Scalable Variational Inference for Probabilistic Boolean Matrix Factorization with Unknown Latent Dimension

Russell Zhang Kunes* Mingzhang Yin† Yuqi Gu*

*Department of Statistics, Columbia University
†Warrington College of Business, University of Florida

Abstract

Boolean matrix factorization (BMF) provides an interpretable representation of binary matrix through a nonlinear Boolean product of two lower-dimensional binary factor matrices and arises in various applications. We study a probabilistic BMF model, the Deterministic Input, Noisy output And gate (DINA) model, in the practically important and challenging setting where the latent dimension is unknown. Existing methods either assume the latent dimension is known or rely on latent-class representations whose complexity grows exponentially with the latent dimension. We propose a scalable mean-field variational inference method that employs a cumulative shrinkage process prior for latent-dimension selection with closed-form coordinate ascent updates, allowing joint estimation of the two binary factor matrices and the unknown latent dimension. The resulting algorithm has per-iteration cost linear in a user-specified upper bound on the latent dimension, making it suitable for high-dimensional, large-scale problems. Theoretically, we establish sample-complexity guarantees for both a full-parametrization variational estimator and the proposed mean-field algorithm, and we prove a minimax lower bound that matches the upper bound up to logarithmic factors in the small-noise regime. Extensive simulation studies illustrate accurate recovery and favorable computational scalability. Applications to single-cell chromatin accessibility data and educational assessment data illustrate cross-domain applicability and uncover interpretable findings.

Keywords: Boolean matrix factorization; Binary latent feature models; Variational inference; Latent-dimension selection; Cumulative shrinkage process; DINA model.

1 Introduction

Boolean matrix factorization (BMF) seeks to represent a binary data matrix through a Boolean product of two lower-dimensional binary factor matrices. Because Boolean products capture nonlinear conjunctive or disjunctive structure while preserving interpretability,

BMF has been used in machine learning and across a range of scientific domains, including recommender systems, genomics, and behavioral data analysis (Miettinen and Vreeken, 2011; Rukat et al., 2017a; Wan et al., 2020; Miettinen and Neumann, 2021; Liang et al., 2020; Haddad et al., 2018; Rukat et al., 2017b; Dzyabura and Hauser, 2011). An important probabilistic instance of BMF is the Deterministic Input, Noisy output And gate (DINA) model (Junker and Sijtsma, 2001), which combines Boolean factorization with observation noise. We use DINA as the primary modeling and theoretical vehicle in this paper.

In educational testing, the DINA model is a fundamental cognitive diagnosis model (CDMs; Rupp and Templin, 2008b; von Davier and Lee, 2019): the observed binary matrix records item responses, one binary factor matrix represents subject-level latent skill attribute mastery, and the other represents item-level attribute requirements, known in psychometrics as the Q-matrix (Tatsuoka, 1983). Identifiability and estimation of the DINA model have received substantial interest (see, e.g. Rupp and Templin, 2008a; de la Torre, 2009; DeCarlo, 2011; Chen et al., 2015; Culpepper, 2015; Chen et al., 2018; Yamaguchi and Okada, 2020b). More generally, however, the same model can be viewed as a structured *conjunctive* Boolean latent feature model for multivariate binary data. This places DINA within probabilistic Boolean matrix factorization rather than a purely psychometric setting.

More broadly, existing approaches to probabilistic Boolean matrix factorization and related binary latent feature models face two recurring obstacles in the setting of unknown latent dimension K and large observed dimension P. First, latent-state configurations often grow exponentially with the number of binary latent attributes, making estimation and computation increasingly difficult as the latent dimension increases. Second, principled mechanisms for selecting the latent dimension while jointly estimating both binary factor matrices remain limited. These challenges are especially acute in high-dimensional binary data, where both the observed dimension and the latent dimension may be substantial.

Existing estimation methods for the DINA model illustrate the broader unknown-dimension

bottleneck described above. Likelihood-based approaches typically marginalize over all latent attributes and are therefore primarily suited to fixed-dimensional settings (e.g., Chen et al., 2015; Xu and Shang, 2018). The joint maximum likelihood approach of Gu and Xu (2023) scales well to high-dimensional data by treating the latent attributes as fixed parameters, but it still assumes that the true latent dimension is known. Bayesian MCMC methods (Culpepper, 2015; Chen et al., 2018) provide a flexible alternative but become computationally burdensome when the sample size (N), observed dimension (P), or latent dimension (K) is large. As a result, scalable joint estimation of the two binary factor matrices together with the unknown latent dimension remains unresolved in this setting.

Variational inference (VI) offers a natural alternative for scalable estimation, but existing VI methods for DINA inherit a related combinatorial difficulty. Current approaches rely on latent-class representations over all 2^K binary attribute patterns, so the variational dimension grows exponentially with the number of latent attributes. For example, the variational Bayes methods of Yamaguchi (2020) and Oka and Okada (2023) enumerate all binary attribute patterns either at the subject level or within item-wise posterior updates, which limits practical use to relatively small K. Related extensions to more general diagnostic classification models and multiple-choice variants retain the same basic dependence on latent-class representations (Yamaguchi and Okada, 2020a; Yamaguchi, 2020). Consequently, existing VI approaches do not provide a scalable solution to unknown-K inference in high-dimensional binary data. This gap is particularly consequential in modern applications, including settings beyond educational testing such as single-cell genomics, where both the observed dimension and the latent dimension may be substantial.

Our approach also differs from much of the broader VI literature for discrete latent variable models. Existing work often addresses the non-differentiability and combinatorial structure of discrete variables through stochastic optimization, including biased continuous relaxations (Jang et al., 2016; Maddison et al., 2016) and unbiased score-function estimators

with variance reduction (Tucker et al., 2017; Yin and Zhou, 2019; Kunes et al., 2023). In contrast, the Boolean factorization structure of DINA permits closed-form coordinate ascent updates through analytic partial integration, in the spirit of the local expectation gradient method (Michalis and Lázaro-Gredilla, 2015). This yields substantially simpler optimization that simultaneously supports scalable inference and latent-dimension selection.

Our Contributions. *Methodologically*, we develop a scalable mean-field variational inference method for probabilistic Boolean matrix factorization with unknown latent dimension, using the DINA model as a structured conjunctive instance. The method combines a cumulative shrinkage process prior for latent-dimension selection with closed-form coordinate ascent updates, yielding per-iteration cost $O(NPK_{\max})$, where $K_{\max}$ is a user-specified upper bound on the latent dimension. This provides a computationally efficient procedure for jointly estimating the latent binary factor matrix, the binary loading matrix, and the latent dimension in large-scale, high-dimensional binary data. *Theoretically*, we analyze both a full-parametrization variational estimator and the proposed mean-field algorithm. We establish sample-complexity guarantees for accurate recovery and derive a minimax lower bound, showing that the resulting rates agree up to logarithmic factors, with near-minimax optimality in the small-noise regime. *Empirically*, we evaluate the method through extensive simulations and two real-data applications from educational assessment and single-cell genomics. Our development is carried out for the DINA model with the Noisy Boolean "And" assumption, while the DINO model (Templin and Henson, 2006) with the Noisy Boolean "Or" assumption is covered as a special case via reparametrization (§ 2). Extensions to other probabilistic BMF models beyond DINA/DINO are discussed in § 7.

The rest of this manuscript is organized as follows. § 2 introduces the model setup and background. § 3 proposes the variational inference method and discusses practical considerations. § 4 presents theoretical guarantees of the proposed method. §§ 5 and 6 provide simulation results and real data applications, respectively. § 7 concludes. Proofs of theoret-

ical results and additional numerical results are included in the Supplementary Material.

2 Model Setup

We adopt the following notations. For any positive integer M, denote $[M] = \{1, \ldots, M\}$. For a matrix $\mathbf{M}$, denote its ith row by $\mathbf{M}_{i,:}$. For two vectors $\boldsymbol{a} = (a_1, \ldots, a_L)$ and $\boldsymbol{b} = (b_1, \ldots, b_L)$ of the same length, we write $\boldsymbol{a} \succeq \boldsymbol{b}$ if $a_l \geq b_l$ for all $l \in [L]$.

We consider an $N \times P$ binary data matrix $\mathbf{Y} = (Y_{ij}) \in \{0,1\}^{N \times P}$, where rows index subjects or samples and columns index items or features. Let $\mathbf{A} = (A_{ik}) \in \{0,1\}^{N \times K}$ denote a binary latent-attribute matrix and let $\mathbf{Q} = (Q_{jk}) \in \{0,1\}^{P \times K}$ denote a binary loading matrix. The DINA model is a structured probabilistic Boolean matrix factorization model in which the *binary ideal response* is generated through a conjunctive Boolean-And relationship between the rows of $\mathbf{A}$ and $\mathbf{Q}$. In educational testing, rows correspond to students, columns to test questions, $\mathbf{A}$ records latent attribute mastery, and $\mathbf{Q}$ is known as the *Q-matrix*. More generally, however, the same formulation applies to multivariate binary data whenever an observed feature is likely to be active only if all required latent attributes are present.

Many standard matrix factorization models are based on inner products and additive latent-factor contributions to the mean of an observed variable. In contrast, the DINA model is governed by a Boolean product, so the latent attributes interact nonlinearly through a highest-order conjunctive rule. This difference is central both statistically and computationally, which produces an interpretable binary loading structure while also leading to a non-Gaussian latent-variable estimation problem.

The *ideal response* of subject i to item j is defined through the Boolean-And gate

$$\mathrm{IR}_{\mathrm{AND}}(\mathbf{A}_{i,:}, \mathbf{Q}_{j,:}) = 1(A_{ik} \geq Q_{jk} \text{ for all } k \in [K]) = \prod_{k=1}^{K} A_{ik}^{Q_{jk}} = \prod_{k=1}^{K} \big(1 - (1 - A_{ik})Q_{jk}\big),$$

where the final equality holds for all $A_{ik}, Q_{jk} \in \{0,1\}$. The above ideal response equals 1 if and only if $A_{ik} = 1$ whenever $Q_{jk} = 1$ for all $k \in [K]$. In educational testing, this means

a student gives the ideal response to an item only when all required latent attributes are mastered. The DINA model likelihood is

$$\mathbb{P}(Y_{ij} = 1 \,|\, A_{i\cdot}, Q_{j\cdot}, s_j, g_j) = \left(\prod_{k=1}^{K} A_{ik}^{Q_{jk}}\right)(1 - s_j) + \left(1 - \left(\prod_{k=1}^{K} A_{ik}^{Q_{jk}}\right)\right) g_j, \tag{1}$$

where s_j and g_j capture the item-specific noise levels, and are interpreted as *slipping* and *guessing* parameters in the educational setting (Junker and Sijtsma, 2001): s_j quantifies the probability of a "careless mistake" despite mastering all required latent skills of item j, and g_j quantifies the probability of a "lucky guess" despite lacking some required skills.

