PSYCHOMETRIKA—VOL. 88, NO. 2, 361–386
JUNE 2023
https://doi.org/10.1007/s11336-023-09904-x

IDENTIFIABILITY OF HIDDEN MARKOV MODELS FOR LEARNING TRAJECTORIES IN COGNITIVE DIAGNOSIS

YING LIU, STEVEN ANDREW CULPEPPER AND YUGUO CHEN

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Hidden Markov models (HMMs) have been applied in various domains, which makes the identifiability issue of HMMs popular among researchers. Classical identifiability conditions shown in previous studies are too strong for practical analysis. In this paper, we propose generic identifiability conditions for discrete time HMMs with finite state space. Also, recent studies about cognitive diagnosis models (CDMs) applied first-order HMMs to track changes in attributes related to learning. However, the application of CDMs requires a known $\boldsymbol{Q}$ matrix to infer the underlying structure between latent attributes and items, and the identifiability constraints of the model parameters should also be specified. We propose generic identifiability constraints for our restricted HMM and then estimate the model parameters, including the $\boldsymbol{Q}$ matrix, through a Bayesian framework. We present Monte Carlo simulation results to support our conclusion and apply the developed model to a real dataset.

Key words: cognitive diagnosis model, DINA model, generic identifiability, hidden Markov model.

Hidden Markov models (HMMs) are widely known for their applications to finance (Sipos, Ceffer, & Levendovszky, 2017), signal processing (Crouse, Nowak, & Baraniuk, 1998), sequence classification (Blasiak & Rangwala, 2011) and many more, partially because HMMs offer inferences about substantively important unobserved states. Recently, HMMs are becoming popular in psychology and education research to model learning trajectories. In particular, researchers in education and psychology use restricted HMMs for discrete data in which the emission probability, which is the conditional probability of a response given the contemporaneous hidden state, is formed by a restricted latent class model (RLCM; Xu, 2017). RLCMs are designed for diagnostic research settings by imposing structure on the emission matrix to infer a collection of underlying skills and attributes. The goal of these applications is to classify respondents according to substantively important latent attribute profiles. Although several studies developed new methods for inferring restricted HMM parameters, issues such as model identifiability are important to examine for restricted HMMs in order to ensure the feasibility of recovering model parameters and conducting diagnostic inferences. In this paper, we focus on discrete time HMMs.

There has been considerable research on the identifiability of HMMs. For an HMM with finite observable and hidden states, prior research (Baras & Finesso, 1992, Lemma 1.2.4) provided an identifiability condition which shows that the distribution of an HMM with r hidden states and κ observable states can be completely determined if the marginal distribution of $2r$ consecutive observed variables is known. Paz (1971) proposed a stronger result in Corollary 3.4: the marginal distribution of $2r-1$ consecutive observed variables uniquely determines the whole HMM distribution. Bonhomme, Jochmans, and Robin (2016) established identifiability of finite HMMs, which include finite observed and hidden states, by imposing restrictions on the structure of the transition and emission matrices, and linked the identifiability problem with the decomposition of a multiway array and simultaneously diagonalizing a collection of matrices. By applying the uniqueness theorem pointed out by Lathauwer, Moor, and Vandewalle (2004), Bonhomme et

Correspondence should be made to Steven Andrew Culpepper, Department of Statistics, University of Illinois at Urbana-Champaign, Computing Applications Building, Room 152, 605 E. Springfield Ave., Champaign, IL 61820, USA. Email: sculpepp@illinois.edu

al. (2016) derived sufficient conditions for this class of finite HMMs to be identifiable up to a permutation of states. Compared with previous results, the identifiability conditions presented in Bonhomme et al. (2016) are easier to verify in practice.

The classical model identifiability studied by the above research is referred to as *strict identifiability*, which guarantees distinct set of parameters correspond to different values for the likelihood function, but these conditions may be too strong in practice. Allman, Matias, and Rhodes (2009) defined *generic identifiability* as "all nonidentifiable parameter choices lie within a proper subvariety, and thus form a set of Lebesgue measure zero," which is enough for practical data analysis purposes. Petrie (1969) identified conditions that the transition matrix and the emission matrix should satisfy to ensure the generic identifiability of HMMs with finite observable and hidden states. Allman et al. (2009) provided a generic identifiable condition for HMMs by applying the fundamental algebraic result in Kruskal (1977) and demonstrated that an HMM with r hidden states and κ observable states are generically identifiable if the marginal distribution of $2k+1$ consecutive observed variables is known, where k satisfies $\binom{k+\kappa-1}{\kappa-1} \geq r$.

The purpose of our paper is to contribute to the body of research on the identifiability of HMMs with particular focus on conditions for identifying parameters of restricted HMMs. The identifiability conditions we propose for restricted HMMs are of interest to wide applications in psychological and educational studies. The main difference between the conventional (i.e., unrestricted) HMMs discussed above and restricted HMMs is that the restricted HMM model parameters are constrained by a latent binary matrix. Consequently, the parameter space falls into a measure zero set with respect to the whole parameter space of an unrestricted model as discussed in Allman et al. (2009), so identifiability conditions mentioned above for conventional HMMs cannot be directly applied to our restricted HMMs.

In this paper, we consider both conventional and restricted HMMs with a time-invariant emission matrix, which governs the probability of observed variables given hidden states. We propose strict and generic identifiability conditions for conventional HMMs with finite observable and hidden states by generalizing the result in Bonhomme et al. (2016). For restricted HMMs, we extend static RLCMs for a single time point by proposing a restricted HMM framework, which allow us to monitor the learning trajectory and uncover the underlying structure between items and skills. We propose strict and generic identifiability conditions for model parameters using both a deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) and the general RLCM (De La Torre, 2011; Henson, Templin, & Willse, 2009; Von Davier, 2008) measurement models. The HMMs discussed in our paper contain finite hidden states, so the identifiability is defined up to a permutation on the state labels. Furthermore, we also propose a Bayesian formulation and present an algorithm for inferring parameters for the popular DINA model restricted HMM (Chen, Culpepper, Wang, & Douglas, 2018).

The remainder of the paper is organized as follows. Section 1 introduces the setup of our two restricted HMMs. Section 2 describes the identifiability issue and shows the strict and generic identifiability conditions of our models. Section 3 proposes the Bayesian formulation for our restricted HMM and provides a summary of the algorithm for posterior inference. Section 4 reports results from a Monte Carlo simulation study demonstrating the accuracy of the proposed algorithm, and Section 5 includes results from a real data application. Lastly, Sect. 6 summarizes the contribution and limitations of this paper and proposes several future research topics. Proofs and other details about the real data are given in Appendices.

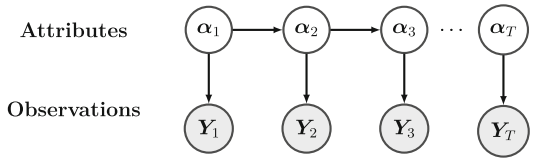

FIGURE 1.
Embedding the HMM into CDMs with learning trajectories.

1. Models and Applications

1.1. Model Setup

Consider a cognitive diagnosis model that allows attribute profiles to change over time. We discuss a model for multiple individuals $i = 1, \ldots, N$, and we suppress the i subscript for individuals in the following discussion to simplify notation. Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T)^\top$ be the learning trajectory of a subject, where $\boldsymbol{\alpha}_t = (\alpha_{1t}, \ldots, \alpha_{Kt})^\top$ represents the corresponding attribute profile at time t ($t = 1, \ldots, T$) and $\alpha_{kt} = 1$ indicates that the subject possesses the k-th attribute and 0 otherwise. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T)^\top$ denote the binary response data of the subject over time, where $\boldsymbol{Y}_t = (Y_{1t}, \ldots, Y_{Jt})^\top$ represents the responses of J items from the subject at time point t and let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)^\top$ be a realization of responses $\boldsymbol{Y}$. Figure 1 presents the HMM we employ to describe dependence in $\boldsymbol{Y}$ and trace the learning trajectory. Specifically, we consider designs with data collected over T time points and we model the dependence over time by introducing a first-order HMM for the association of underlying attributes at each time.

Consider a stationary Markov chain $\{\boldsymbol{\alpha}_t\}$ ($t = 1, 2, \ldots, T$) on state space $\{0, 1\}^K$ with $K < J$, a time-invariant transition matrix $\boldsymbol{\omega}$, and stationary distribution $\boldsymbol{\pi}$. For general HMMs, the responses $\{\boldsymbol{Y}_t\}$ ($t = 1, 2, \ldots, T$) are assumed to be conditionally independent given $\boldsymbol{\alpha}_t$. We let $\boldsymbol{B}$ denote the $2^J \times 2^K$ emission matrix, which contains the emission probabilities $P(\boldsymbol{Y}_t = \boldsymbol{y}_t|\boldsymbol{\alpha}_t)$, where the column entries are indexed by $\boldsymbol{\alpha}_t$ and rows correspond to response patterns $\boldsymbol{y}_t$.

1.2. Applications

Given attribute profile $\boldsymbol{\alpha}_t$, assuming that the binary responses $Y_{1t}, \ldots, Y_{Jt}$ are independent Bernoulli variables with parameter $\theta_{j,\boldsymbol{\alpha}_t}$, then the emission probability $P(\boldsymbol{Y}_t = \boldsymbol{y}_t|\boldsymbol{\alpha}_t)$ is

$$P(\boldsymbol{Y}_t = \boldsymbol{y}_t|\boldsymbol{\alpha}_t, \boldsymbol{\Omega}) = \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}_t}^{y_{jt}} (1 - \theta_{j,\boldsymbol{\alpha}_t})^{1-y_{jt}}, \tag{1}$$

where $\theta_{j,\boldsymbol{\alpha}_t} = P(Y_{jt} = 1|\boldsymbol{\alpha}_t, \boldsymbol{\Omega}_j)$ is the probability of correctly answering item j at time t for a subject with attribute profile $\boldsymbol{\alpha}_t$, and $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_J)$ represents model parameters which are fixed over time.

Equation (1) is an unrestricted model that includes 2^K latent class response probabilities for each item. In our application, we instead consider restricted models, which impose some structure on the $\theta_{j,\boldsymbol{\alpha}_t}$'s.

