Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Performances of Collapsed Gibbs Sampling, Posterior Mode Estimation and Joint Maximum Likelihood Estimation on Diagnostic Classification Models for Boundary Problems

J. Li[0009-0004-3232-2685]†, Z. Lu[0000-0001-9482-1368] * †, M. Madison[0000-0002-2944-7442]†, and W. Shen[0000-0003-3483-0451]†

† Quantitative Methodologies, University of Georgia, Athens, 30602, GA, USA
*Corresponding author. Email: zlu@uga.edu

Abstract

Diagnostic Classification Models (DCMs) assess latent cognitive attributes but often suffer from boundary problems, where slipping and guessing parameters converge to 0 or 1, reducing classification accuracy. This study compares Collapsed Gibbs Sampling (CGS), Posterior Mode Estimation (PME), and Joint Maximum Likelihood Estimation (JMLE) in addressing this issue via Monte Carlo simulation. By varying sample size, item quality, dimensionality, and test length, we examine classification accuracy and boundary problem rates. Results show CGS and JMLE generally yield higher accuracy, while PME consistently maintains the lowest boundary problem occurrence rate. PME is generally computationally efficient in simpler settings, CGS offers stable performance, and JMLE can be competitive under specific conditions.

Keywords: diagnostic classification models, collapsed Gibbs sampling, posterior mode estimation, joint maximum likelihood estimation, boundary problem, Bayesian estimation

1. Introduction

Diagnostic classification models (DCMs), also known as cognitive diagnostic models (CDMs), have gained increasing attention in fields such as educational assessment and psychological evaluation (Chen, 2017). Unlike traditional models like the Rasch model or item response theory (IRT), which assume continuous and unidimensional latent traits, DCMs identify discrete, multidimensional latent skills or attributes required to answer test items correctly (Madison & Bradshaw, 2018; DeCarlo, 2011). Consequently, DCMs yield students' attribute mastery profiles (Oka & Okada, 2022).

In addition to diagnostic profiles, DCMs estimate item and structural parameters that describe the probability of correct responses and the distribution of attribute mastery in the population (Yamaguchi & Templin, 2022). However, these parameter estimates often fall on the boundaries (e.g., 0 or 1), even when the true values lie in the interior of the parameter space (DeCarlo, 2011). These boundary problems are particularly common when the sample size is

small, the number of attributes is large, or the Q-matrix is suboptimal (Clogg & Eliason, 1987; Yamaguchi & Templin, 2022).

To address this, Posterior Mode Estimation (PME) incorporates informative priors to regularize parameter estimates and reduce overfitting (Galindo-Garre & Vermunt, 2006). However, PME still estimates item and structural parameters, which may reduce efficiency (Yamaguchi & Templin, 2022). Collapsed Gibbs Sampling (CGS), in contrast, avoids estimating these parameters by marginalizing them out, focusing instead on directly estimating latent attributes (Porteous et al., 2008).

Despite methodological advancements, limited research has compared the performance of JMLE, PME, and CGS under consistent conditions. This study conducts simulation experiments to evaluate the estimation accuracy, boundary problem occurrence, and computational efficiency of the three methods across varying data conditions. R, Python, and Mplus are used for data generation and analysis.

2. The model and Boundary Problem

This study employs the Deterministic Inputs, Noisy "And" (DINA) model, a widely used framework within DCMs. The DINA model defines the probability of a correct response as a function of latent attribute mastery, specified by a binary Q matrix, and item properties. Suppose N denotes examinee sample size and K denotes the total number of latent attributes that are measured by a test. In the Q matrix, each element q_{ik} indicates whether item i measures attribute *k*. If attribute *k* is measured by item *i,* q_{ik} = 1; otherwise, q_{ik} = 0. Let $\alpha_e = (\alpha_{e1}, \alpha_{e2}, \dots, \alpha_{eK})$ denote the eth examinee's mastery profile, with $\alpha_{ek} = 1$ if examinee e has mastered attribute k, and 0 otherwise, and i = 1, . . . N. In a DINA model, the probability of providing a correct response X to item *i* for each examinee *e* was determined by:

$$P(X_{ei} = 1|\alpha_e, s_i, g_i) = (1 - s_i)^{\eta_{ei}} g_i^{1-\eta_{ei}} \quad (1)$$

where s_i is the slipping parameter, g_i is the guessing probability parameter, and

$$\eta_{ei} = \prod_{k=1}^{K} (a_{ei}{}^{q_{ik}}) \quad (2)$$

is an indicator of whether the examinee e has mastered all required attributes for the item i to answer it correctly (the "deterministic" feature of the model).