The log-likelihood function forms the basis of our optimization objectives. To ease notation, we introduce the separate components of the log-likelihood function:

$$\psi_{1,j}(Y_{ij}) := Y_{ij} \log(1 - s_j) + (1 - Y_{ij}) \log(s_j), \;\; \psi_{2,j}(Y_{ij}) := Y_{ij} \log(g_j) + (1 - Y_{ij}) \log(1 - g_j).$$

where $\psi_{1,j}(Y_{ij})$ and $\psi_{2,j}(Y_{ij})$ are the log-likelihood of Y_{ij} given the ideal response equal to 1 and 0, respectively. Accordingly, the joint log-likelihood function is

$$\mathcal{L}(\mathbf{A}, \mathbf{Q}) := \sum_{i=1}^{N} \sum_{j=1}^{P} \left[\left(\prod_{k=1}^{K} A_{ik}^{Q_{jk}}\right) \psi_{1,j}(Y_{ij}) + \left(1 - \prod_{k=1}^{K} A_{ik}^{Q_{jk}}\right) \psi_{2,j}(Y_{ij}) \right]. \tag{2}$$

The DINO model (Templin and Henson, 2006) provides a complementary *disjunctive* probabilistic Boolean matrix factorization formulation based on a Boolean-Or gate. Given two binary factor matrices $\mathbf{A}$ and $\mathbf{Q}$, the ideal response of subject i to feature j is

$$\mathrm{IR}_{\mathrm{OR}}(\mathbf{A}_{i,:}, \mathbf{Q}_{j,:}) = 1(A_{ik} \geq Q_{jk} \text{ for at least one } k \in [K]) = 1 - \prod_{k=1}^{K} \big[(1 - A_{ik}) Q_{jk}\big].$$

One can verify that

$$\mathrm{IR}_{\mathrm{OR}}(\mathbf{A}_{i,:}, \mathbf{Q}_{j,:}) = 1 - \mathrm{IR}_{\mathrm{AND}}(1_K - \mathbf{A}_{i,:}, \mathbf{Q}_{j,:}),$$

so estimation procedures developed for DINA transfer directly to DINO through reparameterization. This extension is used in §6.2 for analyzing single-cell chromatin accessibility data.

3 New Variational Inference Algorithm

Given a latent variable model with the observed data $\mathbf{Y}$ and unobserved variables $\boldsymbol{\theta}$, variational inference (VI) is a Bayesian inference framework to approximate the posterior distribution $P(\boldsymbol{\theta} \,|\, \mathbf{Y})$ (Blei et al., 2017). VI defines a family of possible approximate distributions $\{V : V \in \mathcal{V}\}$ and selects the closest to the posterior $P(\boldsymbol{\theta} \,|\, \mathbf{Y})$ via optimization. The optimal approximation is obtained by minimizing the KL divergence, $\widehat{V} = \arg\min_{V \in \mathcal{Q}} D_{\mathrm{KL}}\Big(V, P(\boldsymbol{\theta} \,|\, \mathbf{Y})\Big)$.

Cumulative Shrinkage Prior to Select the Latent Dimension. We develop a new scalable mean-field variational inference approach for the DINA model with an unknown number of latent attributes K by introducing an increasing shrinkage prior on the $\mathbf{Q}$ matrix. Specifically, we adopt the cumulative shrinkage process (CSP; Legramanti et al., 2020; Legramanti, 2020), CSP is a spike-and-slab shrinkage prior where the spike probability is stochastically increasing as the column index in the loading matrix (of a linear Gaussian latent factor model) increases. For the binary factor matrix $\mathbf{Q}$, we impose the following prior distributions with an upper bound $K_{\max}$ for its number of columns:

$$Q_{jk} \sim (1 - \pi_k) \cdot \text{Bernoulli}(0.5) + \pi_k \cdot \text{Bernoulli}\,(\delta)\,, \quad k = 1, \ldots, K_{\max}; \tag{3}$$

$$\pi_k = \sum_{l=1}^{k} \omega_l; \quad \omega_l = \nu_l \prod_{m=1}^{l-1} (1 - \nu_m); \quad \nu_l \sim \text{Beta}(1, \kappa), \quad l = 1, \ldots, K_{\max} - 1, \tag{4}$$

where $\delta > 0$ is a small constant close to zero and $\nu_{K_{\max}} = 1$. Eq. (4) gives a truncated stick-breaking construction, and Eq. (3) imposes a spike-and-slab prior for Q_{jk}, with a spike distribution Bernoulli(δ) concentrated at zero and a slab distribution Bernoulli(1/2). Intuitively, the spike models those inactive and redundant columns of $\mathbf{Q}$ and the slab models the active columns. Here, $K_{\max}$ is the maximum possible number of latent components, which

can be set to the observed feature dimension P if without further prior knowledge. Usually, the true number of latent components can be much smaller than $K_{\max}$.

To facilitate a scalable coordinate ascent variational inference algorithm, we introduce auxiliary categorical variables $z_1, \ldots, z_{K_{\max}} \in [K_{\max}]$ similarly as Legramanti et al. (2020):

$$Q_{jk} \sim (1 - \mathbb{1}(z_k \leq k)) \cdot \text{Bernoulli}(0.5) + \mathbb{1}(z_k \leq k) \cdot \text{Bernoulli}\,(\delta)\,, \quad k \in [K_{\max}]; \tag{5}$$

$$\mathbb{P}(z_k = l) = \omega_l, \quad \forall k, l \in [K_{\max}]. \tag{6}$$

The binary indicator $\mathbb{1}(z_k \leq k)$ characterizes the event that the kth column in $\mathbf{Q}$ is redundant. We can derive full conditional distributions $\mathbb{P}(z_k = l \mid -)$ for all $k, l \in [K_{\max}]$.

The spike probability $\pi_k = \mathbb{P}(z_k \leq k) = \sum_{l=1}^{k} \omega_l$ increases monotonically with k, so higher-index columns of $\mathbf{Q}$ are *a priori* more likely to be in the spike (redundant) regime. This is the cumulative shrinkage effect: the model is automatically pushed toward sparse solutions where only the first few columns of $\mathbf{Q}$ are active.

There are two user-defined hyperparameters in Eqs. (3) and (4): κ controls the prior expected number of active columns of $\mathbf{Q}$, and $\delta > 0$ is the Bernoulli probability for entries of $\mathbf{Q}$ in a redundant column. In our numerical experiments, δ is fixed to 0.01 and κ is fixed to 2, while $K_{\max}$ is set as the number of features P, which is a computationally challenging extreme case to avoid under-selection of K. Somewhat surprisingly, we find that in practice, our proposed variational inference algorithm is not only scalable but also effective with minimal hyperparameter tuning. Indeed, recovery of the true latent dimension K is reasonably accurate across all experimental settings with these fixed choices of hyperparameters. With the likelihood in Eq. (1), the joint distribution of DINA model with CSP prior is

$$\begin{aligned} &p(\mathbf{Y}, \mathbf{A}, \mathbf{Q}, z, \nu; s, g) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{P} \left\{ p(Y_{ij} \mid A_{i,:}, Q_{j,:}; s, g) \prod_{k=1}^{K_{\max}} \left[p(Q_{jk} \mid z_k) p(z_k \mid \nu_{1:K}) p(A_{ik}) \right] \right\} \prod_{k=1}^{K_{\max}} p(\nu_k), \end{aligned} \tag{7}$$

with latent variables $(\mathbf{A}, \mathbf{Q}, z, \nu)$ and model parameters (s, g).

Variational Inference Algorithm. The latent variables for the joint model are $\{A_{ik}\}_{i\in[N],k\in[K_{\max}]}$, $\{Q_{jk}\}_{j\in[P],k\in[K_{\max}]}$, as well as $z = \{z_k\}_{k=1}^{K_{\max}}$, $\nu = \{\nu_k\}_{k=1}^{K_{\max}}$. We introduce parameters $\alpha = (\alpha_{ik})_{N\times K}$ and $\gamma = (\gamma_{jk})_{P\times K}$ to approximate the posterior $P(\mathbf{A}, \mathbf{Q} \mid \mathbf{Y})$ with the mean-field approximation $V(\mathbf{A}, \mathbf{Q}; \alpha, \gamma) = \prod_{i=1}^{N}\prod_{k=1}^{K} q(A_{ik}; \alpha_{ik}) \prod_{j=1}^{P}\prod_{k=1}^{K} q(Q_{jk}; \gamma_{jk})$, where $q(A_{ik}; \alpha_{ik})$ and $q(Q_{jk}; \gamma_{jk})$ are Bernoulli distributions with parameters $\alpha_{ik}, \gamma_{jk} \in [0, 1]$. We let the one-hot encoding for the z_k in Eq. (6) to be $\mathbf{z}_k = (z_{k1}, \ldots, z_{kK_{\max}})$, with $z_{kl} = 1$ if $z_k = l$. In addition to the mean-field assumption on $\mathbf{A}$ and $\mathbf{Q}$, we further assume the approximate posteriors for the CSP parameters: $\nu_k \sim \text{Beta}(\pi_{0k}, \pi_{1k})$ independently and $z_k = (z_{k1}, \ldots, z_{kK_{\max}}) \sim \text{Categorical}(\phi_{k1}, \ldots, \phi_{kK_{\max}})$ with $\sum_{l=1}^{K_{\max}} \phi_{kl} = 1$.

Denote $\pi_0 = \{\pi_{0k}\}_{k=1}^{K_{\max}}$, $\pi_1 = \{\pi_{1k}\}_{k=1}^{K_{\max}}$, and $\phi = \{\phi_{kl}\}_{k,l\in[K_{\max}]}$. The variational distribution $V(\mathbf{A}, \mathbf{Q}, z, \nu)$ factorizes as a product of independent Bernoulli, categorical, and Beta distributions over all components; the full densities are given in Supplement S.2.

Together with the DINA model likelihood and the CSP prior, the evidence lower bound (ELBO) for the proposed VI algorithm is:

$$
\begin{aligned}
&\text{ELBO}(\alpha, \gamma, \phi, \pi_0, \pi_1) = \\
&\sum_{i=1}^{N}\sum_{j=1}^{P}\left(\prod_{k=1}^{K_{\max}}(1-(1-\alpha_{ik})\gamma_{jk})\right)\psi_{1,j}(Y_{ij}) + \left(1-\prod_{k=1}^{K_{\max}}(1-(1-\alpha_{ik})\gamma_{jk})\right)\psi_{2,j}(Y_{ij}) \\
&+\sum_{j=1}^{P}\sum_{k=1}^{K_{\max}}\left(\left(1-\sum_{l=1}^{k}\phi_{kl}\right)\log\left(\frac{1}{2}\right) + \left(\sum_{l=1}^{k}\phi_{kl}\right)(\gamma_{jk}\log(\delta) + (1-\gamma_{jk})\log(1-\delta))\right) \\
&+\sum_{k=1}^{K_{\max}}\sum_{l=1}^{K_{\max}}\phi_{kl}\mathbb{E}_{\nu_{1:l}}\log\left(\nu_l\prod_{m=1}^{l-1}(1-\nu_m)\right) + \sum_{k=1}^{K_{\max}}(\kappa-1)\mathbb{E}_{\nu_k}\log(1-\nu_k) + H_V \qquad (8) \\
&=: \tilde{L}(\alpha, \gamma, \phi, \pi_0, \pi_1) + H_V, \qquad (9)
\end{aligned}
$$

where $\tilde{L}$ denotes the expected complete-data log-likelihood under V and $H_V := -\int \log V \, dV$ is the entropy of V. The expected log-likelihood in the first row of Eq. (8) has a closed form

because the Boolean ideal response factors as $\prod_k A_{ik}^{Q_{jk}} = \prod_k (1 - (1 - A_{ik})Q_{jk})$, enabling analytic mean-field integration, see Supplement S.2. Throughout, we treat s and g as non-random fixed parameters estimated by coordinate maximization of the ELBO.

