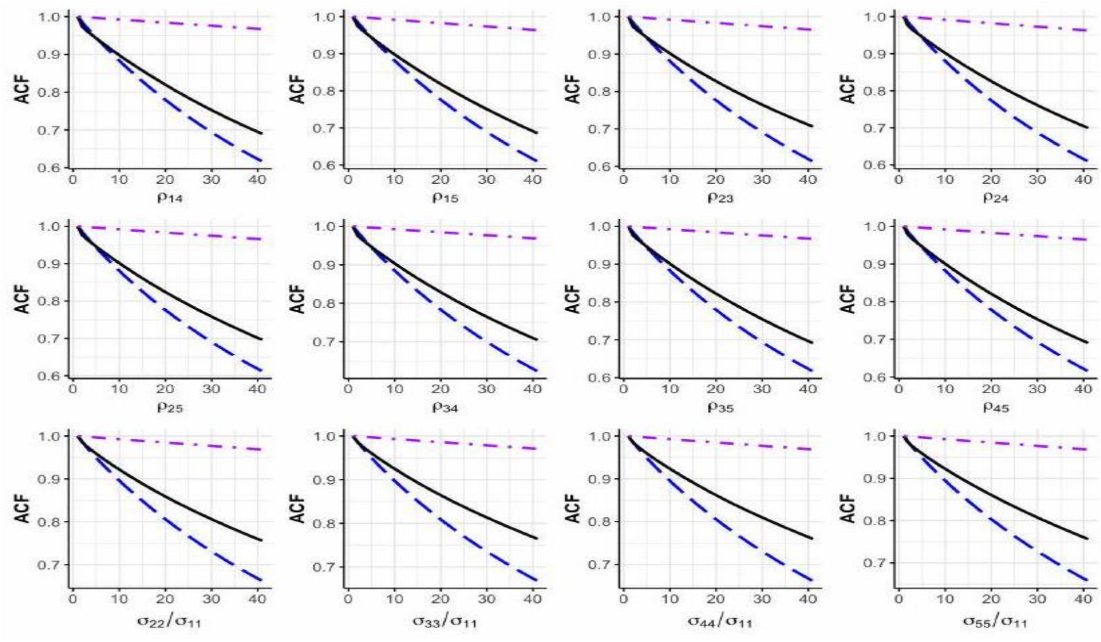


Fig. 2: The ACF plots of the regression parameters, correlations and variances for sample size 2000 based on 100 generated datasets under MAR with 60% missing values