The DINA model assumes a conjunctive relationship between attributes, meaning that an examinee must master all the required attributes of an item to have a higher probability of answering it correctly. This strict assumption makes the model particularly sensitive to data sparsity and response inconsistencies, often leading to extreme parameter estimates when using standard estimation methods. Given an examinee's attribute profile, the probability of correctly answering an item depends on the item's slipping parameter (s_i) and guessing probability parameter (g_i). These parameters introduce stochasticity into the deterministic structure of the model, allowing for response patterns that do not strictly adhere to the mastery assumptions. However, in practice, these parameters are highly susceptible to boundary problems—where estimates converge to extreme values of 0 or 1 due to sparse or unbalanced response patterns. For items that are frequently answered correctly by examinees with only partial attribute mastery, the estimation process tends to drive guessing probabilities toward 1, even when the true guessing rate should be moderate. Similarly, items that are rarely answered correctly can result in slipping probabilities approaching 1, implying that even fully knowledgeable examinees are highly prone to incorrect responses.

3. Estimation Methods and Algorithms

3. 1 Joint Maximum Likelihood Estimation (JMLE)

Joint Maximum Likelihood Estimation (JMLE) is a traditional estimation approach that simultaneously estimates item parameters and latent attribute profiles by maximizing the likelihood function of the observed responses. Generally, its estimation process can be implemented using iterative algorithms such as the Newton-Raphson algorithm and the Expectation-Maximization (EM) algorithm.

The Newton-Raphson algorithm updates item parameters (slipping and guessing) directly by utilizing the gradient (first derivative) and Hessian matrix (second derivative) of the log-likelihood function. This method offers faster convergence under well-behaved likelihood surfaces, but it requires careful numerical implementation due to the risk of instability, especially when parameter estimates are near the boundaries of the parameter space (i.e., close to 0 or 1). The update formula for a generic parameter θ is given by:

$$\theta^{(t+1)} = \theta^{(t)} - \left[\frac{\partial^2 logL(\theta)}{\partial \theta^2}\right]^{-1} \left[\frac{\partial logL(\theta)}{\partial \theta}\right] \quad (3)$$

where *log L(θ)* is the log-likelihood function of the observed data.

Although Newton-Raphson provides faster convergence in theory, in practice the EM algorithm is more commonly used in DCMs due to its numerical stability and the discrete nature of latent attribute profiles. In the EM algorithm, the estimation begins by initializing the item parameters and the latent attribute profiles for all examinees. The algorithm proceeds in two steps, an expectation step (E-step) and a maximizing step (M-step). E-step: The expected complete-data log-likelihood is computed by estimating the probabilities of each examinee's attribute profile given the current estimates of item parameters and the observed responses. This step involves calculating the probability of each latent class membership for each examinee. M-step: The item parameters are updated by maximizing the expected log-likelihood obtained from the E-step. In particular, the slipping and guessing parameters are updated for each item based on the expected classification of examinees, adjusting the probability of observing correct or incorrect responses conditional on attribute mastery. This EM cycle continues iteratively until the parameter estimates converge, typically defined by a sufficiently small change in the log-likelihood or parameter values between iterations.

3. 2 Posterior Mode Estimation (PME)

Posterior Mode Estimation (PME) is a Bayesian approach, defined as:

$$P(\theta|X) \propto P(X|\theta)\, P(\theta) \quad (4)$$

where P(θ|X) represents the posterior distribution, P(X|θ) is the likelihood function, and P(θ) is the prior distribution. In practical applications, PME identifies the mode of the posterior distribution by maximizing the log-posterior:

$$log\, P(\theta|X) = log\, P(X|\theta) + log\, P(\theta) \quad (5)$$

To constrain slipping and guessing probabilities to realistic values, PME employs Beta priors, such as:

$$P(s_i) \sim Beta(\alpha_s, \beta_s) \tag{6}$$

$$P(g_i) \sim Beta(\alpha_g, \beta_g) \tag{7}$$

where α_s, $\beta_s$and α_g, β_g are hyperparameters that shape the prior distribution. These priors prevent parameters from being overly influenced by sparse or unbalanced data. For instance, a Beta(2, 2) prior penalizes boundary values near 0 or 1, encouraging parameter estimates to remain within a reasonable range. By enforcing prior constraints, PME effectively ensures numerical stability and prevents overfitting to rare patterns in the data (Galindo-Garre & Vermunt, 2006), makes it particularly useful for stabilizing estimation in small-sample scenarios or when the Q-matrix is mis-specified.