1.2.1. General Model Chen, Liu, Xu, and Ying (2015) proposed a general alternative representation of CDMs

$$\theta_{j,\boldsymbol{\alpha}_t} = P(Y_{jt} = 1|\boldsymbol{\alpha}_t, \boldsymbol{\beta}_j) = \Psi(\boldsymbol{\alpha}_t^{*\top} \boldsymbol{\beta}_j), \tag{2}$$

where $\Psi(\cdot)$ is a cumulative distribution function (CDF), which we assume is a positive and strictly increasing function in this paper, and

$$\boldsymbol{\alpha}_t^* = \left(1, \alpha_{1t}, \ldots, \alpha_{Kt}, \alpha_{1t}\alpha_{2t}, \ldots, \alpha_{K-1,t}\alpha_{Kt}, \ldots, \prod_{k=1}^{K}\alpha_{kt}\right) \tag{3}$$

is a 2^K-dimensional alternative representation of the binary vector $\boldsymbol{\alpha}_t$ with all the interactions, and

$$\boldsymbol{\alpha}_t^{*\top}\boldsymbol{\beta}_j = \beta_{j,0} + \sum_{k=1}^{K}\beta_{j,k}\alpha_{kt} + \sum_{k>k'}\beta_{j,kk'}\alpha_{kt}\alpha_{k't} + \ldots + \beta_{j,12\ldots K}\prod_{k=1}^{K}\alpha_{kt}. \tag{4}$$

Here $\boldsymbol{\beta}_j$ is a sparse vector of coefficients, in which the nonzero elements represent the impact of latent skills or combinations of skills on the response of item j. The sparsity of $\boldsymbol{\beta}_j$ is represented by a binary vector $\boldsymbol{\gamma}_j \in \{0,1\}^{2^K}$, with 1 implying that the corresponding coefficient is nonzero (active) and 0 implying that the corresponding coefficient is zero (inactive). The intercept $\beta_{j,0}$ is usually assumed to be active.

For the general model, we use $\boldsymbol{\beta}_{J\times 2^K} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)^\top$ to represent the coefficient matrix and $\boldsymbol{\Gamma}_{J\times 2^K} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J)^\top$ to represent its sparsity. Many popular CDMs can be reparameterized from the general model. Next, we introduce the DINA model (Junker & Sijtsma, 2001) as a special case of the general model.

1.2.2. DINA Model We introduce a $J \times K$ binary matrix, $\boldsymbol{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_J)^\top \in \{0,1\}^{J\times K}$, which defines the underlying structure between latent skills and items. Here $\boldsymbol{q}_j^\top = (q_{j1}, \ldots, q_{jK})$ is the j-th row of the $\boldsymbol{Q}$ matrix and $q_{jk} = 1$ indicates item j requires the mastery of skill k and 0 otherwise. In short, RLCMs impose sparsity and are generally more parsimonious than unrestricted models. In this paper, we deploy the DINA model (Junker & Sijtsma, 2001) with parameters $\boldsymbol{\Omega}_j = (\boldsymbol{q}_j, s_j, g_j)$. The probability of a correct response is

$$\theta_{j,\boldsymbol{\alpha}_t} = P(Y_{jt} = 1|\boldsymbol{\alpha}_t, \boldsymbol{q}_j, s_j, g_j) = (1-s_j)^{\eta_{jt}} g_j^{1-\eta_{jt}}, \tag{5}$$

where if let $\mathcal{I}(\cdot)$ denote the indicator function, then $\eta_{jt} = \mathcal{I}\left(\boldsymbol{\alpha}_t^\top \boldsymbol{q}_j \geq \boldsymbol{q}_j^\top \boldsymbol{q}_j\right)$ corresponds to an "and" logic gate that equals one if the subject mastered the required attributes for item j at time t and zero if at least one required attribute was not mastered. Additionally, $\boldsymbol{s}^\top = (s_1, \ldots, s_J)$ and $\boldsymbol{g}^\top = (g_1, \ldots, g_J)$ represent slipping and guessing parameters such that $s_j = P(Y_{jt} = 0|\eta_{jt} = 1)$ and $g_j = P(Y_{jt} = 1|\eta_{jt} = 0)$. Further details about the DINA model are given in Sect. 3.

2. Identifiability

2.1. Identifiable HMMs

As introduced in the previous section, the stationary distribution of attribute profiles is given by the vector $\boldsymbol{\pi} = (\pi_c)^\top \in [0,1]^{2^K}$ with $\sum \pi_c = 1$. Transition matrix $\boldsymbol{\omega}$ is a $2^K \times 2^K$ matrix of first-order transition probabilities between different states. For any time $t > 1$, let $\omega_{c'|c} = P(\boldsymbol{\alpha}_t^\top \boldsymbol{v}_\alpha = c' \mid \boldsymbol{\alpha}_{t-1}^\top \boldsymbol{v}_\alpha = c)$ denote the (c, c') element in $\boldsymbol{\omega}$, which represents the probability

of transitioning from state c to c' between any two consecutive time points, where the vector $\boldsymbol{\upsilon}_\alpha = (2^{K-1}, \ldots, 1)^\top$ is used to create a bijection between the binary attributes $\boldsymbol{\alpha}_t$ and integers $c, c' \in \{0, \ldots, 2^K - 1\}$.

In the context of HMMs, we denote the parameter space of $(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ by

$$\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}) = \{(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}) : \boldsymbol{\pi} \in \boldsymbol{\Omega}(\boldsymbol{\pi}), \boldsymbol{\omega} \in \boldsymbol{\Omega}(\boldsymbol{\omega}), \boldsymbol{B} \in \boldsymbol{\Omega}(\boldsymbol{B})\}, \tag{6}$$

where $\boldsymbol{\Omega}(\boldsymbol{\pi}) = \{\boldsymbol{\pi} \in [0,1]^{2^K} : \sum_c \pi_c = 1\}$, $\boldsymbol{\Omega}(\boldsymbol{\omega}) = \{\boldsymbol{\omega} \in [0,1]^{2^K \times 2^K} : \sum_{c'} \omega_{c'|c} = 1,\ c = 0, \ldots, 2^K - 1\}$ and $\boldsymbol{\Omega}(\boldsymbol{B}) = \{\boldsymbol{B} \in [0,1]^{2^J \times 2^K} : \sum_i B_{ij} = 1,\ j = 1, \ldots, 2^K\}$.

We next discuss model identifiability by considering two discrete time HMMs parameterized by two parameter values $(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ and $(\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{B}})$.

Definition 1. *(Strict Identifiability)* The parameters $(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}) \in \boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ are identifiable when

$$P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}) = P(\boldsymbol{Y} = \boldsymbol{y} \mid \bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{B}}) \text{ if and only if } (\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}) \sim (\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{B}}),$$

where $(\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{B}})$ is another value from the parameter space $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ and "$\sim$" means two parameter values are equivalent up to a permutation of hidden states.

2.2. Generic Identifiability

The identifiability introduced in Definition 1 is referred to as strict identifiability, which could be too restrictive in practice. A weaker notion of identifiability is referred to as generic identifiability, which was first introduced in Allman et al. (2009). Generic identifiability allows the existence of some exceptional values of parameters for which strict identifiability does not hold, however, all non-identifiable parameters should form a Lebesgue measure zero set. Since non-identifiable parameters live in a set of measure zero, one is unlikely to face identifiability problems in performing inference. Thus, generic identifiability is generally sufficient for data analysis purposes. For instance, Allman et al. (2009) showed that the generic identifiability condition requires a fewer number of consecutive observed variables to completely determine the distribution of an HMM in comparison to the strict identifiability condition.

Let $\boldsymbol{\Delta}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ denote the set of non-identifiable parameters from $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$:

$$\begin{aligned}\boldsymbol{\Delta}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}) = \{(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}) :\ & P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}) = P(\boldsymbol{Y} = \boldsymbol{y} \mid \bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{B}}) \text{ for some} \\ & (\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{B}}) \not\sim (\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}),\ (\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}) \in \boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}),\ (\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{B}}) \in \boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})\}.\end{aligned} \tag{7}$$

Based on the definition of generic identifiability, if the non-identifiable parameter set $\boldsymbol{\Delta}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ is of measure zero within parameter space $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$, then we say $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ is a generically identifiable parameter space.

Definition 2. (*Generic Identifiability*) The parameter space $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ is generically identifiable, if the Lebesgue measure of $\boldsymbol{\Delta}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ with respect to parameter space $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ is zero.

2.3. Identifiability Conditions

In this section, we discuss identifiability conditions for both conventional HMMs and restricted HMMs. We start with strict and generic identifiability conditions for conventional HMMs. Consider the bipartition of the set $\mathbb{J} = \{1, 2, \ldots, J\}$ into two disjoint, nonempty subsets $\mathbb{J}_1 = \{1, 2, \ldots, K\}$, $\mathbb{J}_2 = \{K+1, \ldots, J\}$. Then for $t = 1, \ldots, T$, let $\boldsymbol{Y}_t = (\boldsymbol{Y}_t^{\mathbb{J}_1\top}, \boldsymbol{Y}_t^{\mathbb{J}_2\top})^\top$, where $\boldsymbol{Y}_t^{\mathbb{J}_1} = (Y_{1t}, \ldots, Y_{Kt})^\top$ and $\boldsymbol{Y}_t^{\mathbb{J}_2} = (Y_{(K+1)t}, \ldots, Y_{Jt})^\top$. Let $\boldsymbol{B}^{\mathbb{J}_1}$ and $\boldsymbol{B}^{\mathbb{J}_2}$ be the emission matrices for $\boldsymbol{Y}_t^{\mathbb{J}_1}$ and $\boldsymbol{Y}_t^{\mathbb{J}_2}$, respectively, given values of attribute profile $\boldsymbol{\alpha}_t$, and we have $\boldsymbol{B} = \boldsymbol{B}^{\mathbb{J}_1} \odot \boldsymbol{B}^{\mathbb{J}_2}$, where $\odot$ represents the Khatri–Rao, column-wise tensor product defined in Definition 4, Appendix B.

Theorem 1. *(*Strict Identifiability for HMMs*) Any parameter* $(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ *from* $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ *in the HMM is identifiable if* $rank(\boldsymbol{\omega}) = 2^K$, $\pi_c > 0$ *for all c, and*

(a) for $T \geq 3$, $rank(\boldsymbol{B}) = 2^K$ *(Bonhomme et al., 2016);*
(b) for $T = 2$, $rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K$ *and* $rank_K(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2$.

Note $rank_K$ in part (b) of Theorem 1 denotes the Kruskal rank (see Definition 5 in Appendix C). Proof of part (b) is shown in Appendix C. Based on the above result, we propose generic identifiability conditions for conventional HMMs up to a permutation of hidden states.

Theorem 2. *(*Generic Identifiability for HMMs*) For an HMM, the parameter space* $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B})$ *is generically identifiable if* $\pi_c > 0$ *for all c, and*

(a) for $T \geq 3$, $rank(\boldsymbol{B}) = 2^K$;
(b) for $T = 2$, $rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K$ *and* $rank_K(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2$.