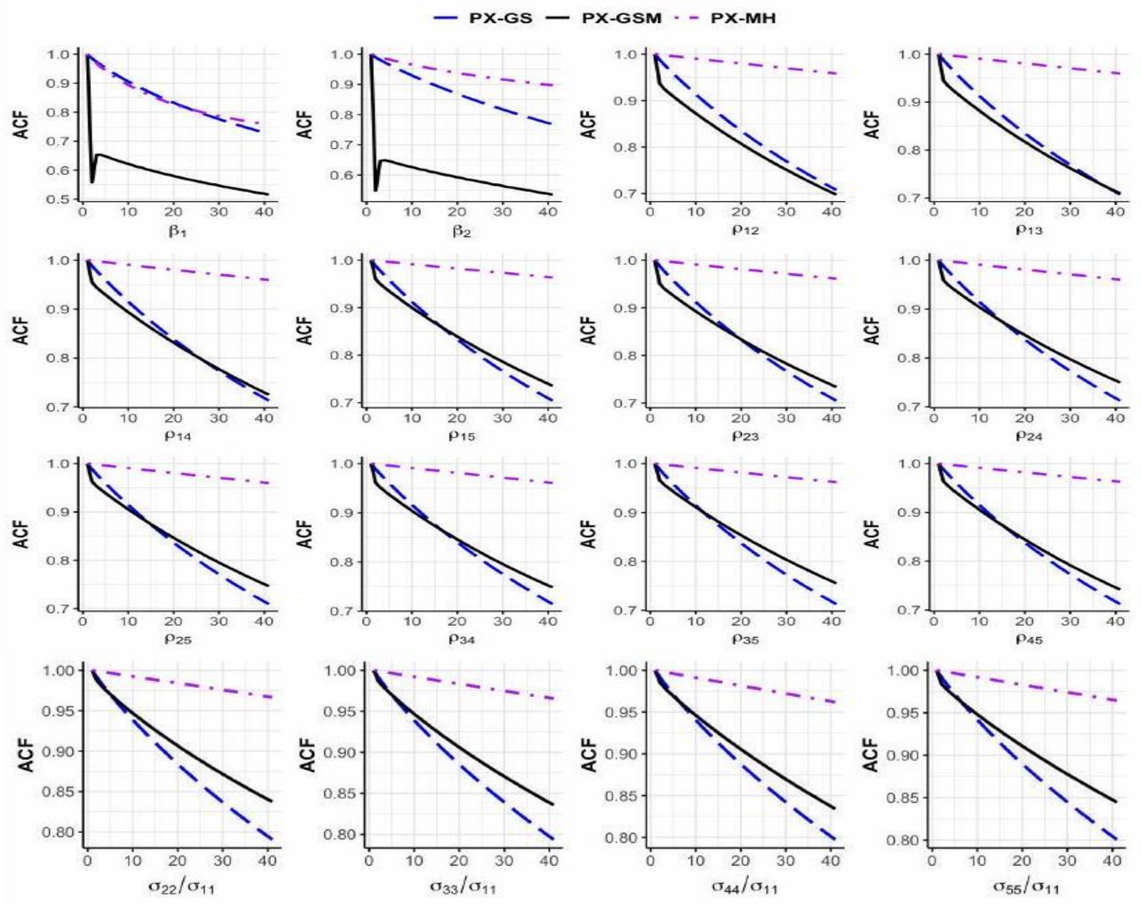


Fig. 3: The ACF plots of the regression parameters, correlations and variances for sample size 4000 based on 100 generated datasets under MAR with 60% missing values

Table 4: Posterior means, standard deviations (s.d) and 95% credible intervals (CI) for regression parameters using PX-MH, PX-GS, PX-GSM

Categories	Parameter	PX-MH		PX-GS		PX-GSM	
		Mean(s.d.)	CI	Mean(s.d.)	CI	Mean(s.d.)	CI
Trauma and stressor-related disorders	β_{11} (Age)	0.00(0.00)	(-0.01, 0.01)	-0.00(0.00)	(-0.01, 0.01)	-0.01(0.00)	(-0.01, 0.00)
	β_{12} (Gender)	0.03(0.08)	(-0.14, 0.19)	0.04(0.08)	(-0.12, 0.20)	0.07(0.07)	(-0.07, 0.20)
Anxiety disorders	β_{21} (Age)	0.01(0.01)	(0.00, 0.03)	0.01(0.00)	(-0.00, 0.02)	0.00(0.00)	(-0.00, 0.01)
	β_{22} (Gender)	-0.40(0.20)	(-0.84, -0.05)	-0.23(0.12)	(-0.47, -0.02)	-0.20(0.13)	(-0.48, -0.01)
Bipolar disorders	$\beta_{31}(Age)$	0.02(0.01)	(0.00, 0.03)	0.01(0.00)	(0.00, 0.02)	0.00(0.00)	(-0.00, 0.01)
	β_{32} (Gender)	-0.04(0.14)	(-0.32, 0.25)	-0.03(0.09)	(-0.22, 0.15)	-0.01(0.05)	(-0.14, 0.09)
Depressive disorders	β_{41} (Age)	0.02(0.01)	(0.01, 0.03)	0.01(0.00)	(0.01, 0.02)	0.00(0.00)	(0.00, 0.01)
	β_{42} (Gender)	-0.17(0.13)	(-0.43, 0.07)	-0.10(0.09)	(-0.28, 0.06)	-0.03(0.03)	(-0.10, 0.03)
Schizophrenia or other psychotic disorders	β_{51} (Age)	0.03(0.01)	(0.02, 0.05)	0.02(0.01)	(0.01, 0.03)	0.00(0.00)	(0.00, 0.01)
	β_{52} (Gender)	0.28(0.17)	(-0.04, 0.63)	0.17(0.11)	(-0.04, 0.40)	0.05(0.04)	(-0.02, 0.13)

Results and Conclusion

In this study, we develop two PX-DA methods - PX-GS and PX-GSM to analyze nominal data with missing values based on the constructed nonidentifiable MNP model. Simulation studies illustrate that the proposed methods outperform PX-MH based on the identifiable MNP model in estimation and convergence properties and can handle missing values with substantial missingness. PX-GSM produces the most accurate estimated values among the compared algorithms. In particular, it exhibits superior ACF plots for regression parameters. Nevertheless, PX-GS displays a slight edge in the convergence of correlations and variances. Overall, PX-MH indicates slow convergence and inferior in estimation compared with PX-GS and PX-GSM. Our real data application shows that model identification exhibited by estimated large variances in correlations is still a challenging issue in the MNP model analysis.