PME is computationally efficient because it builds upon existing MLE algorithms. Specifically, algorithms such as EM or Newton-Raphson can be adapted to include the prior term (log P(θ)) in the objective function, ensuring smooth parameter updates. The PME implementation typically iterates through two main steps: (1) calculate the log-posterior for each parameter by combining the log-likelihood and the log-prior, and (2) use numerical optimization methods to maximize the posterior distribution. Compared to fully Bayesian methods like Markov Chain Monte Carlo (MCMC), PME avoids the computational burden of iterative sampling while still leveraging prior information (Schafer, 1997).

The practical benefits of PME extend beyond its computational efficiency. For example, PME effectively stabilizes parameter estimates in CDMs, particularly under challenging conditions such as sparse response data or complex Q-matrices with overlapping attributes. In these scenarios, MLE often fails due to insufficient observations for certain item-response patterns, resulting in slipping probabilities that approach 1 for difficult items or guessing probabilities that inflate to 1 for overly easy items. By incorporating priors, PME smooths these estimates, ensuring they reflect both theoretical plausibility and empirical data. As demonstrated by DeCarlo (2010), PME outperforms MLE in recovering item parameters under such conditions, yielding more accurate and interpretable estimates of diagnostic classification accuracy.

3. 3 Collapsed Gibbs Sampling (CGS)

While PME effectively addresses boundary problems by incorporating prior distributions into parameter estimation, CGS provides an efficient Bayesian approach for parameter estimation in DCMs, addressing boundary problems by avoiding direct estimation of item-specific parameters such as slipping and guessing. Unlike methods like MLE or PME, CGS focuses on the estimation of examinees' latent attributes while integrating over item parameters during the sampling process. This marginalization inherently reduces the risk of parameter estimates sticking to boundary values.

The key idea of CGS is to reduce the dimensionality of the parameter space by marginalizing item-specific parameters, thereby concentrating the sampling process on latent attribute profiles. The joint posterior distribution of the model is

$$P(\alpha, X, \Theta, \Pi) \propto P(X|\alpha, \Theta)\, P(\Theta) P(\alpha|\Pi) P(\Pi) \tag{8}$$

where α represents the latent attribute profiles, X denotes the observed responses, Θ includes item parameters (e.g., slipping and guessing), and Π represents structural parameters such as

the class mixing probabilities. In CGS, the item parameters Θ and structural parameters Π are integrated out to obtain a collapsed posterior distribution:

$$P(\alpha, X) = \iint P(\alpha, X, \Theta, \Pi) d\Theta d\Pi. \quad (9)$$

The derivation of the parameter marginalization in the model follows the approach outlined in Sato (2016) and Suyama and Sugiyama (2017). This marginalization eliminates the need to directly estimate Θ and Π, which are often prone to extreme boundary values in MLE or PME frameworks. This eliminates focuses estimation solely on latent attribute profiles.

However, the calculation of eq.(9) is not obvious and computationally difficult. Therefore, we use the Gibbs sampling algorithm (Geman & Geman, 1984) to generate random samples of parameters from their conditional distributions. CGS operates through iterative sampling, updating each examinee's latent attribute profile α_e conditioned on the observed responses and the current state of all other examinees' profiles. The conditional posterior for α_e is given by:

$$P(\alpha_e | X_e, \alpha_{-e}) \propto P(X_e | \alpha_e) P(\alpha_e | \alpha_{-e}) \quad (10)$$

where X_e represents the response vector for examinee e, and α_{-e} denotes the latent profiles of all other examinees. This decomposition allows the sampling of α_e based on its contribution to the likelihood $P(X_e|\alpha_e)$ and the prior distribution informed by the collective profile $P(\alpha_e|\alpha_{-e})$. The likelihood $P(X_e|\alpha_e)$ incorporates the influence of slipping and guessing probabilities on examinee responses:

$$P(X_e|\alpha_e) = \prod_{i=1}^{n} [P(X_{ei} = 1|\alpha_e, Q_i)^{X_{ei}} P(X_{ei} = 0|\alpha_e, Q_i)^{1-X_{ei}}] \quad (11)$$

where X_{ei} is the response to item i, and Q_i is the item's Q-matrix row. By marginalizing s_i and g_i, the likelihood avoids boundary issues associated with their direct estimation. The prior $P(\alpha_e|\alpha_{-e})$ reflects the probability of α_e given the population-level latent class distribution. This distribution is governed by mixing proportions Π, which are updated iteratively to ensure coherence with the sampled profiles.