Proof is found in Appendix A (for part (a)) and C (for part (b)).

Remark 1. For conclusions for $T = 2$ case in Theorems 1 and 2, we do not require the attribute profiles sequence to have a stationary distribution. We only need positive initial probabilities for attribute profiles at time point $t = 1$. Proof is given in Appendix G.

Theorem 2 provides generic identifiability conditions for unrestricted HMMs. However, Theorem 2 is not applicable to our setting with a restricted HMM. The parameter space shown in Eq. (6) corresponds to an unrestricted HMM, whereas in our restricted HMMs, the parameter space of the emission matrix $\boldsymbol{B}$ is restricted by some binary structures. For the DINA model, $\boldsymbol{B}$ is restricted by the structure of the $\boldsymbol{Q}$ matrix, while for the general model, $\boldsymbol{B}$ is restricted by the structure of the $\boldsymbol{\Gamma}$ matrix.

2.3.1. General Model We use $\boldsymbol{\beta}_{J\times 2^K}$ to represent the coefficient matrix and $\boldsymbol{\Gamma}_{J\times 2^K}$ to represent its sparsity. For a given sparsity structure $\boldsymbol{\Gamma}$, the parameter space of the restricted HMM is

$$\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\omega}) = \{(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\omega}) : \boldsymbol{\pi} \in \boldsymbol{\Omega}(\boldsymbol{\pi}), \boldsymbol{\omega} \in \boldsymbol{\Omega}(\boldsymbol{\omega}), \boldsymbol{\beta} \in \boldsymbol{\Omega}_{\boldsymbol{\Gamma}}(\boldsymbol{\beta})\}, \tag{8}$$

where $\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}(\boldsymbol{\beta})$ represents the set of coefficient matrices that only have nonzero elements at positions where the corresponding elements in $\boldsymbol{\Gamma}$ are 1.

The following three conditions are needed in Theorems 3 and 4 for the identifiability of restricted HMMs:

($B1$) The true sparsity matrix $\boldsymbol{\Gamma}$ takes the form of $\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{D} \\ \boldsymbol{\Gamma}' \end{pmatrix}_{J \times 2^K}$ after row swapping, where $\boldsymbol{\Gamma}'$ is a $(J - K) \times 2^K$ binary matrix and $\boldsymbol{D}$ is a $K \times 2^K$ binary matrix with the following structure

$$\boldsymbol{D} = \begin{pmatrix} 1\,1\,0 \ldots 0 \ldots 0 \\ 1\,0\,1 \ldots 0 \ldots 0 \\ \vdots\ \vdots \quad \ddots \quad \vdots \\ 1\,0\,0 \ldots 1 \ldots 0 \end{pmatrix}.$$

($B2$) For any two attribute profiles, there exists at least one item associated with a row in $\boldsymbol{\Gamma}'$ such that they have different emission probabilities.

Remark 2. Condition (B1) for $\boldsymbol{\Gamma}$ can be interpreted as requiring that the coefficients $\beta_{k,k}$ are nonzero for $k = 1, \ldots, K$.

Remark 3. We do not require monotonicity where the attribute profiles are positively correlated with the probability of a correct response. Enforcing monotonicity avoids a type of label switching known as attribute-level switching where the meaning of $\alpha_k = 0$ changes from non-master to master or vice versa. Accordingly, our theorems establish identifiability up to label switching.

Theorem 3. *(*Strict Identifiability for Restricted HMMs*) For an HMM with emission probabilities in matrix $\boldsymbol{B}$ formed by the general model as shown in Eq.* (2), the parameter space $\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\omega})$ is identifiable if $rank(\boldsymbol{\omega}) = 2^K$, $\pi_c > 0$ for all $c = 0, \ldots, 2^K - 1$, and

(a) for $T \geq 3$, condition (B1) is satisfied;
(b) for $T = 2$, conditions (B1)-(B2) are satisfied.

Proof is shown in Appendix F.

Theorem 4. *(*Generic Identifiability for Restricted HMMs*) For an HMM with emission probabilities in matrix $\boldsymbol{B}$ formed by the general model as shown in Eq.* (2), the parameter space $\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\omega})$ is generically identifiable if $\pi_c > 0$ for all $c = 0, \ldots, 2^K - 1$, and

(a) for $T \geq 3$, condition (B1) is satisfied;
(b) for $T = 2$, conditions (B1)-(B2) are satisfied.

Proof is shown in Appendix F.

2.3.2. DINA Model We denote the whole parameter space of the restricted HMM by

$$\begin{aligned} \boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}) = \{(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}) : \boldsymbol{\pi} \in \boldsymbol{\Omega}(\boldsymbol{\pi}), \boldsymbol{\omega} \in \boldsymbol{\Omega}(\boldsymbol{\omega}), \boldsymbol{s} \in (0,1)^J, \\ \boldsymbol{g} \in (0,1)^J, \boldsymbol{Q} \in \{0,1\}^{J \times K}\}. \end{aligned} \tag{9}$$

We let $\boldsymbol{\Delta}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q})$ denote the set of non-identifiable parameters from $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q})$:

$$\begin{aligned} \boldsymbol{\Delta}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}) = \{(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}) : P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}) = P(\boldsymbol{Y} = \boldsymbol{y} \mid \bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{Q}}) \\ \text{for some } (\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{Q}}) \not\sim (\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}), \\ (\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}) \in \boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}),\ (\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{Q}}) \in \boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q})\}. \end{aligned} \tag{10}$$

Similar to Definition 2, if the non-identifiable parameter set $\boldsymbol{\Delta}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q})$ is of measure zero with respect to parameter space $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q})$, then we say $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q})$ is a generically identifiable parameter space.

Remark 4. In Eq. (9), we exclude extreme values (0 and 1) for both $\boldsymbol{s}$ and $\boldsymbol{g}$ to make sure that we always have positive response probabilities, which is also required in Theorem 5.

The following two conditions are needed in Theorems 5 and 6 for identifiability of restricted HMMs.

($A1$) $\boldsymbol{Q}$ matrix takes the following form (after a row permutation):

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{I}_K \\ \boldsymbol{Q}^* \end{pmatrix}_{J \times K},$$

where $\boldsymbol{I}_K$ represents a $K \times K$ identity matrix and $\boldsymbol{Q}^*$ can be any form except zero rows.
($A2$) $\boldsymbol{Q}$ matrix takes the following form (after a row permutation):

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{I}_K \\ \boldsymbol{I}_K \\ \boldsymbol{Q}^{**} \end{pmatrix}_{J \times K},$$

where $\boldsymbol{Q}^{**}$ could be any form except zero rows.

Theorem 5. *(*Strict Identifiability for Restricted HMMs-DINA*) For an HMM with emission probabilities in matrix* $\boldsymbol{B}$ *formed by the DINA model as shown in Eq.* (5), the parameter space $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q})$ is identifiable if $rank(\boldsymbol{\omega}) = 2^K$, $g_j \neq 1 - s_j$ for all $j = 1, \ldots, J$, $\pi_c > 0$ for all $c = 0, \ldots, 2^K - 1$, and

(a) for $T \geq 3$, condition (A1) is satisfied;
(b) for $T = 2$, condition (A2) is satisfied.

Proof is found in Appendix B (for part (a)) and D (for part (b)).

Remark 5. Compared with the sufficient and necessary condition for the strict identifiability under the static DINA model (i.e., the case with $T = 1$) shown in Gu and Xu (2021), a common condition is that the $\boldsymbol{Q}$ matrix should be complete, i.e., the $\boldsymbol{Q}$ matrix contains an identity submatrix $\boldsymbol{I}_K$. The conditions required for the remaining rows of $\boldsymbol{Q}$ are different between the static DINA and the restricted HMM formed by the DINA. For example, under the static DINA $\boldsymbol{Q}^*$ must have unique columns, whereas the $\boldsymbol{Q}^*$ matrix for the restricted DINA HMM must have nonzero rows.

Theorem 6. *(*Generic Identifiability for Restricted HMMs-DINA*) For an HMM with emission probabilities in matrix* $\boldsymbol{B}$ *formed by the DINA model as shown in Eq.* (5), the parameter space $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q})$ is generically identifiable if $\pi_c > 0$ for all $c = 0, \ldots, 2^K - 1$, and

(a) for $T \geq 3$, condition (A1) is satisfied;
(b) for $T = 2$, condition (A2) is satisfied.

Proof is found in Appendix B (for part (a)) and D (for part (b)).

Remark 6. Theorems 5 and 6 hold up to label switching of attributes. Conditions (A1) and (A2) exclude the possibility of zero rows in $\boldsymbol{Q}^*$ and $\boldsymbol{Q}^{**}$ so that guessing parameters are identifiable.

3. Bayesian Formulation for the DINA Model

Following the same setting in Sect. 1.2, suppose there are N subjects, J items, K skills and T time points. We use subscript $i = 1, \ldots, N$ to index subjects, $j = 1, \ldots, J$ to index items, $t = 1, \ldots, T$ to index time points, and $c = 0, \ldots, 2^K - 1$ to index latent states. Let $\boldsymbol{\alpha}_{it}$ denote the attribute profile of subject i at time point t, where $\boldsymbol{\alpha}_{it} = (\alpha_{i1t}, \ldots, \alpha_{iKt})^\top$ and $\boldsymbol{\alpha}_i = (\boldsymbol{\alpha}_{i1}, \ldots, \boldsymbol{\alpha}_{iT})^\top$, and Y_{ijt} denote the response of subject i to item j at time point t. The likelihood of observing a sample of N responses to J items with T time points is given by

$$p(\boldsymbol{Y}|\boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\pi}_1, \boldsymbol{\omega}) = \prod_{i=1}^{N} \sum_{\substack{\boldsymbol{\alpha}_{i1} \in \mathcal{A},\ \boldsymbol{\alpha}_{it} \in \mathcal{A}_+^t \\ t=2,\ldots,T}} p(\boldsymbol{Y}_i|\boldsymbol{\alpha}_i, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\pi}_1, \boldsymbol{\omega}) p(\boldsymbol{\alpha}_{i1}|\boldsymbol{\pi}_1) \times \prod_{t=2}^{T} p(\boldsymbol{\alpha}_{it}|\boldsymbol{\alpha}_{i,t-1}, \boldsymbol{\omega}),$$

where $\boldsymbol{\pi}_1$ is the initial distribution of attribute profiles, $\mathcal{A}$ represents the set of all attribute vectors, and $\mathcal{A}_+^t$ represents the set of nondecreasing learning trajectories at time t (Chen, Culpepper, Wang, & Douglas, 2018). The posterior distribution of the parameters for the restricted HMM is

$$p(\boldsymbol{\alpha}, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\omega}, \boldsymbol{\pi}_1|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\pi}_1, \boldsymbol{\omega}) p(\boldsymbol{\alpha}|\boldsymbol{\pi}_1, \boldsymbol{\omega}) p(\boldsymbol{Q}) p(\boldsymbol{\pi}_1) p(\boldsymbol{\omega}) p(\boldsymbol{s}, \boldsymbol{g}).$$