Closed-form Coordinate Ascent Updates. Despite the seemingly heavy notation, each group of variational parameters permits a closed-form coordinate ascent update, see Supplement S.2. First-order conditions lead to the following updates for the posterior variational Beta parameters of ν:

$$\pi_{0,k}^{(t)} = 1 + \sum_{l=1}^{K_{\max}} \phi_{l,k}^{(t-1)}, \qquad \pi_{1,k}^{(t)} = \kappa + \sum_{l=1}^{K_{\max}} \sum_{m=k+1}^{K_{\max}} \phi_{l,k}^{(t-1)},$$

for $k = 1, \ldots, K_{\max} - 1$. For the categorical parameters, first-order conditions lead to the following update:

$$\begin{aligned} \phi_{k,l}^{(t)} &= \mathbb{P}(z_k = l \mid -) \\ &\propto \begin{cases} \exp\left(\mathbb{E}_{\pi^{(t-1)}}(\log \omega_l) + \sum_{j=1}^{P} \gamma_{j,k}^{(t-1)} \log(\frac{\delta}{1-\delta}) + P \log(1-\delta)\right), & \text{if } l \leq k; \\ \exp\left(\mathbb{E}_{\pi^{(t-1)}}(\log \omega_l) + P \log(1/2)\right), & \text{if } l > k. \end{cases} \end{aligned} \tag{10}$$

The proportionality in Eq. (10) is across $l \in [K_{\max}]$ for each fixed k, with the normalizing constant $\sum_{l'=1}^{K_{\max}}[\cdot]$ making $\phi_{k,\cdot}^{(t)}$ a proper probability vector. The two cases reflect the two regimes introduced by the CSP prior: when $z_k = l \leq k$, column k is in the spike regime ($Q_{jk} \sim \text{Bernoulli}(\delta)$), and the data evidence $\sum_j \gamma_{jk} \log(\delta/(1-\delta))$, which is large and negative when column k is active, enters the update; when $z_k = l > k$, column k is in the slab regime ($Q_{jk} \sim \text{Bernoulli}(1/2)$) and the data contribute only the flat term $P \log(1/2)$. Together, the update $\phi_{k,l}^{(t)}$ weighs the prior ω_l against the data evidence for column k being active or redundant, and the resulting $\sum_{l=1}^{k} \phi_{k,l}^{(t)} = \mathbb{P}(z_k \leq k \mid -)$ is the posterior probability that column k is redundant. Denote by $\sigma(\cdot)$ the sigmoid function with $\sigma(x) = 1/(1 + e^{-x})$. For

the variational parameters (γ_{jk}) of the binary matrix $\mathbf{Q}$, the first-order conditions give:

$$\begin{aligned}\gamma_{jk}^{(t)} &= \sigma\Bigg(\sum_{i=1}^{N}\Bigg(-(1-\alpha_{ik}^{(t-1)})\prod_{l\neq k}(1-(1-\alpha_{il}^{(t-1)})\gamma_{jl}^{(t-1)})\Bigg)\psi_{1,j}(Y_{ij}) \\ &\quad + \Bigg((1-\alpha_{ik}^{(t-1)})\prod_{l\neq k}(1-(1-\alpha_{il}^{(t-1)})\gamma_{jl}^{(t-1)})\Bigg)\psi_{2,j}(Y_{ij}) + \Bigg(\sum_{l=1}^{k}\phi_{kl}^{(t-1)}\Bigg)\log\left(\frac{\delta}{1-\delta}\right)\Bigg) \\ &= \sigma\Big(\nabla_{\gamma_{jk}}\tilde{L}\Big(\alpha^{(t-1)},\gamma^{(t-1)},\phi^{(t-1)},\pi_0^{(t-1)},\pi_1^{(t-1)}\Big)\Big). \qquad (11)\end{aligned}$$

The term $\log(\delta/(1-\delta))$ in Eq. (11) is large and negative (since $\delta \approx 0$), acting as a column-wise sparsity penalty: when $\sum_{l=1}^{k}\phi_{k,l}^{(t-1)} \approx 1$ (column k likely redundant), it drives all $\gamma_{jk}^{(t)}$ toward zero. Conversely, small $\gamma_{jk}^{(t-1)}$ across j causes Eq. (10) to concentrate mass on $l \leq k$, reinforcing the penalty in the next iteration. This adaptive feedback resembles non-convex penalized methods (Fan and Li, 2001; Zou, 2006; Ročková and George, 2016).

Finally, our updates for the variational parameters (α_{ik}) for the binary matrix $\mathbf{A}$ are $\alpha_{ik}^{(t)} = \sigma\Big(\nabla_{\alpha_{ik}}\tilde{L}(\alpha^{(t-1)},\gamma^{(t-1)},\phi^{(t-1)},\pi_0^{(t-1)},\pi_1^{(t-1)})\Big)$, where importantly, the gradient term $\nabla_{\alpha_{ik}}\tilde{L}(\alpha^{(t-1)},\gamma^{(t-1)},\phi^{(t-1)},\pi_0^{(t-1)},\pi_1^{(t-1)})$ has no dependence on $\alpha_{ik}^{(t-1)}$, which facilitates computation of the update.

The closed-form updates for s and g are determined by coordinate maximization of the ELBO with all variational parameters (α,γ,ϕ,π) held fixed, see Supplement S.2; this is an exact M-step for (s_j,g_j) given the current variational distribution. The update formulas are given explicitly in Algorithm 1 (lines 10–11). To maintain identifiability and ensure $\rho = \log((1-s_j)/g_j) > 0$, we require $s_j, g_j \in (0, 1/2)$; in practice we project each update onto this interval after each ELBO maximization step. We terminate Algorithm 1 when the relative change in ELBO falls below a small threshold (e.g., 10^{-4}). We summarize the complete updates in Algorithm 1, which we call the *DINA-CSP* algorithm.

Spectral Initialization with Varimax Rotation. The ELBO is highly nonconvex, so a wise initialization is crucial to achieving a desirable convergence performance of the VI

Algorithm 1: Coordinate Ascent VI for DINA with a CSP Prior (DINA-CSP)

Data: Input data $\mathbf{Y} = (Y_{ij})_{N \times P}$, δ, κ, $K_{\max}$

Result: $\widehat{\gamma}$, $\widehat{\alpha}$, $\widehat{\pi}$, $\widehat{\phi}$, $\widehat{s}$, $\widehat{g}$

Initialization via top-$K_{\max}$ SVD followed by Varimax rotation;

while *not converged* **do**

  for $(j,k) \in [P] \times [K_{\max}]$ **do**

    $\gamma_{jk}^{(t)} = \sigma\Big(\nabla_{\gamma_{jk}} \tilde{L}(\alpha, \gamma, \phi, \pi_0, \pi_1)\Big)$;

  for $(i,k) \in [N] \times [K_{\max}]$ **do**

    $\alpha_{ik}^{(t)} = \sigma\Big(\nabla_{\alpha_{ik}} \tilde{L}(\alpha, \gamma, \phi, \pi_0, \pi_1)\Big)$;

  for $k \in [K_{\max}]$ **do**

    for $l \in [K_{\max}]$ **do**

      if $l \geq k+1$ **then**

        $\phi_{k,l}^{(t)} = \exp\Big(\mathbb{E}_{\pi^{(t-1)}}(\log \omega_l) + P \log(1/2)\Big)$;

      if $l < k+1$ **then**

        $\phi_{k,l}^{(t)} = \exp\Big(\mathbb{E}_{\pi^{(t-1)}}(\log \omega_l) + \sum_{j=1}^{P} \gamma_{j,k}^{(t-1)} \log(\delta) + (P - \sum_{j=1}^{P} \gamma_{j,k}^{(t-1)}) \log(1-\delta)\Big)$;

    normalize $\phi_k^{(t)} \leftarrow \phi_k^{(t)} / \sum_{l=1}^{K_{\max}} \phi_{kl}^{(t)}$;

    $\pi_{0,k}^{(t)} = 1 + \sum_{l=1}^{K_{\max}} \phi_{l,k}^{(t-1)}$;

    $\pi_{1,k}^{(t)} = \kappa + \sum_{l=1}^{K_{\max}} \sum_{m=k+1}^{K_{\max}} \phi_{l,k}^{(t-1)}$;

  `// Updates for slipping and guessing parameters` s `and` g

  for $j \in [P]$ **do**

    $s_j^{(t)} = \frac{\sum_{i=1}^{N} (1 - Y_{ij}) \Big(\prod_{k=1}^{K_{\max}} (1 - (1 - \alpha_{ik}^{(t-1)}) \gamma_{jk}^{(t-1)}) \Big)}{\sum_{i=1}^{N} \prod_{k=1}^{K_{\max}} (1 - (1 - \alpha_{ik}^{(t-1)}) \gamma_{jk}^{(t-1)})}$;

    $g_j^{(t)} = \frac{\sum_{i=1}^{N} Y_{ij} \Big(1 - \prod_{k=1}^{K_{\max}} (1 - (1 - \alpha_{ik}^{(t-1)}) \gamma_{jk}^{(t-1)}) \Big)}{\sum_{i=1}^{N} \Big(1 - \prod_{k=1}^{K_{\max}} (1 - (1 - \alpha_{ik}^{(t-1)}) \gamma_{jk}^{(t-1)}) \Big)}$;

return $\widehat{\gamma}$, $\widehat{\alpha}$, $\widehat{\pi}$, $\widehat{\phi}$, $\widehat{s}$, $\widehat{g}$;

algorithm (Yin et al., 2020). We propose to initialize the variational parameters γ of the item-attribute matrix $\mathbf{Q}$ by a spectral method-based factor analysis with Varimax rotation to encourage sparsity (Barber (2012), *Algorithm 21.1*). We then initialize the variational parameters α of the matrix $\mathbf{A}$ by applying the update for α_{ik} in Algorithm 1 until convergence with γ fixed, as detailed in Supplement S.5. Recently, Rohe and Zeng (2023) showed that applying Varimax after a spectral method can recover the true latent structure under certain

semiparametric factor models with a linear latent structure.

When every row of $\mathbf{Q}$ is a canonical basis vector, the Boolean product (i.e., DINA ideal response) reduces exactly to a linear factor model with binary latent factors, see Supplement S.4, making Varimax-rotated principal components a principled initialization. Empirically, this initialization leads to vast improvements over random initialization across all settings, even when $\mathbf{Q}$ is denser than the simple-structure case. Finally, the binary latent attributes in the vector A_i may be correlated, and such a correlation structure can be incorporated in the variational distribution by modeling $q(A_i)$ either as the marginal distribution of a hierarchical model, or by modeling it with full parameterization. However, due to the intensive computation of such parameterizations, we defer the details to Supplement S.3.