Our investigation assumes MCAR and MAR mechanisms. We generate missing values for MCAR based on the uniform distribution and for MAR based on the pattern of the covariates generated by Bernoulli distributions. However, it is worthwhile to consider other possible methods to construct covariates, such as based on normal distribution besides Bernoulli distribution or considering more than one type of distribution, to be able to conduct further investigation regarding missingness mechanisms. In the near future, extending our proposed methods to the NMAR mechanism is also worth our effort. Given the promising performance of PX-GS and PX-GSM, we are motivated to further extend those methods to analyze multivariate nominal measures with missing values.

Acknowledgment

The authors gratefully acknowledge support from the National Institutes of Health (NIH). Also, the authors would like to thank the Editors and reviewers for their valuable comments and suggestions that improve the quality of the paper significantly.

Funding Information

The project is supported by NIH grant R15GM151700.

Author's Contributions

Suwash Silwal: Conducted the simulation studies, performed the analysis of real data, and drafted the manuscript.

Ziao Zhang: Contributed to the development of the methodology, designed the simulation study framework, and participated in the writing and revision of the manuscript.

Ethics

This article is original and contains unpublished materials. The corresponding author has read and approved the manuscript, and no ethical issues are involved.

References

Albert, J. H., & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, *88*(422), 669–679. https://doi.org/10.1080/01621459.1993.10476321

Ben-Akiva, M., Lerman, S. R., & Uncles, M. D. (2018). *Discrete Choice Analysis: Theory and Application to Travel Demand. 38*. https://doi.org/10.2307/2582065.

Berndt, E. R., Hall, B. H., Hall, R. E., & Hausman, J. A. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, *3*, 653–666.

Bunch, D. S. (1991). Estimability in the multinomial probit model. *Transportation Research Part B: Methodological*, *25*(1), 1–12. https://doi.org/10.1016/0191-2615(91)90009-8

Bunch, D. S., & Kitamura, R. (1991). *Probit model estimation revisited: Trinomial models of household car ownership*.

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(63), 1–67.

Buuren, S. van. (2018). *Flexible Imputation of Missing Data, Second Edition*. https://doi.org/10.1201/9780429492259

Carpenter, J. R., & Smuk, M. (2021). Missing data: A statistical framework for practice. *Biometrical Journal*, *63*(5), 915–947. https://doi.org/10.1002/bimj.202000196

Clark, C. E. (1961). The Greatest of a Finite Set of Random Variables. *Operations Research*, *9*(2), 145–162. https://doi.org/10.1287/opre.9.2.145

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, *85*(410), 398–409. https://doi.org/10.1080/01621459.1990.10476213

Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of constraint probabilities. *Computing Science and Statistics: Proceedings of the Symposium on the Interface*, 613672.

Geweke, J., Keane, M. P., & Runkle, D. E. (1994). Recursively simulating multinomial multiperiod probit probabilities MPRA Paper. *80, 1*, 1–46.

Gilks, W. R., & Richardson, S., D.J. Spiegelhalter, (1995). *Markov Chain Monte Carlo in Practice*. https://doi.org/10.1201/b14835

Hajivassiliou, V. A., & McFadden, D. L. (1998). The Method of Simulated Scores for the Estimation of LDV Models. *Econometrica*, *66*(4), 863. https://doi.org/10.2307/2999576

Hajivassiliou, V., McFadden, D., & Ruud, P. (1996). Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results. *Journal of Econometrics*, *72*(1–2), 85–134. https://doi.org/10.1016/0304-4076(94)01716-6

Harel, O., & Zhou, X. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine*, *26*(16), 3057–3077. https://doi.org/10.1002/sim.2787

Hausman, J. A., & Wise, D. A. (1978). A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica*, *46*(2), 403. https://doi.org/10.2307/1913909

Horowitz, J. L., Sparmann, J. M., & Daganzo, C. F. (1982). An Investigation of the Accuracy of the Clark Approximation for the Multinomial Probit Model. *Transportation Science*, *16*(3), 382–401. https://doi.org/10.1287/trsc.16.3.382