We formulate the DINA Bayesian model as follows:

$$Y_{ijt}|\boldsymbol{\alpha}_{it}, \boldsymbol{q}_j, s_j, g_j \sim Bernoulli\left((1-s_j)^{\eta_{ijt}} g_j^{(1-\eta_{ijt})}\right), \quad \eta_{ijt} = \mathcal{I}\left(\boldsymbol{\alpha}_{it}^\top \boldsymbol{q}_j \geq \boldsymbol{q}_j^\top \boldsymbol{q}_j\right), \tag{11}$$

$$\begin{aligned} p(\boldsymbol{\alpha}_i|\boldsymbol{\pi}_1, \boldsymbol{\omega}) &= p(\boldsymbol{\alpha}_{i1}|\boldsymbol{\pi}_1) \prod_{t=2}^{T} p\left(\boldsymbol{\alpha}_{it}|\boldsymbol{\alpha}_{i,t-1}, \boldsymbol{\omega}\right) \\ &= \left(\prod_{\boldsymbol{\alpha}_c \in \mathcal{A}} \pi_{1c}^{\mathcal{I}(\boldsymbol{\alpha}_{i1}=\boldsymbol{\alpha}_c)}\right) \prod_{t=2}^{T} \prod_{\boldsymbol{\alpha}_c \in \mathcal{A}_+^{t-1}} \prod_{\boldsymbol{\alpha}_{c'} \in \mathcal{A}_+^t} \omega_{c'|c}^{\mathcal{I}(\boldsymbol{\alpha}_{it}=\boldsymbol{\alpha}_{c'})\mathcal{I}(\boldsymbol{\alpha}_{i,t-1}=\boldsymbol{\alpha}_c)}, \end{aligned} \tag{12}$$

$$\boldsymbol{\pi}_1 = \left(\pi_{1,0}, \ldots, \pi_{1,2^K-1}\right) \sim \text{Dirichlet}\left(\boldsymbol{\delta}_0\right), \ \boldsymbol{\delta}_0 = \left(\delta_{0,0}, \ldots, \delta_{0,2^K-1}\right), \tag{13}$$

$$\boldsymbol{\omega}_c = \left(\omega_{0|c}, \ldots, \omega_{2^K-1|c}\right) \sim \text{Dirichlet}\left(\boldsymbol{\delta}_c\right), \ \boldsymbol{\delta}_c = \left(\delta_{0|c}, \ldots, \delta_{2^K-1|c}\right), \tag{14}$$

$$p\left(s_j, g_j\right) \propto s_j^{\alpha_s - 1} \left(1 - s_j\right)^{\beta_s - 1} g_j^{\alpha_g - 1} \left(1 - g_j\right)^{\beta_g - 1} \mathcal{I}\left(0 < g_j < 1 - s_j < 1\right), \tag{15}$$

$$p(\boldsymbol{Q}) \propto \mathcal{I}(\boldsymbol{Q} \in \mathcal{Q}). \tag{16}$$

We add restriction '$g_j < 1 - s_j$' in Eq. (15) to avoid the label switching issue for attributes. Equation (16) shows a uniform prior for the $\boldsymbol{Q}$ matrix in the space $\mathcal{Q}$ of identifiable models under the generic identifiability condition given in Conditions (A1) or (A2).

The Gibbs sampling algorithm is implemented to sample from the posterior distribution. Full conditional distributions of parameters shown above are included in Appendix E. Furthermore, we apply constrained Gibbs sampling method for the $\boldsymbol{Q}$ matrix, which was discussed in Chen, Culpepper, Chen, and Douglas (2018). The full sampling steps of all parameters are shown in Algorithm 1.

Algorithm 1

Initialize with an identifiable $\boldsymbol{Q}^{(0)}$ matrix, attribute profiles $\boldsymbol{\alpha}^{(0)}$, attribute categorical probabilities $\boldsymbol{\pi}_1^{(0)}$, transition matrix $\boldsymbol{\omega}^{(0)}$, and other item parameters $\boldsymbol{s}^{(0)}$ and $\boldsymbol{g}^{(0)}$.
for r in $1:R$ **do**
  for all j in $1:J$ and k in $1:K$ **do**
    if q_{jk} is the element in a row vector $\boldsymbol{e}_k$ (a 0-1 vector with a single 1 in the k-th position) where there is only one $\boldsymbol{e}_k$ (if $T \geq 3$) or two $\boldsymbol{e}_k$'s (if $T = 2$) in the current $\boldsymbol{Q}$ **then**
      Let $q_{jk}^{(r)} = q_{jk}^{(r-1)}$.
    else
      Update $q_{jk}^{(r)} = i$ $(i = 0, 1)$ with weight proportional to $\prod_{i=1}^{N}\prod_{t=1}^{T} p(\boldsymbol{Y}_{ijt}|s_j^{(r-1)}, g_j^{(r-1)}, \boldsymbol{\alpha}_{it}^{(r-1)}, \boldsymbol{Q}_{new}^{(r)}, q_{jk}^{(r)} = i, \boldsymbol{Q}_{old}^{(r-1)})$, where $\boldsymbol{Q}_{new}^{(r)}$ represents the entries of current $\boldsymbol{Q}$ that have already been updated, and $\boldsymbol{Q}_{old}^{(r-1)}$ represents the entries of current $\boldsymbol{Q}$ that have not been updated.
    end if
  end for
  for i in $1:N$ **do**
    for t in $1:T$ **do**
      if $t = 1$ **then**
        Given $\boldsymbol{\alpha}_{i2}^{(r-1)} = \boldsymbol{\alpha}_{c2}$, update $\boldsymbol{\alpha}_{i1}^{(r)}$ to $\boldsymbol{\alpha}_{l1}$ with weight proportional to $p(\boldsymbol{Y}_{i1}|\boldsymbol{\alpha}_{l1}, \boldsymbol{Q}^{(r)}, \boldsymbol{s}^{(r-1)}, \boldsymbol{g}^{(r-1)}) \cdot \pi_{1,l} \cdot \omega_{c|l}$.
      else if $1 < t < T$ **then**
        Given $\boldsymbol{\alpha}_{i,t-1}^{(r)} = \boldsymbol{\alpha}_{c,t-1}$ and $\boldsymbol{\alpha}_{i,t+1}^{(r-1)} = \boldsymbol{\alpha}_{c',t+1}$, update $\boldsymbol{\alpha}_{it}^{(r)}$ to $\boldsymbol{\alpha}_{lt}$ with weight proportional to $p(\boldsymbol{Y}_{it}|\boldsymbol{\alpha}_{lt}, \boldsymbol{Q}^{(r)}, \boldsymbol{s}^{(r-1)}, \boldsymbol{g}^{(r-1)}) \cdot \omega_{l|c} \cdot \omega_{c'|l}$.
      else
        Given $\boldsymbol{\alpha}_{i,T-1}^{(r-1)} = \boldsymbol{\alpha}_{c,T-1}$, update $\boldsymbol{\alpha}_{iT}^{(r)}$ to $\boldsymbol{\alpha}_{lT}$ with weight proportional to $p(\boldsymbol{Y}_{iT}|\boldsymbol{\alpha}_{lT}, \boldsymbol{Q}^{(r)}, \boldsymbol{s}^{(r-1)}, \boldsymbol{g}^{(r-1)})\omega_{l|c}$.
      end if
    end for
  end for
  Update $\boldsymbol{\pi}_1^{(r)}|\boldsymbol{\alpha}^{(r)} \sim Dirichlet(\tilde{\boldsymbol{N}}_0 + \boldsymbol{\delta}_0)$, where $\tilde{\boldsymbol{N}}_0 = (\tilde{N}_{0,1}, \ldots, \tilde{N}_{0,2^K})^\top$ represents the frequencies of each initial attribute pattern $\boldsymbol{\alpha}_{c1}$, $c = 0, \ldots, 2^K - 1$.
  For $c = 0, \ldots, 2^K - 1$, update $\boldsymbol{\omega}_c^{(r)}|\boldsymbol{\alpha}^{(r)} \sim Dirichlet(\tilde{\boldsymbol{N}}_c + \boldsymbol{\delta}_c)$, where $\tilde{\boldsymbol{N}}_c = (\tilde{N}_{0|c}, \ldots, \tilde{N}_{2^K-1|c})^\top$ represents the number of subjects that changed their attribute profiles to $\boldsymbol{\alpha}_{c,t+1}$ at time $t+1$, where $t = 1, \ldots, T-1$.
  For $j = 1, \ldots, J$, update $s_j^{(r)}, g_j^{(r)}|\boldsymbol{Y}, \boldsymbol{\alpha}^{(r)}, \boldsymbol{Q}^{(r)} \sim Beta(a_s, b_s)Beta(a_g, b_g)\mathcal{I}(0 < g_j < 1 - s_j < 1)$, i.e., sample $s_j^{(r)}$ and $g_j^{(r)}$ independently from $Beta(a_s, b_s)$ and $Beta(a_g, b_g)$ truncated in the region $0 < g_j < 1 - s_j < 1$. Expressions for a_s, b_s, a_g and b_g are in Equation (12) of Culpepper (2015) (a_s, b_s, a_g and b_g here are corresponding to $\tilde{\alpha}_s$, $\tilde{\beta}_s$, $\tilde{\alpha}_g$ and $\tilde{\beta}_g$ in Equation (12)).
end for

4. Monte Carlo Simulation Study

4.1. Settings

We next report results from a Monte Carlo experiment to evaluate the performance of Algorithm 1. We conducted the simulation study under different sample size (i.e., $N = 500$, 1000, and 2000), numbers of attributes (i.e., $K = 3$, 4, and 5), the length of the time period (i.e., $T = 2$, 3,

4, and 5), and correlations among the attributes at the first time point (i.e., $\rho = 0$, $\rho = 0.25$ and 0.5).