4 Theoretical Guarantees

Previous studies proposed the necessary and sufficient conditions for the *population identifiability* and estimability of the DINA model with either a known or unknown item-attribute matrix $\mathbf{Q}$ (Gu and Xu, 2021). Under that traditional identifiability notion, the latent attributes are treated as random variables and marginalized out. However, in the high-dimensional (large P) regime in this work, we need to consider a different notion of estimability, when the latent attributes $\mathbf{A}$ are treated as fixed parameters to be estimated. We develop several theoretical results: (i) the sparsity pattern of $\mathbf{Q}$ is preserved by the best linear approximation (Proposition 1); (ii) the sample complexity for accurate recovery as fixed points of the coordinate ascent algorithm (Theorem 1); and (iii) a minimax lower bound confirming rate optimality up to logarithmic factors (Theorem 2).

Proposition 1 (Linear approximation to the Boolean ideal response)**.** *If* $A_{ik} \sim_{iid} Bernoulli(1/2)$*, and define* γ_N *as the* $P \times K$ *matrix that minimizes the error between the Boolean ideal response*

matrix $(\prod_{k=1}^{K}(1-(1-A_{ik})Q_{jk}))_{i\in[N],j\in[P]}$ *and the low-rank approximation* $(\sum_{k=1}^{K}A_{ik}\gamma_{jk})_{i\in[N],j\in[P]}$*:*

$$\gamma_N = \underset{\gamma\in\mathbb{R}^{P\times K}}{\arg\min}\frac{1}{NP}\sum_{i=1}^{N}\sum_{j=1}^{P}\left(\prod_{k=1}^{K}(1-(1-A_{ik})Q_{jk}) - \sum_{k=1}^{K}A_{ik}\gamma_{jk}\right)^2$$

Then $\gamma_N \to \gamma^* = (\gamma^*_{jk})_{j\in[P],k\in[K]}$ *as* $N\to\infty$ *almost surely, with* $\mathbf{1}\{\gamma^*_{jk}>0\} = Q_{jk}$ *for all* $j = 1,\ldots,P$ *and* $k = 1,\ldots,K$*. Specifically, for* j *such that* $\sum_{k=1}^{K}Q_{jk} = 1$*, we have* $\gamma^*_{jk} = Q_{jk}$ *for all* $k\in[K]$*.*

The proof is given in Supplement S.1. Proposition 1 establishes two things precisely. First, for all items j, the sign pattern of the population minimizer matches $\mathbf{Q}$: $\mathbf{1}\{\gamma^*_{jk} > 0\} = Q_{jk}$. Second, for simple-structure items (i.e., items loading on exactly one attribute with $\sum_k Q_{jk} = 1$), the population minimizer equals $\mathbf{Q}$ exactly: $\gamma^*_{jk} = Q_{jk}$. The result is asymptotic ($N\to\infty$) and applies to the population-level solution γ^*, not directly to the finite-sample Varimax estimator. The following proposition establishes the corresponding finite-sample guarantee for simple-structure items.

Proposition 2 (Spectral initialization guarantee with simple-structure $\mathbf{Q}$)**.** *Suppose* $\mathbf{Q}$ *has a simple structure with every row being a canonical basis vector,* $s_j = g_j = s < 1/2$*, Assumptions 1 and 2 hold with constant* Ξ*, and* $N \asymp P$*. Then with probability at least* $1-\xi$*, the Varimax estimator* $\widehat{\gamma}_{\text{Varimax}}$ *satisfies the initialization condition of Theorem 1,*

$$\max_{(j,k)\in[P]\times[K]}|\widehat{\gamma}^{(0)}_{jk} - Q_{jk}| \le 1 - \left(1-\tfrac{\epsilon\Xi}{2}\right)^{1/K},$$

provided $N = \tilde{\Omega}\left(\frac{K^3}{\epsilon^4\Xi^4}\right)$*. This sample complexity is the dominant constraint for a complete end-to-end guarantee: Theorem 1 itself requires only* $N = \tilde{\Omega}(1/(\epsilon\Xi\min(\epsilon,\rho)))$*, so Proposition 2 is binding when* K *is large relative to* $\epsilon\Xi$*. The proof, which uses Wedin's theorem and matrix Bernstein, is given in Supplement S.1.*

For general $\mathbf{Q}$ with rows beyond canonical basis vectors, the highly nonlinear Boolean

product makes verifying the initialization condition of Theorem 1 challenging, so it remains an open problem for future research. In practice, however, the spectral initialization achieves good empirical performance across all settings (§ 5; see also Supplementary Figure S.2).

Next, we present theoretical results on estimating general $\mathbf{A}$ and $\mathbf{Q}$ matrices under the noisy setting with slipping and guessing parameters s and g. We begin by proposing identifiability conditions for $\mathbf{A}$ and $\mathbf{Q}$. To estimate A_{ik}, we need $\Omega(P)$ questions that student i can "almost answer" (mastering all required skills except skill k), formalized in Assumption 1. To estimate Q_{jk}, we analogously need $\Omega(N)$ students who possess all skills required by item j except skill k, formalized in Assumption 2. Throughout, $f(n) = \Omega(g(n))$ means $f(n) \geq c\, g(n)$ for some constant $c > 0$ independent of N, P, K.

Assumption 1. *For all $(i,k) \in [N] \times [K]$, $\sum_{j=1}^{P} Q_{jk} \prod_{l \neq k} A_{il}^{Q_{jl}} = \Omega(P)$. This states that for any student i and any latent skill k, there are $\Omega(P)$ number of questions in the exam that the student can "almost answer" in the sense that they possess all latent skills besides the skill k. Note that this is automatically satisfied if for each skill k, there are $\Omega(P)$ questions that solely measure that skill. Further, for sufficiently large P, assume $\frac{1}{P}\sum_{j=1}^{P} Q_{jk} \prod_{l \neq k} A_{il}^{Q_{jl}} \geq \Xi$ for some positive constant $\Xi \in (0,1)$.*

Assumption 2. *For all $(j,k) \in [P] \times [K]$, $\sum_{i=1}^{N} (1 - A_{ik}) \prod_{l \neq k} A_{il}^{Q_{jl}} = \Omega(N)$. Further, for sufficiently large N let $\frac{1}{N}\sum_{i=1}^{N} (1 - A_{ik}) \prod_{l \neq k} A_{il}^{Q_{jl}} \geq \Xi$ for some positive Ξ. This is the analogous assumption to Assumption 1.*

These are conditions on the true data-generating parameters $(\mathbf{A}, \mathbf{Q})$ and are not directly verifiable from the observed data alone. They can be interpreted as structural diversity conditions ensuring that each student-skill and item-skill pair is adequately represented; Examples 1–3 illustrate structures under which they hold.

Assumption 1 and Assumption 2 can be thought of as identifiability assumptions. Suppose $\sum_{j=1}^{P} Q_{jk} \prod_{l \neq k} A_{il}^{Q_{jl}} = 0$. Let $\tilde{A}_{i'k'} = A_{i'k'}$ for all $(i',k') \neq (i,k)$ and $\tilde{A}_{ik} = 1$, $A_{ik} = 0$. Further, let $\tilde{\mathbf{A}}$ be the matrix with entry (i,k) equal to $\tilde{A}_{ik}$. Then: $\mathbb{P}(\mathbf{Y}; \mathbf{A}, \mathbf{Q}) = \mathbb{P}(\mathbf{Y}; \tilde{\mathbf{A}}, \mathbf{Q})$.

Hence the data distribution under $A_{ik} = 1$ is indistinguishable from the data distribution under $A_{ik} = 0$. Assumptions 1 and 2 are very mild, covering the identifiability condition in Gu and Xu (2023) as a special case but more general than that. Specifically, the following examples illustrate what structures of $\mathbf{A}$ and $\mathbf{Q}$ satisfy these assumptions.

Example 1. *Suppose* $\mathbf{Q}$ *vertically stacks* C *copies of identity submatrices* I_K*. For* $C = 2$*, we have* $\mathbf{Q} = (I_K; I_K)^\top$*. For a given* k*, we have* $\prod_{l \neq k} A_{il}^{Q_{jk}} = 1$ *for the rows* Q_j *with* $Q_{jk} = 1$*. Hence, Assumption* 1 *is satisfied for any choice of* A*, as* $\frac{1}{P}\sum_{j=1}^{P} Q_{jk} \prod_{l \neq k} A_{il}^{Q_{jl}} \geq C/P = \Xi$*, where* $\Xi := C/P$*. This example covers the identifiability condition in Gu and Xu (2023).*

Assumptions 1 and 2 do not require $\mathbf{Q}$ to contain any identity submatrix I_K; Supplement S.1gives two further examples (a dense $\mathbf{A}$ with arbitrary $\mathbf{Q}$, and a paired-skill structure $\mathbf{Q} = \mathbf{A}$) that satisfy both assumptions. Such relaxed conditions depart significantly from existing identifiability analyses of the DINA model and related cognitive diagnostic models in high-dimensional cases (Gu and Xu, 2021). Note that $\prod_{l \neq k}(1 - A_{il}^{Q_{jl}})$, the leave-one-out ideal response, can be equivalently written as $\mathbb{1}(A_{il} \geq Q_{jl}$ for all $l \in [K] \setminus \{k\})$, so it indicates whether student i masters all required skills of question j besides the kth skill. Ξ can be thought of as the minimum (across i, k) fraction of questions (resp. students) that can almost be answered by student i if not considering skill k.

We present two complementary theoretical results on estimating $\mathbf{A}$ and $\mathbf{Q}$, stated jointly in Theorem 1. Part (A) is a *statistical estimator result* for the full-parametrization variational algorithm described in Supplement S.2, which represents each row of $\mathbf{A}$ by a distribution over all 2^K binary patterns. This formulation yields an explicit estimator whose statistical accuracy can be analyzed in the usual sense. Part (B) establishes the *fixed-point sample complexity* for the proposed Algorithm 1: it characterizes how many samples are needed for the true parameters $(\mathbf{A}, \mathbf{Q})$ to emerge as fixed points of the coordinate ascent procedure.

Because the ELBO is non-convex, fixed points of Algorithm 1 need not be global maximizers. Nevertheless, characterizing the sample complexity of fixed points is a principled and

established approach in the analysis of iterative statistical algorithms. The EM algorithm, which shares the coordinate-ascent structure of Algorithm 1, has been studied via exactly this framework (Balakrishnan et al., 2017): one establishes that when N and P exceed a threshold, the truth becomes a fixed point, and argues that proper initialization then leads the algorithm to return it as the estimate. We adopt the same strategy here. Establishing *global* convergence guarantees (analogous to the spectral convergence theory of Zhang et al. (2016) for simpler linear latent class models) requires controlling spectral properties of the DINA likelihood that are technically prohibitive due to the Boolean higher-order interaction structure, and we leave this as an important direction for future work.