Regarding the detailed algorithm, CGS approach builds upon the standard Gibbs sampling but modifies the process by collapsing over item parameters, reducing the dimensionality of the estimation problem. It begins by assigning initial values to latent attribute profiles, either randomly or based on prior knowledge. Given these profiles, class proportions are assigned Dirichlet priors to maintain a probabilistic structure in the latent classification process. Unlike PME, which explicitly estimates slipping and guessing probabilities, CGS conditions latent attribute updates on observed responses while integrating over the unknown item parameters. For each iteration, the sampling process first updates the latent attribute profile of each examinee based on the conditional posterior distribution, incorporating the responses and the current state of other examinees' profiles. Next, class proportions are updated using a Dirichlet posterior, ensuring a smooth and stable representation of latent class distributions. To ensure robust estimation, CGS employs a burn-in period, during which initial samples are discarded to allow the Markov Chain to converge to a stationary distribution. Once convergence is achieved, the posterior samples are aggregated to obtain the final estimates of latent attribute mastery.

4. Simulation Design

To evaluate the performance of JMLE, CGS and PME in mitigating boundary problems within the DINA model, a series of Monte Carlo simulations were conducted. The study systematically varied key experimental conditions, including sample size (N), number of attributes (K), number of items (I), and item quality, to assess how these factors influence classification accuracy and parameter stability.

First, dichotomous response data were generated based on the DINA model described in Eq. (1) for each simulated dataset under various conditions, including four key factors. First, item quality (slipping and guessing probabilities) was assigned. High-quality items were characterized by s_i = g_i = 0.1, ensuring strong discriminative power, whereas low-quality items had s_i = g_i = 0.3, increasing response variability and reducing classification precision. Second, sample size was examined at two levels (N = 100 and N = 1000) to assess the impact of data availability on estimation accuracy. Third, the number of attributes was set at either K = 4 or K = 8, with higher attribute dimensionality expected to exacerbate boundary issues due to increased model complexity. Fourth, the number of items varied between I = 20 and I = 40 to explore how test length influences estimation performance. These factors were fully crossed, yielding 16 experimental conditions, with multiple replications per condition to ensure statistical reliability.

After the response data were generated, latent attribute profiles were estimated using DCMs. Three estimation approaches were applied to each dataset. Estimation procedures recover the underlying true attribute profiles, allowing for direct comparison between estimated and true classifications. Classification accuracy was assessed using the Correct Classification Rate (CCR), defined as the proportion of examinees whose estimated attribute profiles perfectly matched their true profiles:

$$CCR = \frac{1}{N}\sum_{e=1}^{N} I(\hat{\alpha}_e = \alpha_e) \tag{12}$$

where I(·) is an indicator function that equals 1 when the estimated profile matches the true profile and 0 otherwise. To further investigate the occurrence of boundary problems, each experimental condition was monitored for instances in which estimated s or g values exceeded 0.9999 or fell below 0.0001. The boundary problem occurrence rate (BPOR) was computed as:

$$BPOR = \frac{1}{I}\sum_{e=1}^{N} I(s_i < 0.0001 \; or \; s_i > 0.9999 \; or \; g_i < 0.0001 \; or \; g_i > 0.9999) \tag{13}$$

where the numerator counts the number of items for which boundary estimates occurred, and the denominator represents the total number of items in the dataset. This metric was averaged across multiple replications to obtain an overall estimate of the frequency with which boundary issues arose for each estimation method.

5. Result and Conclusion

Simulation results show that JMLE and CGS generally yield higher correct classification rates (CCR) than PME under low-dimensional settings (e.g., K = 4), especially when item quality is high. For instance, with N = 1000 and I = 40, CGS and JMLE achieved CCRs above 0.90, while PME remained near 0.52.

In contrast, under high-dimensional conditions (e.g., K = 8), all methods performed poorly, though PME showed slightly more stable performance in some scenarios. Regarding boundary problems, PME consistently achieved the lowest boundary problem occurrence rate

(BPOR), benefiting from prior regularization. CGS avoided boundary issues entirely due to marginalization of item parameters. In terms of computational efficiency, PME was generally the fastest in simple conditions, while CGS demonstrated stable but moderate runtime. JMLE showed higher computational cost in complex scenarios but was occasionally competitive in favorable settings.