For the $\rho = 0$ case, the attribute profile $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{K1})^\top$ is generated uniformly from all possible 2^K cases. For the $\rho > 0$ case, the dependence among attribute profiles is introduced using the method of Chiu, Douglas, and Li (2009). Suppose $\boldsymbol{Z} = (Z_1, \ldots, Z_K)^\top$ follows a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with unit variance and correlation ρ, where $\boldsymbol{\Sigma} = (1-\rho)\boldsymbol{I}_K + \rho \mathbf{1}_K \mathbf{1}_K^\top$ and $\mathbf{1}_K$ is a column vector of 1 with length K. Then, the attribute profile $\boldsymbol{\alpha}_1$ is given by $\alpha_{k1} = \mathcal{I}(Z_k \geq \Phi^{-1}(\frac{k}{K+1}))$, $k = 1, \ldots, K$, where Φ is the cumulative distribution function of the standard normal distribution.

We let the slipping and guessing parameters be 0.2 and 0.3, respectively. Our true $\boldsymbol{Q}$ has $J = 18$ items for $K = 3, 4$ and $J = 20$ items for $K = 5$. Also, as discussed in Chen, Culpepper, Wang, and Douglas (2018), we assume that attributes are non-decreasing over time, so the transition matrix $\boldsymbol{\omega}$ is an upper triangular matrix. We sample the true $\boldsymbol{\omega}$ from the prior under the assumption of non-decreasing learning trajectories. The simulation study specified the true unknown $\boldsymbol{Q}$ matrices, $\boldsymbol{Q}_{K=3}$, $\boldsymbol{Q}_{K=4}$ and $\boldsymbol{Q}_{K=5}$, in Eq. (17), which all satisfy the identifiability constraints given in Conditions $(A1)$ and $(A2)$.

We use a Markov chain of length of 20,000 with a 10,000 burn-in period for $K = 3$, a chain of length of 30,000 with a 20,000 burn-in period for $K = 4$, and a chain of length of 40,000 with a 30,000 burn-in period for $K = 5$.

$$\boldsymbol{Q}_{K=3} = \begin{pmatrix} 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&1&0\\ 1&0&1\\ 0&1&1\\ 1&1&0\\ 1&0&1\\ 0&1&1\\ 1&1&1\\ 1&1&1\\ 1&1&1 \end{pmatrix}, \quad \boldsymbol{Q}_{K=4} = \begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\\ 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\\ 1&1&0&0\\ 1&0&1&0\\ 1&0&0&1\\ 0&1&1&0\\ 0&1&0&1\\ 0&0&1&1\\ 1&1&1&0\\ 1&1&0&1\\ 1&0&1&1\\ 0&1&1&1 \end{pmatrix}, \quad \boldsymbol{Q}_{K=5} = \begin{pmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 0&0&0&1&1\\ 0&1&0&0&1\\ 1&0&0&0&1\\ 1&0&1&0&0\\ 1&1&0&0&0\\ 0&0&1&1&1\\ 0&1&0&1&1\\ 0&1&1&0&1\\ 1&0&0&1&1\\ 1&1&1&0&0 \end{pmatrix}, \tag{17}$$

4.2. *Results*

We repeated the simulation study 100 times for each setting. We use several metrics to evaluate parameter recovery. In particular, we report the average element-wise accuracy rate (EAR) for $\boldsymbol{Q}$ by comparing the estimated $\hat{\boldsymbol{Q}}$ and the true $\boldsymbol{Q}$ matrix, where $\hat{\boldsymbol{Q}}$ is the mode of all samples after the burn-in period. Furthermore, we compute the average root mean squared error (RMSE) to assess the accuracy of the estimated transition matrix $\hat{\boldsymbol{\omega}}$, where $\hat{\boldsymbol{\omega}}$ is the mean of all samples

TABLE 1.
Summary of simulation performance for restricted HMM.

K	T	ρ	$N = 500$			$N = 1000$			$N = 2000$		
			$\hat{\boldsymbol{Q}} = \boldsymbol{Q}$	EAR	RMSE	$\hat{\boldsymbol{Q}} = \boldsymbol{Q}$	EAR	RMSE	$\hat{\boldsymbol{Q}} = \boldsymbol{Q}$	EAR	RMSE
3	2	0.00	99	0.9981	0.0714	100	1.0000	0.0515	100	1.0000	0.0379
3	3	0.00	96	0.9907	0.0622	93	0.9856	0.0562	98	0.9963	0.0320
3	4	0.00	91	0.9826	0.0643	94	0.9883	0.0502	90	0.9815	0.0507
3	5	0.00	95	0.9911	0.0591	87	0.9722	0.0675	85	0.9678	0.0629
4	2	0.00	98	0.9997	0.0912	99	0.9989	0.0743	99	0.9983	0.0585
4	3	0.00	96	0.9944	0.0785	99	0.9982	0.0599	93	0.9885	0.0525
4	4	0.00	94	0.9901	0.0775	97	0.9953	0.0601	94	0.9900	0.0527
4	5	0.00	93	0.9876	0.0785	94	0.9882	0.0626	90	0.9846	0.0579
5	2	0.00	98	0.9976	0.0956	99	0.9993	0.0861	93	0.9922	0.0750
5	3	0.00	95	0.9947	0.0811	93	0.9916	0.0757	97	0.9962	0.0605
5	4	0.00	90	0.9868	0.0796	92	0.9909	0.0734	91	0.9879	0.0631
5	5	0.00	91	0.9902	0.0827	91	0.9892	0.0727	87	0.9819	0.0661
3	2	0.25	94	0.9909	0.1135	97	0.9939	0.0856	94	0.9869	0.0710
3	3	0.25	97	0.9961	0.0834	94	0.9874	0.0699	93	0.9856	0.0577
3	4	0.25	98	0.9972	0.0750	94	0.9844	0.0663	94	0.9893	0.0522
3	5	0.25	100	1.0000	0.0732	97	0.9928	0.0595	95	0.9896	0.0518
4	2	0.25	74	0.9865	0.1301	90	0.9858	0.1142	85	0.9717	0.1036
4	3	0.25	95	0.9908	0.0969	88	0.9796	0.0865	86	0.9757	0.0790
4	4	0.25	88	0.9821	0.0952	88	0.9788	0.0841	82	0.9699	0.0774
4	5	0.25	81	0.9700	0.0991	84	0.9728	0.0889	83	0.9703	0.0787
5	2	0.25	49	0.9783	0.1041	59	0.9540	0.1026	52	0.9409	0.0983
5	3	0.25	71	0.9682	0.0991	70	0.9579	0.0958	56	0.9353	0.0949
5	4	0.25	75	0.9721	0.0978	75	0.9670	0.0917	63	0.9508	0.0894
5	5	0.25	76	0.9707	0.0970	77	0.9698	0.0913	64	0.9498	0.0926
3	2	0.5	84	0.9835	0.1435	95	0.9894	0.1153	88	0.9759	0.0988
3	3	0.5	92	0.9898	0.0912	97	0.9933	0.0781	90	0.9835	0.0651
3	4	0.5	96	0.9907	0.0874	96	0.9911	0.0721	93	0.9876	0.0593
3	5	0.5	91	0.9869	0.0929	94	0.9865	0.0743	91	0.9824	0.0654
4	2	0.5	56	0.9811	0.1399	73	0.9596	0.1364	69	0.9461	0.1281
4	3	0.5	83	0.9788	0.1097	78	0.9610	0.1019	68	0.9421	0.1000
4	4	0.5	82	0.9729	0.1042	77	0.9643	0.0983	78	0.9596	0.0844
4	5	0.5	81	0.9706	0.1040	69	0.9490	0.1039	77	0.9599	0.0879
5	2	0.5	16	0.9602	0.1073	25	0.9393	0.1068	35	0.9184	0.1057
5	3	0.5	57	0.9690	0.1000	67	0.9565	0.0984	51	0.9273	0.1008
5	4	0.5	74	0.9760	0.0978	67	0.9565	0.0975	54	0.9254	0.0990
5	5	0.5	75	0.9740	0.0975	70	0.9590	0.0951	60	0.9452	0.0958

$\hat{\boldsymbol{Q}} = \boldsymbol{Q}$ indicates the number of times the estimated $\hat{\boldsymbol{Q}}$ equals to the true $\boldsymbol{Q}$ out of 100 repetitions in each case; EAR = element-wise accuracy rate, averaged over 100 repetitions; RMSE = root mean squared error for $\boldsymbol{\omega}$, averaged over 100 repetitions.

after the burn-in period. Simulation results in Table 1 show a good recovery for $\boldsymbol{Q}$ matrix. It also suggests that for fixed K, the RMSE becomes smaller as the time period gets longer, since we have more attribute samples for estimating the transition matrix $\boldsymbol{\omega}$; the EARs are high for most of the settings, especially when $\rho = 0$; as ρ becomes larger, there is a slight impact on the EAR. Table 2 shows the average computation time for our simulation study using a MacBook Pro with 2.3 GHz Intel Core i5 processor.

TABLE 2.
Computation time (minutes) per replication.

K	T	$N = 500$	$N = 1000$	$N = 2000$
3	2	4.6	8.6	18.0
3	3	6.0	12.5	25.7
3	4	8.7	17.7	34.5
3	5	10.5	21.8	40.0
4	2	11.4	21.4	38.0
4	3	16.1	28.3	56.9
4	4	20.0	39.5	77.3
4	5	25.0	48.6	94.8
5	2	26.5	47.6	94.1
5	3	39.1	70.5	132.2
5	4	48.8	90.0	174.8
5	5	60.0	111.9	220.3

5. Real Data Analysis

In this section, we apply Algorithm 1 to the Problems in Elementary Probability Theory data set (Heller & Wickelmaier, 2013).

5.1. Problems in Elementary Probability Theory

This data set contains responses to two sets of $J = 12$ questions in elementary probability theory observed before and after some instructions. All 504 participants completed the first set of questions, but only 345 of them completed the second set of questions, so we have $N = 345$ and $T = 2$. We compared models with $K = 2$, 3, and 4 and set $K = 3$ based on results for the log-likelihood from a 10-fold cross-validation method. We ran five Markov chains with $K = 3$ for convergence diagnostics of the Markov chain. Figure 2 shows the plot of maximum proportional scale reduction factor (PSRF) (Brooks & Gelman, 1998) for checking the convergence of Markov chain with multivariate parameters. The approximate convergence is achieved after 10, 000 iterations since the maximum PSRF remains below 1.1 after that. So we ran 100 Markov chains of length 20, 000 (with 10, 000 as burn-in) with $K = 3$ to estimate the parameters, and the results are shown in Table 3.