Throughout, we make the following simplifying assumptions to facilitate the analysis: (i) the slipping and guessing rates s_j and g_j are known (this simplification is standard in the DINA theory literature; extending the analysis to unknown per-item rates is future work); (ii) $s_j = g_j =: s < \frac{1}{2}$; (iii) $K = K_{\max}$, since A_{ik} is non-identifiable when $Q_{jk} = 0$ for all $j \in [P]$ (without loss of generality in practice: Proposition 3 below shows that when the true dimension $K < K_{\max}$, the CSP prior automatically identifies redundant columns at the CAVI fixed point and drives their variational mass to zero). We write $\epsilon := \frac{1}{2} - s > 0$ for the signal level and $\rho := \log \frac{1-s}{s} > 0$ for the log-odds discriminability. Both parts of Theorem 1 are proved via a Bernstein inequality and union-bound argument, following the proof style of Zhang et al. (2016); the minimax lower bound in Theorem 2 is proved via Le Cam's method.

Theorem 1 (Sample complexity of Algorithms 1 and 2)**.** *Adopt the notation ϵ, ρ, and Ξ from the preceding discussion. Let $\mathcal{M} > 0$ be a fixed target accuracy and $\xi \in (0, 1)$ be a target failure probability. We state two results, for Algorithm 2 and Algorithm 1, respectively:*

(A) Suppose that for all $(j, k) \in [P] \times [K]$, it holds that $|\gamma_{jk}^{(0)} - Q_{jk}| \leq 1 - \left(1 - \frac{\epsilon\Xi}{2}\right)^{1/K}$. Then for estimates $\widehat{\gamma}_{jk}$ and $\widehat{\chi}_i$ obtained via a finite number of iterations of Algorithm 2, where $\chi_i \in \Delta^{2^K-1}$ is the variational distribution over all 2^K binary patterns for student i under the full parametrization, with probability at least $1 - \xi$, for all $(i, j, k) \in [N] \times$

$[P] \times [K]$, *it holds that* $|\widehat{\gamma}_{jk} - Q_{jk}| < \mathcal{M}$ *and* $m_i := \arg\max(\widehat{\chi}_i) = A_i$, *provided that* $N = \tilde{\Omega}\Big(\frac{1}{\epsilon\Xi \min(\epsilon,\rho)}\Big)$, *and* $P = \tilde{\Omega}\Big(\frac{1}{\epsilon\Xi \min(\epsilon,\rho)}\Big)$. *The dependence on* $\mathcal{M}$ *and* ξ *is hidden in logarithmic factors.*[1]

(B) Suppose that for all $(j,k) \in [P] \times [K]$ *and* $(i,k) \in [N] \times [K]$,

$$|\alpha_{ik}^{(0)} - A_{ik}| \leq 1 - \Big(1 - \frac{\epsilon\Xi}{2}\Big)^{1/K}, \quad |\gamma_{jk}^{(0)} - Q_{jk}| \leq 1 - \Big(1 - \frac{\epsilon\Xi}{2}\Big)^{1/K}.$$

Then for estimates $\widehat{\alpha}_{ik}$ *and* $\widehat{\gamma}_{jk}$ *obtained via a finite number of iterations of Algorithm 1, with probability at least* $1 - \xi$, *for all* $(i,j,k) \in [N] \times [P] \times [K]$,

$$|\widehat{\gamma}_{jk} - Q_{jk}| < \mathcal{M}, \quad |\widehat{\alpha}_{ik} - A_{ik}| < \mathcal{M},$$

provided $N = \tilde{\Omega}\Big(\frac{1}{\epsilon\Xi \min(\epsilon,\rho)}\Big)$ *and* $P = \tilde{\Omega}\Big(\frac{1}{\epsilon\Xi \min(\epsilon,\rho)}\Big)$. *The dependence on* $\mathcal{M}$ *and* ξ *is hidden in logarithmic factors.*

Proofs of Parts (A) and (B) are given in Supplement S.1. The two parts together show that both the full-parametrization Algorithm 2 and the mean-field Algorithm 1 require $\tilde{\Omega}(1/(\epsilon\Xi \min(\epsilon,\rho)))$ samples (the same rate) to accurately recover $\mathbf{A}$ and $\mathbf{Q}$. This shared rate reflects that the mean-field restriction does not inflate the sample complexity compared to the unconstrained variational family.

Theorem 1 is stated under assumption (iii) ($K = K_{\max}$), meaning all $K_{\max}$ columns of $\mathbf{Q}$ are active. The following result shows this is without loss of generality: when the true dimension $K < K_{\max}$, the CSP mechanism in Algorithm 1 automatically identifies the K active columns and suppresses the remaining $K_{\max} - K$ redundant ones, recovering the correct dimension at the fixed point.

[1] Formally, $f(n) = \tilde{\Omega}(g(n))$ means that there exist constants $c > 0$, $k \geq 0$, and n_0 such that for all $n \geq n_0$, $f(n) \geq c\, g(n)\, (\log n)^{-k}$. In other words, $f(n)$ is asymptotically bounded below by $g(n)$ up to polylogarithmic factors.

Proposition 3 (CSP Column Selection Consistency). *Let* $\Sigma_k := \sum_{j=1}^{P} \gamma_{jk}$. *Suppose* $\delta < 2^{-1/\Xi}$ *and that* $(\alpha^*, \gamma^*, \phi^*)$ *is a fixed point of Algorithm 1 at which* $\Sigma_k \geq \Xi P$ *for all* $k \leq K$ *(active columns) and* $\Sigma_k \leq M$ *for all* $k > K$ *(redundant columns), for some* $0 \leq M < P\Xi$. *Then:*

(i) For each active column $k \leq K$, *the posterior probability that the column is redundant under the CSP prior satisfies* $\sum_{l=1}^{k} \phi_{kl}^* \leq K_{\max} \exp\big(-P[\Xi \log(1/\delta) - \log 2]\big)$.

(ii) For each redundant column $k > K$, *the posterior probability that the column is active satisfies* $\sum_{l=k+1}^{K_{\max}} \phi_{kl}^* \leq K_{\max} \exp\big(-P[\log 2 - (M/P)\log(1/\delta)]\big)$.

Both bounds decay exponentially in P, *suggesting consistency of latent dimension selection.*

Corollary 1 (Consistent Identification of the Latent Dimension). *Under the conditions of Theorem 1 (Parts A or B) and Proposition 3, with* $\delta < 2^{-1/\Xi}$, *the estimator* $\widehat{K} = \big|\{k \in [K_{\max}] : \Sigma_k \geq \tau\}\big|$ *with any threshold* $\tau \in (M, \Xi P)$ *satisfies* $\widehat{K} = K$ *with probability at least* $1 - \xi$ *when* $N, P = \tilde{\Omega}(1/(\epsilon\Xi \min(\epsilon, \rho)))$.

Next, we establish the minimax lower bound of the estimation problem, which shows that accurate recovery using $o(\frac{1}{\epsilon\rho\Xi})$ samples is impossible.

Theorem 2 (Minimax lower bound). *We show that* $\epsilon\rho\Xi$ *is the fundamental quantity that determines the sample complexity. Let the parameter space* Θ *be the set of* $\mathbf{A}$ *and* $\mathbf{Q}$ *such that Assumptions* 1 *and* 2 *hold (with constant* Ξ*), assumed non-empty. Then:*

- *If* $P \leq \kappa \frac{1}{\epsilon\rho\Xi}$ *for* $0 < \kappa < 1$, *we have* $\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A},\mathbf{Q}\in\Theta} \mathbb{P}\Big(\widehat{\mathbf{A}} \neq \mathbf{A}\Big) \geq \frac{1}{4}(1 - \sqrt{\kappa})$;
- *If* $N \leq \kappa \frac{1}{\epsilon\rho\Xi}$ *for* $0 < \kappa < 1$, *we have* $\inf_{\widehat{\mathbf{Q}}} \sup_{\mathbf{A},\mathbf{Q}\in\Theta} \mathbb{P}\Big(\widehat{\mathbf{Q}} \neq \mathbf{Q}\Big) \geq \frac{1}{4}(1 - \sqrt{\kappa})$.

The proof uses Le Cam's method to construct two hypotheses that are close in total variation yet far in parameter space. Combining Theorems 1 and 2, we establish that Algorithm 1 achieves the minimax rate up to a factor of $\rho/\min(\rho, \epsilon)$ and logarithmic terms.

Since $\rho = 4\epsilon + O(\epsilon^3)$, as shown in Supplement S.1, this gap is bounded above by $\rho/\epsilon \to 4$ as $\epsilon \to 0$. Two regimes arise: when $\rho < \epsilon$ (moderate to large noise), the gap equals 1 and the rates in Theorems 1 and 2 match exactly up to logarithmic factors; when $\rho \geq \epsilon$ (small noise, s near $1/2$), the gap converges to 4, establishing near-minimax optimality in this regime. We also note that the lower bound in Theorem 2 holds for all estimators without an initialization assumption, while Theorem 1 requires a specific initialization quality; both characterize the fundamental difficulty of recovering $\mathbf{A}$ and $\mathbf{Q}$ in this model.

Two aspects of the analysis can potentially be strengthened in future work. First, we assume that s_j and g_j are known; in practice, accurate estimation of these parameters is feasible once (α, γ) is initialized close to $(\mathbf{A}, \mathbf{Q})$, and incorporating them into the theoretical analysis is a natural extension. Second, establishing global convergence guarantees for Algorithm 1 (rather than fixed-point sample complexity) requires controlling the spectral structure of the DINA likelihood Hessian through its Boolean nonlinearity, which we leave as a technically challenging open problem.

5 Simulation Studies

We examine the performance of DINA-CSP for various sample size $N \in \{500, 1000, 2000\}$, the true underlying latent dimension $K \in \{3, 4, 5, 10\}$, and the attribute correlation $r \in \{0, 0.25, 0.5\}$. The slipping and guessing parameters are set to 0.2. The number of latent components is assumed to be unknown.

For the matrix $\mathbf{A}$, we set $A_{ik} \sim \text{Bernoulli}(0.5)$ for $r = 0$, and for $r > 0$, we follow Chen et al. (2018) by first generating a vector $\boldsymbol{\theta}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_K)$ where $\boldsymbol{\Sigma}_K = (1-r)\mathbf{I}_K + r(\mathbf{1}_K\mathbf{1}_K^\top)$ is the covariance matrix with the diagonal elements being 1 and all off-diagonal elements being r. Following Chen et al. (2021), each binary latent vector $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iK})$ follows $\alpha_{ik} = \mathbb{I}\left\{\theta_{ik} > \Phi^{-1}\left(\frac{k}{K+1}\right)\right\}$, where Φ^{-1} is the inverse cumulative distribution function of a standard normal distribution. From the cognitive diagnostic modeling perspective, the above construction encodes correlated and heterogeneous difficulty levels of the latent skill

attributes, where attributes with a larger index k are more difficult to master.

For $K \leq 5$, we specify the matrix $\mathbf{Q}$ in the same way as Table 1 in Chen et al. (2021), which corresponds to $(K, P) = (3, 18), (4, 18), (5, 20)$, respectively. For $K > 5$, we specify the matrix $\mathbf{Q}$ following the settings of Gu and Xu (2023) with $(K, P) = (10, 30)$. In each simulation setting, the true $\mathbf{Q}$ matrix contains $\lfloor P/(2K) \rfloor$ copies of I_K as sub-matrices stacking vertically, and $K_{\max}$ is set as P.