Imai, K., & van Dyk, D. A. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, *124*(2), 311–334. https://doi.org/10.1016/j.jeconom.2004.02.002

Kamakura, W. A., & Russell, G. J. (1989). A Probabilistic Choice Model for Market Segmentation and Elasticity Structure. *Journal of Marketing Research*, *26*(4), 379. https://doi.org/10.2307/3172759

Keane, M. P. (1992). A Note on Identification in the Multinomial Probit Model. *Journal of Business & Economic Statistics*, *10*(2), 193–200. https://doi.org/10.1080/07350015.1992.10509898

Keane, M. P. (1994). A Computationally Practical Simulation Estimator for Panel Data. *Econometrica*, *62*(1), 95. https://doi.org/10.2307/2951477

Lawrence, E., Bingham, D., Liu, C., & Nair, V. N. (2008). Bayesian Inference for Multivariate Ordinal Data Using Parameter Expansion. *Technometrics*, *50*(2), 182–191. https://doi.org/10.1198/004017008000000064

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*.

Liu, J. S., & Wu, Y. N. (1999). Parameter Expansion for Data Augmentation. *Journal of the American Statistical Association*, *94*(448), 1264–1274. https://doi.org/10.1080/01621459.1999.10473879

MacEachern, S. N. (2007). Comment on article by Jain and Neal. *Bayesian Analysis*, *2*(3), 483–494.

McCulloch, R. E., Polson, N. G., & Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, *99*(1), 173–193. https://doi.org/10.1016/s0304-4076(00)00034-8

McCulloch, R., & Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, *64*(1–2), 207–240. https://doi.org/10.1016/0304-4076(94)90064-7

McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica*, *57*(5), 995. https://doi.org/10.2307/1913621

Meng, X.-L., & Dyk, DA van. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, *86*(2), 301–320. https://doi.org/10.1093/biomet/86.2.301

Nobile, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statistics and Computing*, *8*(3), 229–242. . https://doi.org/10.1023/a:1008905311214

Nobile, A. (2000). Comment: Bayesian multinomial probit models with a normalization constraint. *Journal of Econometrics*, *99*(2), 335–345. https://doi.org/10.1016/s0304-4076(00)00035-x

Pakes, A., & Pollard, D. (1989). Simulation and the Asymptotics of Optimization Estimators. *Econometrica*, *57*(5), 1027. https://doi.org/10.2307/1913622

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Samhsa. (2022). *Mental Health Client-Level Data (MH-CLD)*.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. *72*.

Tanner, M. A., & Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540. https://doi.org/10.1080/01621459.1987.10478458

Train, K. E. (2009). *Discrete Choice Methods with Simulation*.

Wijesuriya, R., Moreno-Betancur, M., Carlin, J. B., White, I. R., Quartagno, M., & Lee, K. J. (2025). Multiple Imputation for Longitudinal Data: A Tutorial. *Statistics in Medicine*, *44*(3–4), e10274. https://doi.org/10.1002/sim.10274

Yucel, R. M. (2011). State of the Multiple Imputation Software. *Journal of Statistical Software*, *45*(1), 1–7.

Zhang, X. (2020). Parameter-expanded data augmentation for analyzing correlated binary data using multivariate probit models. *Statistics in Medicine*, *39*(25), 3637–3652. https://doi.org/10.1002/sim.8685

Zhang, X. (2022). Bayesian Analysis of Longitudinal Ordinal Data Using Non-Identifiable Multivariate Probit Models. *Journal of Mathematics and Statistics*, *18*(1), 163–175. https://doi.org/10.3844/jmssp.2022.163.175

Zhang, X. (2025). Bayesian Analysis of Longitudinal Ordinal Data with Missing Values Using Multivariate Probit Models. *Journal of Statistics Applications & Probability*, *14*(3), 337–352. https://doi.org/10.18576/jsap/140302

Zhang, X. (2026). Parameter-expanded data augmentation for analyzing multinomial probit models. *Communications in Statistics - Theory and Methods*, *55*(9), 2724–2737. https://doi.org/10.1080/03610926.2025.2559116

Zhang, X., Boscardin, W. J., & Belin, T. R. (2008). Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. *Computational Statistics & Data Analysis*, *52*(7), 3697–3708. https://doi.org/10.1016/j.csda.2007.12.012