Table 3 reports the estimated $\hat{\boldsymbol{Q}}$ matrix, an expert-specified $\boldsymbol{Q}$ matrix, and item parameters. Note that we constructed the estimated $\hat{\boldsymbol{Q}}$ by first finding the posterior mode of $\hat{\boldsymbol{Q}}$ for each of the 100 repetitions and then selecting the value with the highest log-likelihood.

The estimated $\hat{\boldsymbol{Q}}$ matrix shares some common interpretation with the expert $\boldsymbol{Q}$ matrix. The expert $\boldsymbol{Q}$ matrix includes four attributes: (1) calculate the classic probability of an event (pb); (2) probability of the complement of an event (cp); (3) the union of two disjoint events (un); and (4) the probability of two independent events (id). By referring to the two problem sets (given in Appendix H) and the expert $\boldsymbol{Q}$ matrix, we conclude that Attribute 1 in our estimated $\hat{\boldsymbol{Q}}$ is related to calculations of classical probability (pb) (Question 1) and the understanding of independence in probability theory (id) (Questions 4, 10, 11, 12), Attribute 2 represents the mastery of applying probability models, including probabilities of the union of two disjoint events (un) (Questions 3, 7, 8, 12) and the complement of an event (cp) (Questions 2, 5, 6), and Attribute 3 is related to problems about a standard deck of cards (Questions 5, 9).

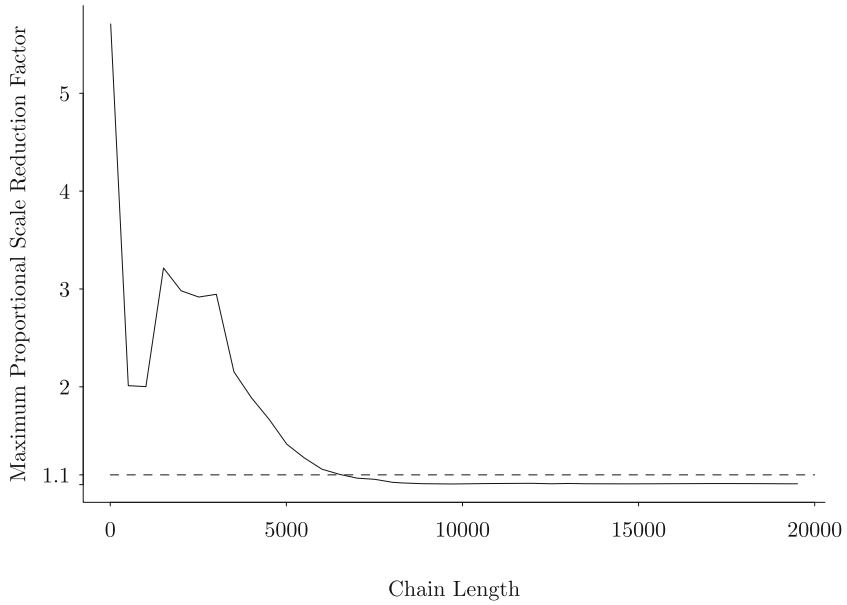

FIGURE 2.
The maximum PSRF for Problems in Elementary Probability Theory data.

TABLE 3.
Estimated $\hat{\boldsymbol{Q}}$, slipping $\hat{s}_j$ and guessing $\hat{g}_j$ parameters for Problems in Elementary Probability Theory data.

Question	Expert $\boldsymbol{Q}$				$\hat{\boldsymbol{Q}}$			$\hat{s}_j$	$\hat{g}_j$
	pb	cp	un	id	$A1$	$A2$	$A3$		
1	1	0	0	0	1	0	0	0.041	0.751
2	0	1	0	0	0	1	0	0.014	0.817
3	0	0	1	0	0	1	0	0.032	0.590
4	0	0	0	1	1	0	0	0.030	0.274
5	1	1	0	0	0	0	1	0.079	0.646
6	1	1	0	0	0	1	0	0.029	0.534
7	1	0	1	0	0	1	0	0.031	0.638
8	1	0	1	0	0	1	0	0.027	0.770
9	1	0	0	1	0	0	1	0.029	0.260
10	0	1	0	1	1	0	0	0.114	0.087
11	1	1	0	1	1	0	0	0.219	0.105
12	1	0	1	1	1	1	0	0.109	0.156

The expert $\boldsymbol{Q}$ can be found in Heller and Wickelmaier (2013).

Table 4 shows the estimated transition matrix $\hat{\boldsymbol{\omega}}$ and proportions of each attribute pattern. Learning trajectories can be inferred by the transition probabilities shown in the table. For example, the 5-th row of $\hat{\boldsymbol{\omega}}$ shows the probabilities that students who have mastered skill 1 would transfer to other states next time. Students mastered skill 1 are more likely to master more skills next time compared with those who mastered skill 2 or 3 only, and the most likely skill to be mastered next time is skill 2, and then skill 3 in the end. Similarly, we can deduce learning trajectories for students who mastered only skill 2 or 3 in the beginning.

TABLE 4.
Estimated distribution of initial attributes and transition matrix $\boldsymbol{\omega}$ for for Problems in Elementary Probability Theory data.

c	$\boldsymbol{\alpha}_c$	$\hat{\boldsymbol{\pi}}_1$	$\hat{\boldsymbol{\omega}}_{1\|c}$	$\hat{\boldsymbol{\omega}}_{2\|c}$	$\hat{\boldsymbol{\omega}}_{3\|c}$	$\hat{\boldsymbol{\omega}}_{4\|c}$	$\hat{\boldsymbol{\omega}}_{5\|c}$	$\hat{\boldsymbol{\omega}}_{6\|c}$	$\hat{\boldsymbol{\omega}}_{7\|c}$	$\hat{\boldsymbol{\omega}}_{8\|c}$
0	000	0.089	0.566	0.062	0.109	0.049	0.061	0.035	0.057	0.061
1	001	0.021		0.317	0.000	0.221	0.000	0.180	0.000	0.282
2	010	0.071			0.309	0.196	0.000	0.000	0.289	0.206
3	011	0.076				0.380	0.000	0.000	0.000	0.620
4	100	0.019					0.268	0.192	0.309	0.231
5	101	0.025						0.377	0.000	0.623
6	110	0.145							0.720	0.280
7	111	0.554								1.000

The attributes represented by $\boldsymbol{\alpha}_c$ are labeled in the order of $A1$, $A2$ and $A3$. $\hat{\boldsymbol{\omega}}_{c'|c}$ refers to the transition probability from state c to state c'.

6. Discussion

This paper focuses on the identifiability issue of discrete time HMMs with finite hidden states. We proposed generic identifiability conditions by generalizing the strict identifiability condition discussed in Bonhomme et al. (2016), which may be too strong in practical analysis. Then, we proved the strict and generic identifiability conditions for our restricted HMMs, in which the emission probability is formed by two kinds of restricted latent class models. Also, we developed a Bayesian formulation for the restricted HMM where the generic identifiability conditions are taken into consideration. The simulation results show that our algorithm can efficiently estimate model parameters under different model settings for restricted HMMs.

In educational studies, researchers usually impose a non-decreasing pattern on the restricted HMM transition matrix $\boldsymbol{\omega}$. However, this format implies that the last state is an absorbing state, and only the probability of the last state is nonzero in the stationary distribution, which does not satisfy our strict and generic identifiability conditions in Theorems 5 and 6. The generic identifiability proposed in our paper is a sufficient condition, so there might exist a weaker condition on the stationary distribution $\boldsymbol{\pi}$ which would allow the non-decreasing configuration. Therefore, one direction for future research would be the derivation of necessary and sufficient generic identifiability conditions for discrete time HMMs. Understanding necessary conditions is more difficult when using the Kruskal condition (Kruskal, 1977) for the uniqueness of three-way arrays as it offers a general sufficient condition for establishing uniqueness. Gu and Xu (2021) established necessary and sufficient conditions for the static DINA model. Future research may be able to extend their proof technique to understand necessary conditions for restricted HMMs.

In this paper, we assume that the number of attributes, K, is fixed and known. However, the prior knowledge for K may not be available in some cases. In real data analysis, we chose K by applying cross-validation. It is also possible to assume that K is an unknown parameter that needs to be inferred. However, one challenge is that an unknown K implies that the dimensions of attribute profiles and transition matrix are no longer available. Future research should consider methods for accurately inferring both $\boldsymbol{Q}$ and K (Chen, Liu, Culpepper, & Chen, 2021).

Funding The authors gratefully acknowledge the financial support of the NSF Grant Nos. SES-1758631, SES-1951057, and SES 21-50628.

Appendix A: Proof of Part (a) of Theorem 2 ($T \geq 3$ case)

We start with introducing some basic terminology and facts from algebraic geometry.

Definition 3. (Cox, Little, and O'Shea (2015)) An algebraic variety V is defined as the simultaneous zero-set of a finite collection of multivariate polynomials $\{f_i\}_{i=1}^n \subset \mathbb{C}[x_1, \ldots, x_k]$,

$$V = V(f_1, \ldots, f_n) = \{\boldsymbol{a} \in \mathbb{C}^k | f_i(\boldsymbol{a}) = 0, 1 \leq i \leq n\}. \tag{A1}$$

Here $\mathbb{C}[x_1, \ldots, x_k]$ represents the set of all polynomials in $x_1, \ldots, x_k$ with coefficients in $\mathbb{C}$, and $\mathbb{C}^k$ is the set of k-dimensional complex numbers.

Lemma 1. *(Allman et al. (2009)) A variety is all of $\mathbb{C}^k$ only when all f_i are 0; otherwise, a variety is called a proper subvariety and must be of dimension less than k, and of Lebesgue measure 0 in $\mathbb{C}^k$.*

Remark 7. In Lemma 1, analogous statements still hold if we replace $\mathbb{C}^k$ by $\mathbb{R}^k$.

In order to show generic identifiability of model parameters, we can prove that all nonidentifiable parameter choices lie within a proper subvariety, and thus form a set of Lebesgue measure zero based on Lemma 1.