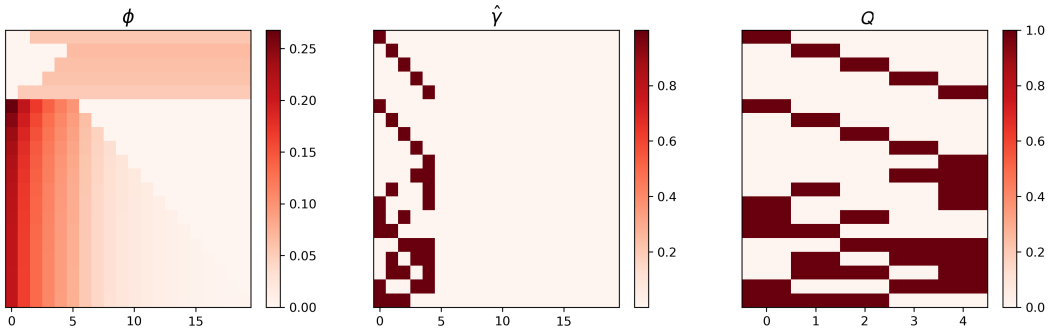

Figure 1: Example run of our algorithm, $s = 0.2$, $g = 0.2$, $N = 1000$, $P = 20$, $K = 5$, $K_{\max} = 20$. *Left:* Estimated ϕ variational parameter, indicating 5 columns are selected. Under the variational parameterization, the row sum $\sum_{l=1}^{k} \phi_{kl}$ of the lower diagonal matrix of ϕ is the probability that the kth column in the Q-matrix is redundant, which corresponds to the rows $k \geq 6$ for the left panel. *Middle:* Estimated γ variational parameters. *Right:* Ground truth $\mathbf{Q}$ matrix.

The average recovery rates are presented in Table 1 and an example run is shown in Fig. 1. All the results are averaged over 100 independent trials in each simulation setting. We adopt the elementwise accuracy rate (EAR) and the number of over-specified elements (NOSE) as evaluation criteria (Chen et al., 2021), defined as

$$\text{EAR} := \min_{\sigma:[K]\to[K]} \frac{1}{PK} \sum_{j=1}^{P} \sum_{k=1}^{K} \mathbb{1}[\widehat{Q}_{jk} = Q_{j\sigma(k)}], \quad \text{NOSE} := \sum_{j=1}^{P} \sum_{k=K+1}^{\widehat{K}} \widehat{Q}_{jk}. \tag{12}$$

EAR is the proportion of the original true $\mathbf{Q}$ matrix entries that are estimated correctly, under the best permutation $\sigma(\cdot)$ of the K columns. NOSE measures the false discovery in excessively estimating the redundant elements in the Q-matrix.

In addition, for a column k in the estimated $\widehat{\mathbf{Q}}$, we regard it as redundant if $\sum_j \mathbb{I}[\widehat{Q}_{jk} >$

K	N	$\widehat{K}_{est.}$	EAR	NOSE	K	N	$\widehat{K}_{est.}$	EAR	NOSE
$r = 0$									
3	500	95%	99.30%	0.09	5	500	100%	98.50%	0
	1000	99%	100.00%	0.02		1000	100%	99.50%	0
	2000	100%	100.00%	0		2000	100%	99.90%	0
4	500	100%	98.20%	0	10	500	100%	99.40%	0
	1000	100%	99.50%	0		1000	100%	99.70%	0
	2000	100%	100.00%	0		2000	100%	100.00%	0
K	N	$\widehat{K}_{est.}$	EAR	NOSE	K	N	$\widehat{K}_{est.}$	EAR	NOSE
$r = 0.25$									
3	500	99%	86.30%	0.06	5	500	88%	88.80%	0.18
	1000	100%	86.80%	0		1000	99%	89.60%	0
	2000	100%	86.80%	0		2000	100%	89.40%	0
4	500	95%	86.10%	0.09	10	500	39%	92.50%	5
	1000	99%	86.20%	0		1000	67%	95.10%	1.16
	2000	100%	85.80%	0		2000	73%	96.10%	0.07

Table 1: Average recovery rate under each condition. The EAR (higher is better) and NOSE (lower is better)are defined in Eq. (12). A high EAR corresponds to an accurate estimation of Q-matrix, while a lower NOSE reflects a more accurate estimation of the number of attributes K.

K	N	$\widehat{K}_{est.}$ CSP	$\widehat{K}_{est.}$ Crimp	K	N	$\widehat{K}_{est.}$ CSP	$\widehat{K}_{est.}$ Crimp	K	N	$\widehat{K}_{est.}$ CSP	$\widehat{K}_{est.}$ Crimp
$r = 0$				$r = 0.25$				$r = 0.5$			
3	500	96%	96%	3	500	99%	82%	3	500	100%	79%
	1000	99%	92%		1000	100%	85%		1000	100%	74%
	2000	100%	90%		2000	100%	87%		2000	100%	74%
4	500	100%	99%	4	500	95%	80%	4	500	84%	76%
	1000	100%	90%		1000	99%	68%		1000	87%	65%
	2000	100%	76%		2000	100%	65%		2000	95%	37%
5	500	100%	90%	5	500	88%	41%	5	500	62%	37%
	1000	100%	65%		1000	99%	32%		1000	75%	20%
	2000	100%	57%		2000	100%	21%		2000	79%	15%

Table 2: Comparison of K estimate.

$0.5] = 0$, which means that no item is estimated by the posterior mode to require this attribute, and we compare the estimated attribute number $\widehat{K}$ with the ground truth. Table 2 compares the estimation accuracy of the latent attributes K between Algorithm 1 and the Crimp sampler (Chen et al., 2021), a computationally intensive MCMC method.

Comparison with existing variational inference methods. The closest methodological competitors to DINA-CSP are the variational Bayes approaches of Yamaguchi and Okada (2020b) and Oka and Okada (2023). Neither provides a publicly available implementation, so direct comparison on matched datasets is not possible. Both methods represent each subject's latent profile across all 2^K binary attribute patterns, giving per-iteration costs of $O(N \cdot 2^K)$ and $O(P \cdot (2^K - 1))$ respectively; both assume K is known and are practically limited to $K \leq 8$ in their original implementations. By contrast, DINA-CSP has per-iteration cost $O(NPK)$, does not require K to be pre-specified, achieves EAR $\geq 98.2\%$ at $K \in \{3, 4, 5\}$ with $N = 500$ and $r = 0$ while simultaneously recovering $\widehat{K}$ with at least 95% accuracy (Tables 1 and 2), and scales to $K \in \{15, 25, 35\}$ where neither competitor is feasible (Table 3). A detailed breakdown of the comparison is provided in Supplement S.7.

We further examine the scalability of DINA-CSP for a large number of latent attributes $K \in \{15, 25, 35\}$. This corresponds to very challenging simulation scenarios, both statistically and computationally, due to the large latent dimensions. The results are presented in Table 3 and Fig. 2. We consider the conditions $(K, P) \in \{(15, 50), (25, 80), (35, 100)\}$ and $r \in \{0, 0.25\}$. For all conditions, we set $N = 2000$ and $K_{\max} = 50$.

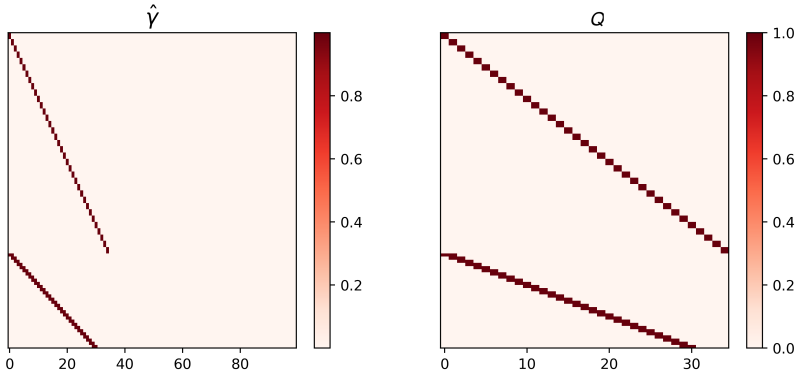

Figure 2: Example simulation run of variational inference procedure on large-scale data, demonstrating perfect recovery in a setting with typical levels of noise ($s = 0.2$, $g = 0.2$, $N = 3000$, $P = 100$, $K = 35$, $K_{\max} = 100$). *Left:* Estimated γ variational parameters. *Right:* Ground truth $\mathbf{Q}$ matrix.

K	$\widehat{K}_{est.}$	EAR	NOSE	K	$\widehat{K}_{est.}$	EAR	NOSE
$r = 0$				$r = 0.25$			
15	100%	100.00%	0	15	12%	94.30%	35.48
25	92%	100.00%	0.08	25	24%	97.50%	25.36
35	100%	100.00%	0	35	0%	98.80%	6.333

Table 3: Results for the setting with a large number of latent attributes K.

Failure of $\widehat{K}$ estimation under large K and attribute correlation. Table 3 reveals a clear limitation: when K is large and attributes are correlated ($r = 0.25$), the CSP dimension estimator $\widehat{K}$ degrades substantially, reaching 0% accuracy at $(K, r) = (35, 0.25)$ with NOSE = 6.333. Notably, the EAR remains high across all large-K, correlated settings (94.3%–98.8%), indicating that Q-matrix recovery per se does not fail; the method recovers the item-attribute relationships accurately but inflates the estimated number of attributes.

The mechanism underlying this failure is the interaction between attribute correlation and the CSP's column-wise evidence. The CSP identifies redundant columns by comparing $\Sigma_k = \sum_j \gamma_{jk}$ against a threshold: large Σ_k signals an active column, while small Σ_k signals redundancy. Under Assumption 1, the effective discrimination constant Ξ quantifies how well items separate students who have mastered attribute k from those who have not. When attributes are highly correlated, students with high ability profiles tend to master multiple attributes simultaneously, reducing the diversity of partial-mastery patterns across items. This shrinks the effective Ξ and compresses the evidence gap between active and redundant columns, making it harder for the CSP threshold to correctly separate them. At large K, this compression is amplified because many attributes share similar mastery patterns, causing several redundant columns to accumulate non-negligible slab evidence.

Practitioners encountering this regime should be aware that $\widehat{K}$ will tend to overestimate the true dimension when K is large and $r > 0$. Practical remedies include increasing N (larger samples sharpen the column-wise evidence), tightening the CSP spike parameter δ toward zero (making the prior more aggressive about suppressing low-evidence columns), or post-hoc pruning of estimated columns by applying a stricter threshold on Σ_k/P. We discuss

this as a known limitation and direction for future work in § 7.

Furthermore, the variational inference strategy is substantially faster than the Crimp sampler, with experiments for $K \leq 5$ running in less than 20 seconds on average on a laptop (Fig. 3, *left*). The dependence of run time on K scales linearly with the mean-field parameterization (Fig. 3, *right*) rather than exponentially like in previous studies that adopt a full parametrization (Yamaguchi and Okada, 2020a; Hijikata et al., 2023).