Overall, CGS offers a good balance between classification accuracy and computational stability, while PME provides robustness against boundary issues, especially in high-dimensional or sparse-data conditions.

Table 1. Correct Classification Rate (CCR) and Boundary Problem Occurrence Rate (BPOR) Across Simulation Conditions for JMLE, CGS and PME

	CCR			BPOR[a]	
Conditions	JMLE	CGS	PME	JMLE	PME
K = 4, N = 100, High-Quality[b], I = 20	0.83	0.82	0.48	0.31	0.28
I = 40	0.88	0.88	0.46	0.28	0.21
Low-Quality, I = 20	0.35	0.28	0.33	0.45	0.34
I = 40	0.51	0.46	0.34	0.44	0.36
N = 1000, High-Quality, I = 20	0.82	0.82	0.51	0.36	0.22
I = 40	0.91	0.92	0.52	0.23	0.17
Low-Quality, I = 20	0.40	0.35	0.38	0.38	0.25
I = 40	0.53	0.52	0.45	0.37	0.23
K = 8, N = 100, High-Quality, I = 20	0.28	0.29	0.35	0.46	0.37
I = 40	0.31	0.29	0.26	0.42	0.33
Low-Quality, I = 20	0.04	0.02	0.11	0.51	0.38
I = 40	0.09	0.08	0.15	0.52	0.31
N = 1000, High-Quality, I = 20	0.29	0.27	0.30	0.39	0.28
I = 40	0.32	0.33	0.37	0.42	0.35
Low-Quality, I = 20	0.04	0.05	0.06	0.53	0.39
I = 40	0.07	0.06	0.11	0.49	0.36

a. No BPOR for CGS estimation because the parameters have been integrated out.
b. "High-Quality" refers to high-quality items (s = g = 0.1); "Low-Quality" refers to low-quality items (s = g = 0.3).

Table 2. Computation Efficiency (CE) Accounted by Seconds Across Simulation Conditions for JMLE, CGS and PME

	CE[a]		
Conditions	JMLE	CGS	PME
K = 4, N = 100, High-Quality[b], I = 20	0.64	0.23	0.20
I = 40	0.58	0.19	0.24
Low-Quality, I = 20	3.03	3.70	3.68
I = 40	0.77	0.50	0.47
N = 1000, High-Quality, I = 20	0.59	0.21	0.42
I = 40	0.69	0.29	0.27
Low-Quality, I = 20	2.79	1.22	1.18
I = 40	1.01	0.75	0.89
K = 8, N = 100, High-Quality, I = 20	0.74	0.46	0.37

		I = 40	0.96	0.79	0.59
	Low-Quality,	I = 20	1.88	1.33	1.43
		I = 40	1.22	1.21	1.67
N = 1000,	High-Quality,	I = 20	3.52	3.28	3.15
		I = 40	2.59	3.25	3.14
	Low-Quality,	I = 20	14.10	15.08	16.02
		I = 40	45.33	50.53	51.32

[a] "CE" is defined as the average seconds spend (repeated 10 times).

Acknowledgement

Funding Statement: None

References

Chen, Y., Li, X., Liu, J., & Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis. Psychometrika, 82, 660-692.

Chung, M. (2019). A Gibbs sampling algorithm that estimates the Q-matrix for the DINA model. Journal of Mathematical Psychology, 93, 102275.

Clogg, C. C., & Eliason, S. R. (1987). Some common problems in log-linear analysis. Sociological Methods & Research, 16(1), 8-44.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. Applied Psychological Measurement, 35(1), 8-26.

Galindo Garre, F., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. Behaviormetrika, 33, 43-59.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(6), 721–741.

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. Psychometrika, 83, 963-990.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. Applied Psychological Measurement, 29(4), 262-277.

Oka, M., & Okada, K. (2021). Assessing the performance of diagnostic classification models in small sample contexts with different estimation methods. arXiv preprint arXiv:2104.10975.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008, August). Fast collapsed gibbs sampling for latent dirichlet allocation. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 569-577).

Sato, I. (2016). Bayesian Nonparametrics. Kodansha.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. CRC press.

Suyama, A., & Sugiyama, M. (2017). Introduction to Machine Learning by Bayesian Inference. Kodansha.

Yamaguchi, K., & Templin, J. (2022). Direct estimation of diagnostic classification model attribute mastery profiles via a collapsed Gibbs sampling algorithm. Psychometrika, 87(4), 1390-1421.