Proposition 1. *$rank(\boldsymbol{B}) = rank(\boldsymbol{\omega}) = 2^K$ if and only if $rank(\boldsymbol{B} \cdot \boldsymbol{\omega}) = 2^K$.*

Proof. By Sylvester's rank inequality (Matsaglia & Styan, 1974), we have

$$rank(\boldsymbol{B}) + rank(\boldsymbol{\omega}) - 2^K \leq rank(\boldsymbol{B} \cdot \boldsymbol{\omega}) \leq min\{rank(\boldsymbol{B}), rank(\boldsymbol{\omega})\},$$

so the proposition holds. □

By Proposition 1 and part (a) of Theorem 1, we only need to show that $rank(\boldsymbol{B} \cdot \boldsymbol{\omega}) = 2^K$ holds almost everywhere in $\boldsymbol{\Omega}_{\boldsymbol{\omega},\boldsymbol{B}} = \{(\boldsymbol{\omega}, \boldsymbol{B}) : \boldsymbol{\omega} \in \boldsymbol{\Omega}(\boldsymbol{\omega}), \boldsymbol{B} \in \boldsymbol{\Omega}(\boldsymbol{B})$ and $rank(\boldsymbol{B}) = 2^K\}$.
Let M be a subset of $\{1, \ldots, 2^J\}$ with 2^K elements, and then, let $[\boldsymbol{B} \cdot \boldsymbol{\omega}]_M$ denote the minor of a submatrix in $\boldsymbol{B} \cdot \boldsymbol{\omega}$ that corresponds to the rows with indices in M. Let

$$f(\boldsymbol{B}, \boldsymbol{\omega}) = \sum_M ([\boldsymbol{B} \cdot \boldsymbol{\omega}]_M)^2 : \boldsymbol{\Omega}_{\boldsymbol{\omega},\boldsymbol{B}} \to \mathbb{R} \tag{A2}$$

denote the summation of all squared minors of order 2^K of matrix $\boldsymbol{B} \cdot \boldsymbol{\omega}$.

Since $f(\boldsymbol{B}, \boldsymbol{\omega})$ is a polynomial function of $\boldsymbol{B}$ and $\boldsymbol{\omega}$, and we know that the rank of $\boldsymbol{B} \cdot \boldsymbol{\omega}$ is the maximal order of a nonzero minor of $\boldsymbol{B} \cdot \boldsymbol{\omega}$, then by Proposition 1, we can write the zero set of $f(\boldsymbol{B}, \boldsymbol{\omega})$ as:

$$\begin{aligned} \boldsymbol{Z}_f &= \{(\boldsymbol{\omega}, \boldsymbol{B}) : (\boldsymbol{\omega}, \boldsymbol{B}) \in \boldsymbol{\Omega}_{\boldsymbol{\omega},\boldsymbol{B}} \text{ and } f(\boldsymbol{B}, \boldsymbol{\omega}) = 0\} \\ &= \{(\boldsymbol{\omega}, \boldsymbol{B}) : (\boldsymbol{\omega}, \boldsymbol{B}) \in \boldsymbol{\Omega}_{\boldsymbol{\omega},\boldsymbol{B}} \text{ and } [\boldsymbol{B} \cdot \boldsymbol{\omega}]_M = 0 \text{ for all possible } M\} \\ &= \{(\boldsymbol{\omega}, \boldsymbol{B}) : \ \boldsymbol{\omega} \in \boldsymbol{\Omega}(\boldsymbol{\omega}), \ \boldsymbol{B} \in \boldsymbol{\Omega}(\boldsymbol{B}), \ rank(\boldsymbol{B}) = 2^K \text{ and } rank(\boldsymbol{\omega}) < 2^K\}. \end{aligned} \tag{A3}$$

In the following, we will show that $f(\boldsymbol{B}, \boldsymbol{\omega})$ is not a constant zero function.

Proposition 2. *If* $rank(\boldsymbol{B}) = 2^K$*, then there exists some nonsingular* $\boldsymbol{\omega}$*, such that* $f(\boldsymbol{B}, \boldsymbol{\omega}) \neq 0$.

Proof. Given a full column rank $\boldsymbol{B}$, there must exist a nonzero minor of order 2^K in $\boldsymbol{B}$. Without loss of generality, we assume that the first 2^K rows of $\boldsymbol{B}$, denoted by $\boldsymbol{B}^*$, satisfy $\det(\boldsymbol{B}^*) \neq 0$; then, $\det(\boldsymbol{B}^*)$ is a nonzero minor of order 2^K. Let $\boldsymbol{B} = (\boldsymbol{B}^{*\top}, \boldsymbol{B}'^{\top})^\top$. In order to show that $f(\boldsymbol{B}, \boldsymbol{\omega}) \neq 0$ for some nonsingular $\boldsymbol{\omega}$, it is enough to show that $\boldsymbol{B} \cdot \boldsymbol{\omega}$ has full column rank for some specific choice of nonsingular $\boldsymbol{\omega}$, since that will establish that some minors of order 2^K of $\boldsymbol{B} \cdot \boldsymbol{\omega}$ are nonzero polynomials in the entries of $\boldsymbol{B}$ and $\boldsymbol{\omega}$.
For any nonsingular $\boldsymbol{\omega}$, we have

$$\boldsymbol{B} \cdot \boldsymbol{\omega} = \begin{bmatrix} \boldsymbol{B}^* \\ \boldsymbol{B}' \end{bmatrix} \cdot \boldsymbol{\omega} = \begin{bmatrix} \boldsymbol{B}^* \cdot \boldsymbol{\omega} \\ \boldsymbol{B}' \cdot \boldsymbol{\omega} \end{bmatrix}. \tag{A4}$$

Since $\det(\boldsymbol{B}^*)$ is a nonzero minor of $\boldsymbol{B}$ and $\boldsymbol{\omega}$ is a nonsingular matrix, then $rank(\boldsymbol{B}^* \cdot \boldsymbol{\omega}) = 2^K$. Therefore, $\det(\boldsymbol{B}^* \cdot \boldsymbol{\omega})$ is a nonzero minor of $\boldsymbol{B} \cdot \boldsymbol{\omega}$, which implies $rank(\boldsymbol{B} \cdot \boldsymbol{\omega}) = 2^K$ and $f(\boldsymbol{B}, \boldsymbol{\omega}) \neq 0$. □

Therefore, by Lemma 1, the zero set $\boldsymbol{Z}_f$ has measure zero within $\boldsymbol{\Omega}_{\boldsymbol{\omega},\boldsymbol{B}}$. The HMM with $T \geq 3$ is generically identified.

Appendix B: Proof of Part (a) of Theorems 5 and 6 ($T \geq 3$ case)

We first show that if emission matrix $\boldsymbol{B}$ is identified, then parameters $\boldsymbol{s}$, $\boldsymbol{g}$, and $\boldsymbol{Q}$ in a restricted HMM can also be identified.

Proposition 3. *For any* $\boldsymbol{B}, \boldsymbol{B}' \in \boldsymbol{\Omega}(\boldsymbol{B})$, $\boldsymbol{s}, \boldsymbol{s}' \in (0,1)^J$, $\boldsymbol{g}, \boldsymbol{g}' \in (0,1)^J$ *and* $\boldsymbol{Q}, \boldsymbol{Q}' \in \{0,1\}^{J \times K}$, *we have*

$$\boldsymbol{B} = \boldsymbol{B}' \text{ if and only if } (\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}) = (\boldsymbol{s}', \boldsymbol{g}', \boldsymbol{Q}').$$

Proof. It suffices to show that given $\boldsymbol{B} = \boldsymbol{B}'$, we must have $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}) = (\boldsymbol{s}', \boldsymbol{g}', \boldsymbol{Q}')$.
For $j \in \{1, 2, \ldots, J\}$, let $\boldsymbol{D}_j$ be the matrix such that $\boldsymbol{D}_j \boldsymbol{B}$ and $\boldsymbol{D}_j \boldsymbol{B}'$ reduce to the 2×2^K matrix of conditional probabilities for Y_j given $\boldsymbol{\alpha}_t$. For instance, the second row of $\boldsymbol{D}_j \boldsymbol{B}$ is $P(Y_j = 1 \mid \boldsymbol{\alpha}_t)$:

$$(g_j^{1-\eta_{j0}}(1-s_j)^{\eta_{j0}}, \ldots, g_j^{1-\eta_{j,2^K-1}}(1-s_j)^{\eta_{j,2^K-1}}),$$

where $\eta_{jc} = \mathcal{I}\left(\boldsymbol{\alpha}_t^\top \boldsymbol{q}_j \geq \boldsymbol{q}_j^\top \boldsymbol{q}_j,\ \boldsymbol{\alpha}_t^\top \boldsymbol{v} = c\right), c = 0, \ldots, 2^K - 1$. Note that $\eta_{j,2^K-1} = \eta'_{j,2^K-1} = 1$ and the assumption that $\boldsymbol{q}_j \neq \mathbf{0}$ and $\boldsymbol{q}'_j \neq \mathbf{0}$ implies $\eta_{j0} = \eta'_{j0} = 0$. Therefore, $\boldsymbol{D}_j \boldsymbol{B} = \boldsymbol{D}_j \boldsymbol{B}'$ implies that $g_j = g'_j, s_j = s'_j$. Also, for $c \in \{1, \ldots, 2^K - 2\}$ we have

$$g_j^{1-\eta_{jc}}(1-s_j)^{\eta_{jc}} = g_j^{1-\eta'_{jc}}(1-s_j)^{\eta'_{jc}},$$

and $g_j \neq 1 - s_j$ implies that $\eta_{jc} \neq \eta'_{jc}$ is not possible, so $\boldsymbol{q}_j = \boldsymbol{q}'_j$. □