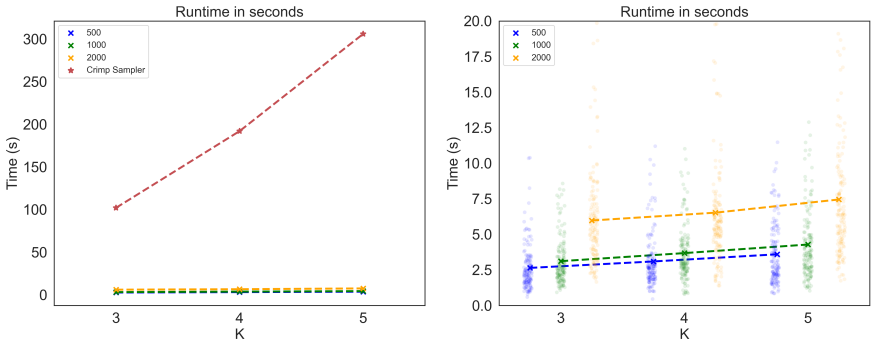

Figure 3: Run time of the proposed algorithm relative to Crimp sampler. *Left:* Run times for our algorithm are substantially shorter than the Crimp sampler across the same experimental contexts. *Right:* Run time increases slowly with both K and N.

Finally, we explore the effect of initialization with Varimax rotation by comparing it with random initialization. Across 100 independent trials and a range of sample sizes and latent dimensions, Varimax initialization consistently improves both negative log-likelihood and elementwise accuracy rate over random initialization (Supplementary Figure S.2). Supplement S.7 presents additional comparisons across initialization strategies, variational family specifications, and stochastic optimization variants. The mean-field variational family offers advantages not only in scalability but also in accuracy when the sample size is small relative to the complexity of more flexible parameterizations.

6 Real Data Applications

We illustrate the proposed method on two substantively different data settings, corresponding to the two probabilistic Boolean factorization structures studied in this paper. Section 6.1 focuses on an educational-assessment problem under the DINA model, where the Boolean

"And" assumption is natural and interpretability of the loading structure is central. Section 6.2 focuses on a modern single-cell genomics problem under the DINO model, where the Boolean "Or" assumption is more appropriate and high dimensionality is central. Taken together, these applications show that our method can be used for both interpretable structure refinement and scalable latent-structure discovery in distinct large-scale data settings.

6.1 Item Response Data in Large-scale Educational Assessments

The Trends in International Mathematics and Science Study (TIMSS), administered by the International Association for the Evaluation of Educational Achievement (IEA), is a large-scale international assessment measuring fourth and eighth-grade students' mathematics and science achievement across participating countries every four years since 1995. Cognitive diagnostic models such as the DINA model have been previously applied to TIMSS student response data to identify latent skill structures (Gu and Xu, 2023). We apply the proposed method to the TIMSS 2011 Austrian dataset containing students' responses to a set of math questions. Previous studies used subsets of the data for exploratory item-attribute $\mathbf{Q}$-matrix estimation with $N = 1010$ and $P = 47$; however, thanks to the scalability of the proposed algorithm, here we fit an exploratory DINA item-attribute matrix using the full dataset of $P = 174$ questions across $N = 4668$ students. Each student answered an average of 25 questions, leading to a total number of 115983 binary observations of Y_{ij}. We handle the missing data similarly as done in Gu and Xu (2023) for the DINA model, by writing the joint likelihood only over the observed entries in the $N \times P$ data matrix under the ignorable missingness assumption. All the coordinate ascent steps in the proposed algorithm can be easily adapted to this case. We initialize the item-attribute matrix using an expert-derived item-attribute matrix (Fig. 4, *left*), which is available from the TIMSS study design.

The expert-defined item-attribute Q-matrix consists of 9 columns corresponding to 9 expert-defined latent attributes based on the pairwise interactions between three content skills: data (D), geometry (G), numbers (N), and three cognitive skills: knowing (K), apply-

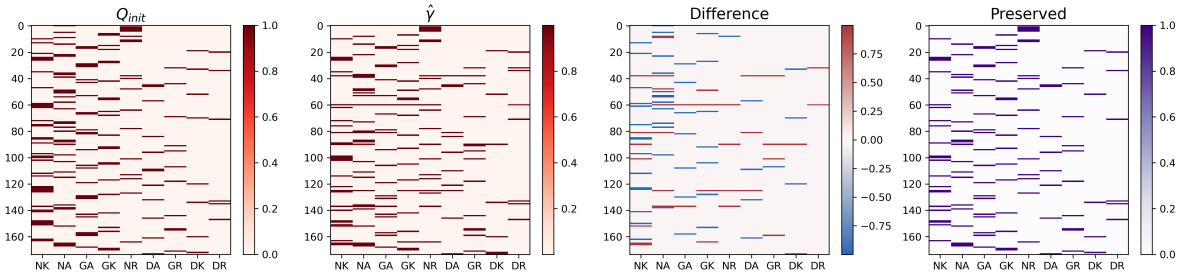

Figure 4: TIMSS 2011 item-attribute matrix after refinement by variational inference estimation algorithm. *Left:* Original expert item-attribute Q-matrix. *Middle left:* Refined item-attribute matrix. *Middle right:* Difference between the original one and refined one. *Right*: Preserved entries in the refined matrix.

ing (A), and reasoning (R). So, the 9 expert-defined attributes are DK, DA, DR, GK, GA, GR, NK, NA, NR. Each row in the expert Q-matrix is a canonical basis vector, meaning that each item requires exactly one of the nine possible skill combinations. Without using any information of this expert Q-matrix including its number of columns, we apply the proposed algorithm by setting a maximum number of latent attributes $K_{\max} = 35$. Interestingly, our estimated number of latent attributes is equal to $K = 9$, matching the expert defined latent dimension. In total, 5.5% of the entries in the original 174×9 item-attribute matrix are modified by our estimator, where 3.3% involve an entry modified from 1 to 0, while 2.2% involve an entry modified from 0 to 1 (Fig. 4, *middle right*)). In the estimated $\mathbf{Q}$ (Q-matrix), DR (*data reasoning*) was the least modified latent attribute with 1.1% of item-attribute matrix entries changing, while NK (*numerical knowing*) was the most modified one with 12.6% of item-attribute matrix entries changing (Fig. 4, *right*). We observe that a single question, M031185, is fully reassigned to be measuring a different latent attribute–it is moved from NR to NA. This question involves converting distances on a map to their respective true distances by multiplying by a scale factor. It is reasonable that this question requires the *applying* cognitive domain. This justifies the interpretability of our estimation result.

Items requiring multiple latent skills. In the expert-defined item-attribute matrix $\mathbf{Q}_{\text{expert}}$, each question requires only a single skill to answer. Hence there is a total of 174

positive entries for the 174 questions. After applying the DINA-CSP algorithm, 16 questions are found to require multiple skills, see Supplementary Table S.3. These questions generally have the highest difficulty and the fewest number of students answering them correctly (Supplementary Figure S.1, *left*). These items are sensibly found to require additional skills. For example, item M041284 involves classifying shapes into a table based on the number of sides and whether the sides have the same length and is estimated by our algorithm to require both attributes GR and DR; the organization of data into a table makes this question solvable mainly by students that are proficient in the *data reasoning* skill. Another example is that a geometry question, M031297, which involves computing the area of a shaded region, is found to also require NA in addition to GA, possibly due to the computation with fractions required to arrive at the correct answer. Finally, question M031016 involves writing all integers between 1 and 3000 that end in 112 and is found to require both NR and NA attributes. We also observe that students with the highest number of latent attributes, as measured by the estimate $\sum_{k=1}^{K} \alpha_{ik}$, score highest on the exam (Supplementary Figure S.1, *right*). This reflects that the estimated latent attributes indeed capture the target skills the exam questions are designed to measure.

6.2 Single-cell ATACseq Data

To demonstrate the scalability and wide-applicability of our approach beyond educational testing data, we also apply our method to single cell assay for transposase accessible chromatin sequencing (ATACseq) data (Buenrostro et al., 2015; Grandi et al., 2022). Single-cell ATACseq measures how "open" a region of DNA is via the ability of a hyperactive Tn5 transposase to insert sequencing adapters in the DNA. Functional regions of DNA (e.g. sequences acting as promoters or enhancers) can be identified via the enrichment of reads mapping to a particular stretch of the genome. After processing the raw sequencing reads, common practice is to produce a cells-by-peaks count matrix, which counts the number of reads per cell that coincide with each peak. Peaks are typically defined as 500 base pair segments of

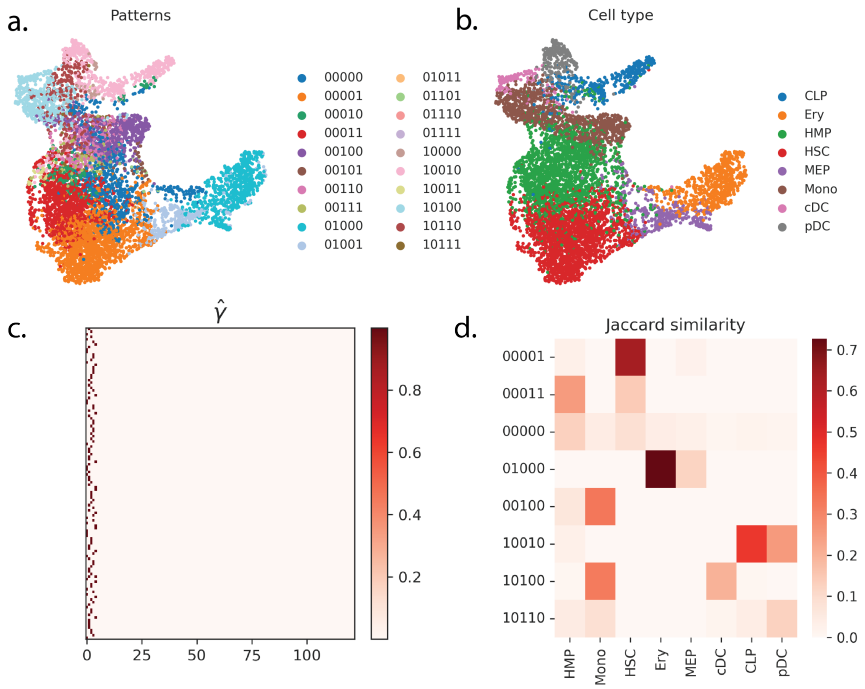

Figure 5: *(a.)* 2D projection (UMAP) of cells colored by latent binary pattern α_i. *(b.)* 2D projection (UMAP) of cells colored by cell type annotation *(c.)* Estimated item-attribute matrix, γ, showing the inferred dimension K is equal to 5. *(d.)* Pairwise Jaccard similarity between most common latent binary patterns and cell type labels.

DNA with a high density of sequencing reads, as determined by a peak calling algorithm such as MACS2 (Gaspar, 2018; Granja et al., 2021).

The DINO model with the Boolean "Or" assumption is particularly well suited to describe the single-cell ATAC-seq cells-by-peaks count matrix. First, the data itself is inherently binary; canonically, a peak that represents a functional region of the genome has two states: "open" (representing an active state) and "closed" (representing an inactive state). Second, a single genomic location can be active as part of multiple contexts or cell states; for example, many enhancers bound by STAT transcription factors are reused across multiple immune cell types (Yoshida et al., 2019). Thus, an "or" relationship is natural to describe the underlying process that determines a given peak's binary state. This disjunctive structure, where a peak is accessible if activated by *any* relevant regulatory program, has been explicitly modeled in probabilistic Boolean logic approaches for transcription factor activity inference from single-

cell genomics data (Arriojas et al., 2023). To demonstrate the utility of our approach, we fit a DINO model to a single cell ATACseq dataset (Persad et al., 2023) of CD34+ bone marrow cells collected from healthy human donors. These cells are the stem and progenitor cells that differentiate into various blood cell types, including erythrocytes, B cells, T cells, monocytes and dendritic cells (Fig. 5, b). Under the DINO model, the ideal response matrix entries become: $J_{ij} := 1 - \prod_{k=1}^{K}(1 - A_{ik}Q_{jk})$. This model is mathematically equivalent to the DINA model after a certain reparameterization and can be fit by our VI algorithm.

Using the preprocessed data of the original paper, we have a binarized count matrix of 6881 cells and 246113 peaks. Most peaks are non-informative, so for each cell type, we take the 16 most informative features, leading to a dataset with $N = 6656$ cells (after removing 225 low-coverage cells; see Supplementary Material for details) and $P = 122$ features. We fit the DINO model using CAVI-CSP with $K_{\max} = P = 122$. The estimated latent dimension $\widehat{K}$ is 5 (Fig. 5, c). Fig. 5(a-d) show that the learned binary patterns have high agreement with the cell type annotations, which are held-out biological knowledge. In addition, the discovered binary patterns provide even more fine-grained subgrouping of the cells. An interpretable feature of the result is that we can track how related cell types are based on their binary patterns. For example, monocytes are represented by two binary patterns, 10100 and 00100, and HSCs are represented by two patterns 00001 and 00011. HMPs and HSCs are the most stem-like cell populations, and are both found to express the 5th binary latent (Fig. 5, d). Additionally, the 3rd binary latent is specifically expressed by dendritic cells and monocytes, and hence is specific to myeloid identity cells. These biologically interpretable results imply that it would be promising to further look into each learned latent dimension to interpret their unique biological meanings.

7 Discussion

We have developed a scalable variational inference framework for probabilistic Boolean matrix factorization with unknown latent dimension, using the DINA/DINO family as a struc-

tured and theoretically tractable modeling class. Relative to latent-class formulations whose variational dimension grows exponentially with K, the proposed mean-field formulation uses $\mathcal{O}(K)$ variational parameters and, with suitable initialization, attains competitive empirical performance at substantially lower computational cost. In addition, we establish theoretical guarantees in the large-scale and high-dimensional regime. The sample complexity $\tilde{\Omega}(1/(\epsilon\Xi\min(\epsilon,\rho)))$ in Theorem 1 has a clear interpretation: $\epsilon = \frac{1}{2} - s$ is the signal level, $\rho = \log\frac{1-s}{s}$ is the log-odds discriminability, and Ξ is the minimum fraction of identifiable rows or columns under Assumptions 1 and 2. When the signal is weak ($\epsilon \to 0$) or the identifiable fraction Ξ is small, recovery becomes statistically harder and requires more data. The minimax lower bound in Theorem 2 shows that no estimator can improve this rate beyond logarithmic factors. Our current analysis assumes $s_j = g_j$ for all j; extending the guarantees to general per-item slipping and guessing rates is a natural direction for future work. Extensive simulations and two substantively different real-data applications further demonstrate the effectiveness of the proposed method.

The modeling and variational ideas developed here may also be useful for other structured binary latent-feature models, including probabilistic Boolean factorization models with alternative logic gates, more general discrete latent-variable models, and multilayer binary latent-structure models. Cognitive diagnostic models that involve main effects and other interaction effects of latent attributes form one important subclass of this broader family (von Davier, 2008; Henson et al., 2009; de la Torre, 2011; von Davier and Lee, 2019). Similar ideas may also be relevant for deep discrete latent-variable models that require selecting the number of latent variables in each layer, including Bayesian pyramids for multivariate categorical data proposed by Gu and Dunson (2023) and deep discrete encoders for rich data types with many discrete latent layers proposed by Lee and Gu (2025).

A distinctive feature of the proposed variational inference algorithm is that the mean-field approximation and the factorized Boolean likelihood yield closed-form coordinate-wise

updates. By contrast, many recent approaches to variational inference with discrete latent variables rely on approximate gradient estimators and simultaneous stochastic-gradient updates. Some methods reduce variance at the cost of introducing bias in the gradient estimates (Jang et al., 2016), while others preserve unbiasedness using the score-function estimator and mitigate variance through techniques such as control variates (Titsias and Shi, 2022), antithetic sampling (Yin and Zhou, 2019; Kunes et al., 2023), and Rao–Blackwellization (Dong et al., 2020). Future work may combine such ideas with probabilistic BMF or extend the proposed method to block coordinate-ascent schemes for improved scalability.

Data Availability Statement. The data supporting the findings of this study are derived from public domain resources. Reproducibility materials, including code and supporting files used in the analysis, will be made publicly available upon publication.

References

Arriojas, A., Patalano, S., Macoska, J., and Zarringhalam, K. (2023). A bayesian noisy logic model for inference of transcription factor activity from single cell and bulk transcriptomic data. *NAR Genomics and Bioinformatics*, 5(4):lqad106.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1):77–120.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*.

Chen, Y., Culpepper, S. A., Chen, Y., and Douglas, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika*, 83:89–108.

Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*.

Chen, Y., Liu, Y., Culpepper, S. A., and Chen, Y. (2021). Inferring the number of attributes for the exploratory DINA model. *Psychometrika*, 86(1):30–64.

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics*, 40(5):454–476.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1):115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76:179–199.

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*.

Dong, Z., Mnih, A., and Tucker, G. (2020). DisARM: An antithetic gradient estimator for binary latent variables. *Advances in Neural Information Processing Systems*.

Dzyabura, D. and Hauser, J. R. (2011). Active machine learning for consideration heuristics. *Marketing Science*, 30(5):801–819.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Gaspar, J. M. (2018). Improved peak-calling with macs2. *BioRxiv*, page 496521.

Grandi, F. C., Modi, H., Kampman, L., and Corces, M. R. (2022). Chromatin accessibility profiling by ATAC-seq. *Nature protocols*, 17(6):1518–1552.

Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y., and Greenleaf, W. J. (2021). Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53(3):403–411.

Gu, Y. and Dunson, D. B. (2023). Bayesian pyramids: Identifiable multilayer discrete latent structure models for discrete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):399–426.

Gu, Y. and Xu, G. (2021). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinica*, 31:449–472.

Gu, Y. and Xu, G. (2023). A joint MLE approach to large-scale structured latent attribute analysis. *Journal of the American Statistical Association*, 118(541):746–760.

Haddad, A., Shamsi, F., Zhu, L., and Najafizadeh, L. (2018). Identifying dynamics of brain function via Boolean matrix factorization. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 661–665. IEEE.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*.

Hijikata, K., Oka, M., Yamaguchi, K., and Okada, K. (2023). variationalDCM: An R package for variational Bayesian inference in diagnostic classification models.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*.

Kunes, R. Z., Yin, M., Land, M., Haviv, D., Pe'er, D., and Tavaré, S. (2023). Gradient estimation for binary latent variables via gradient variance clipping. In *AAAI*.

Lee, S. and Gu, Y. (2025). Deep discrete encoders: Identifiable deep generative models for rich data with discrete latent layers. *Journal of the American Statistical Association*.

Legramanti, S. (2020). Variational bayes for gaussian factor models under the cumulative shrinkage process. In *Book of Short Papers, SIS 2020*, pages 416–420. SIS = Italian Statistical Society.

Legramanti, S., Durante, D., and Dunson, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, 107(3):745–752.

Liang, L., Zhu, K., and Lu, S. (2020). BEM: mining coregulation patterns in transcriptomics via Boolean matrix factorization. *Bioinformatics*, 36(13):4030–4037.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Michalis, T. and Lázaro-Gredilla, M. (2015). Local expectation gradients for black box variational inference. *Advances in neural information processing systems*, 28.

Miettinen, P. and Neumann, S. (2021). Recent developments in Boolean matrix factorization. In *IJCAI*.

Miettinen, P. and Vreeken, J. (2011). Model order selection for Boolean matrix factorization. In *ACM SIGKDD*.

Oka, M. and Okada, K. (2023). Scalable Bayesian approach for the DINA Q-matrix estimation combining stochastic optimization and variational inference. *Psychometrika*.

Persad, S., Choo, Z.-N., Dien, C., Sohail, N., Masilionis, I., Chaligné, R., Nawy, T., Brown, C. C., Sharma, R., Pe'er, I., et al. (2023). Seacells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nature Biotechnology*.

Ročková, V. and George, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622.

Rohe, K. and Zeng, M. (2023). Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society Series B*.

Rukat, T., Holmes, C. C., Titsias, M. K., and Yau, C. (2017a). Bayesian Boolean matrix factorisation. In *ICML*.

Rukat, T., Lange, D., and Archambeau, C. (2017b). An interpretable latent variable model for attribute applicability in the Amazon catalogue. *arXiv*.

Rupp, A. A. and Templin, J. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, 68(1):78–96.

Rupp, A. A. and Templin, J. L. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*.

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20:345–354.

Templin, J. L. and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11(3):287.

Titsias, M. and Shi, J. (2022). Double control variates for gradient estimation in discrete latent variable models. In *Artificial Intelligence and Statistics*.

Tucker, G., Mnih, A., Maddison, C. J., Lawson, J., and Sohl-Dickstein, J. (2017). Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. *Advances in Neural Information Processing Systems*, 30.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61:287–307.

von Davier, M. and Lee, Y.-S. (2019). Handbook of diagnostic classification models. *Cham: Springer International Publishing*.

Wan, C., Chang, W., Zhao, T., Li, M., Cao, S., and Zhang, C. (2020). Fast and efficient Boolean matrix factorization by geometric segmentation. In *AAAI*.

Xu, G. and Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, 113(523):1284–1295.

Yamaguchi, K. (2020). Variational Bayesian inference for the multiple-choice DINA model. *Behaviormetrika*, 47(1):159–187.

Yamaguchi, K. and Okada, K. (2020a). Variational Bayes inference algorithm for the saturated diagnostic classification model. *Psychometrika*, 85(4):973–995.

Yamaguchi, K. and Okada, K. (2020b). Variational Bayes inference for the DINA model. *Journal of Educational and Behavioral Statistics*, 45(5):569–597.

Yin, M., Wang, Y. R., and Sarkar, P. (2020). A theoretical case study of structured variational inference for community detection. In *International conference on artificial intelligence and statistics*, pages 3750–3761. PMLR.

Yin, M. and Zhou, M. (2019). Arm: Augment-reinforce-merge gradient for stochastic binary networks. In *International Conference on Learning Representations*.

Yoshida, H., Lareau, C. A., Ramirez, R. N., Rose, S. A., Maier, B., Wroblewska, A., Desland, F., Chudnovskiy, A., Mortha, A., Dominguez, C., et al. (2019). The cis-regulatory atlas of the mouse immune system. *Cell*, 176(4):897–912.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. (2016). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102):1–44.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.