The emission matrix $\boldsymbol{B}$ is of size $2^J \times 2^K$, and we use $\boldsymbol{B}_{\boldsymbol{y}_t, \boldsymbol{\alpha}_t}$ to denote the element corresponding to the row with response pattern $\boldsymbol{y}_t$ (we refer to it as the $\boldsymbol{y}_t$-th row) and column with attribute profile $\boldsymbol{\alpha}_t$ (we refer to it as the $\boldsymbol{\alpha}_t$-th column), so $\boldsymbol{B}_{\boldsymbol{y}_t, \boldsymbol{\alpha}_t}$ is the emission probability

$$P(\boldsymbol{Y}_t = \boldsymbol{y}_t | \boldsymbol{\alpha}_t, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}) = \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}_t}^{y_{jt}} \left[1 - \theta_{j,\boldsymbol{\alpha}_t}\right]^{1-y_{jt}}, \tag{B1}$$

where $\theta_{j,\boldsymbol{\alpha}_t} = (1-s_j)^{\eta_{jt}} g_j^{(1-\eta_{jt})}$ and $\eta_{jt} = \mathcal{I}\left(\boldsymbol{\alpha}_t^\top \boldsymbol{q}_j \geq \boldsymbol{q}_j^\top \boldsymbol{q}_j\right)$.
As mentioned in Sect. 2.3, we have a bipartition of the set $\mathbb{J} = \{1, 2, \ldots, J\}$ into two disjoint, nonempty subsets $\mathbb{J}_1 = \{1, 2, \ldots, K\}$, $\mathbb{J}_2 = \{K+1, \ldots, J\}$. Then, let $\boldsymbol{Y}_t = (\boldsymbol{Y}_t^{\mathbb{J}_1 \top}, \boldsymbol{Y}_t^{\mathbb{J}_2 \top})^\top$, where $\boldsymbol{Y}_t^{\mathbb{J}_1} = (Y_{1t}, \ldots, Y_{Kt})^\top$ and $\boldsymbol{Y}_t^{\mathbb{J}_2} = (Y_{(K+1)t}, \ldots, Y_{Jt})^\top$. Assuming that the $\boldsymbol{Q}$ matrix has the form shown in condition $(A1)$, let

$$\boldsymbol{Q}_{J\times K} = \begin{pmatrix} \boldsymbol{Q}^{\mathbb{J}_1} \\ \boldsymbol{Q}^{\mathbb{J}_2} \end{pmatrix}, \tag{B2}$$

and without loss of generality, let $\boldsymbol{Q}^{\mathbb{J}_1} = \boldsymbol{I}_K$ and $\boldsymbol{Q}^{\mathbb{J}_2} = \boldsymbol{Q}^*$. Then, the emission probability can be decomposed into two parts since the components of $\boldsymbol{Y}_t$ are independent given profile $\boldsymbol{\alpha}_t$:

$$P(\boldsymbol{Y}_t = \boldsymbol{y}_t | \boldsymbol{\alpha}_t, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}) = P(\boldsymbol{Y}_t^{\mathbb{J}_1} = \boldsymbol{y}_t^{\mathbb{J}_1} | \boldsymbol{\alpha}_t, \boldsymbol{Q}^{\mathbb{J}_1} = \boldsymbol{I}_K, \boldsymbol{s}, \boldsymbol{g}) \cdot P(\boldsymbol{Y}_t^{\mathbb{J}_2} = \boldsymbol{y}_t^{\mathbb{J}_2} | \boldsymbol{\alpha}_t, \boldsymbol{Q}^*, \boldsymbol{s}, \boldsymbol{g}). \tag{B3}$$

Similarly, the emission matrix $\boldsymbol{B}$ can also be decomposed into two parts. Let $\boldsymbol{B}^{\mathbb{J}_1}$ be a matrix of size $2^K \times 2^K$, where its $\boldsymbol{y}_t^{\mathbb{J}_1}$-th row and $\boldsymbol{\alpha}_t$-th column element is $P(\boldsymbol{Y}_t^{\mathbb{J}_1} = \boldsymbol{y}_t^{\mathbb{J}_1} | \boldsymbol{\alpha}_t, \boldsymbol{I}_K, \boldsymbol{s}, \boldsymbol{g})$; and let $\boldsymbol{B}^{\mathbb{J}_2}$ be a matrix of size $2^{(J-K)} \times 2^K$, where its $\boldsymbol{y}_t^{\mathbb{J}_2}$-th row and $\boldsymbol{\alpha}_t$-th column element is $P(\boldsymbol{Y}_t^{\mathbb{J}_2} = \boldsymbol{y}_t^{\mathbb{J}_2} | \boldsymbol{\alpha}_t, \boldsymbol{Q}^*, \boldsymbol{s}, \boldsymbol{g})$. Therefore, the emission matrix $\boldsymbol{B}$ can be decomposed as

$$\boldsymbol{B} = \boldsymbol{B}^{\mathbb{J}_1} \odot \boldsymbol{B}^{\mathbb{J}_2}, \tag{B4}$$

where $\odot$ represents column-wise tensor product, which is defined next.

Definition 4. (Khatri–Rao product; Khatri and Rao (1968)) Given matrices $\boldsymbol{U} \in \mathbb{R}^{m_1 \times n}$ and $\boldsymbol{V} \in \mathbb{R}^{m_2 \times n}$ with columns $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$, respectively, their Khatri–Rao tensor product is denoted by $\boldsymbol{U} \odot \boldsymbol{V}$. The result is a matrix of size $(m_1 m_2) \times n$

$$\boldsymbol{U} \odot \boldsymbol{V} = \begin{bmatrix} \mathbf{u}_1 \otimes \mathbf{v}_1 & \mathbf{u}_2 \otimes \mathbf{v}_2 & \cdots & \mathbf{u}_n \otimes \mathbf{v}_n \end{bmatrix}.$$

Remark 8. If $\mathbf{u}$ and $\mathbf{v}$ are vectors, then the Khatri–Rao product and Kronecker product are identical, i.e., $\mathbf{u} \odot \mathbf{v} = \mathbf{u} \otimes \mathbf{v}$.

We can represent $\boldsymbol{B}^{\mathbb{J}_1}$ as the Kronecker product of K 2×2 sub-matrices (Chen, Culpepper, & Liang, 2020)

$$\boldsymbol{B}^{\mathbb{J}_1} = \bigotimes_{j=1}^{K} \begin{bmatrix} 1-g_j & s_j \\ g_j & 1-s_j \end{bmatrix} := \bigotimes_{j=1}^{K} \boldsymbol{B}_j^{\mathbb{J}_1}. \tag{B5}$$

Condition '$g_j \neq 1 - s_j$' in Theorem 5 implies that $rank(\boldsymbol{B}_j^{\mathbb{J}_1}) = 2$ for all j. Then, according to the property of the rank of a Kronecker product, we have $rank(\boldsymbol{B}^{\mathbb{J}_1}) = \prod_{j=1}^{K} rank(\boldsymbol{B}_j^{\mathbb{J}_1}) = 2^K$, which implies that $\boldsymbol{B}^{\mathbb{J}_1}$ is a full rank matrix.
For the decomposition in Eq. (B4), we have

$$\boldsymbol{B}^{\mathbb{J}_1} \odot \boldsymbol{B}^{\mathbb{J}_2} = \begin{bmatrix} \boldsymbol{B}^{\mathbb{J}_1} \boldsymbol{D}_1(\boldsymbol{B}^{\mathbb{J}_2}) \\ \boldsymbol{B}^{\mathbb{J}_1} \boldsymbol{D}_2(\boldsymbol{B}^{\mathbb{J}_2}) \\ \vdots \\ \boldsymbol{B}^{\mathbb{J}_1} \boldsymbol{D}_{2^J-K}(\boldsymbol{B}^{\mathbb{J}_2}) \end{bmatrix}, \tag{B6}$$

where $\boldsymbol{D}_k(\boldsymbol{B}^{\mathbb{J}_2})$ denotes the diagonal matrix with the k-th row of $\boldsymbol{B}^{\mathbb{J}_2}$ lying on its diagonal. Here $\boldsymbol{D}_1(\boldsymbol{B}^{\mathbb{J}_2})$ has full rank since $s_j, 1-s_j, g_j, 1-g_j$ are nonzero, which implies that

$$rank(\boldsymbol{B}^{\mathbb{J}_1} \boldsymbol{D}_1(\boldsymbol{B}^{\mathbb{J}_2})) = \min(rank(\boldsymbol{D}_1(\boldsymbol{B}^{\mathbb{J}_2})),\ rank(\boldsymbol{B}^{\mathbb{J}_1})) = 2^K, \tag{B7}$$

then $\det(\boldsymbol{B}^{\mathbb{J}_1} \boldsymbol{D}_1(\boldsymbol{B}^{\mathbb{J}_2}))$ is a nonzero minor of $\boldsymbol{B}^{\mathbb{J}_1} \odot \boldsymbol{B}^{\mathbb{J}_2}$ with order 2^K, so we have

$$rank(\boldsymbol{B}) = rank(\boldsymbol{B}^{\mathbb{J}_1} \odot \boldsymbol{B}^{\mathbb{J}_2}) = 2^K.$$

Also $\pi_c > 0$ for all c in Theorem 5. Therefore, the strict identifiability condition in Theorem 1 is satisfied, and the restricted HMM with $T \geq 3$ is identified. This completes the proof of part (a) of Theorem 5.
Without the condition '$g_j \neq 1 - s_j$', $\boldsymbol{B}$ has full column rank unless there exists at least one j^*, such that $g_{j^*} = 1 - s_{j^*}$. Then, the dimension of this exceptional set is less than the dimension of $\boldsymbol{\Omega}(\boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q})$, hence of Lebesgue measure zero. Therefore, the generic identifiability condition in Theorem 2 is satisfied, and the restricted HMM with $T \geq 3$ is generically identified. This completes the proof of part (a) of Theorem 6.

Appendix C: Proof of Part (b) of Theorems 1 and 2 ($T = 2$ case)

The proof is based on Kruskal (1977) for the uniqueness of three-way arrays and its application on the identifiability conditions of three-variate latent class models discussed in Allman et al. (2009).

We start from representing the marginal distribution of $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)^\top$ as a three-way array by decomposing $\boldsymbol{Y}_2$ into two parts as shown in Eq. (B3):

$$\begin{aligned}\boldsymbol{T}(\boldsymbol{y}_1, \boldsymbol{y}_2^{\mathbb{J}_1}, \boldsymbol{y}_2^{\mathbb{J}_2}) &= P(\boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2^{\mathbb{J}_1} = \boldsymbol{y}_2^{\mathbb{J}_1}, \boldsymbol{Y}_2^{\mathbb{J}_2} = \boldsymbol{y}_2^{\mathbb{J}_2} \mid \boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{B}) \\ &= \sum_{\boldsymbol{\alpha}_2} \pi_{\boldsymbol{\alpha}_2} P(\boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2^{\mathbb{J}_1} = \boldsymbol{y}_2^{\mathbb{J}_1}, \boldsymbol{Y}_2^{\mathbb{J}_2} = \boldsymbol{y}_2^{\mathbb{J}_2} \mid \boldsymbol{\omega}, \boldsymbol{B}, \boldsymbol{\alpha}_2) \\ &= \sum_{\boldsymbol{\alpha}_2} \pi_{\boldsymbol{\alpha}_2} P(\boldsymbol{Y}_1 = \boldsymbol{y}_1 \mid \boldsymbol{\omega}, \boldsymbol{B}, \boldsymbol{\alpha}_2) P(\boldsymbol{Y}_2^{\mathbb{J}_1} = \boldsymbol{y}_2^{\mathbb{J}_1} \mid \boldsymbol{\omega}, \boldsymbol{B}, \boldsymbol{\alpha}_2) P(\boldsymbol{Y}_2^{\mathbb{J}_2} = \boldsymbol{y}_2^{\mathbb{J}_2} \mid \boldsymbol{\omega}, \boldsymbol{B}, \boldsymbol{\alpha}_2).\end{aligned} \tag{C1}$$

As shown in Bonhomme et al. (2016), we let $\boldsymbol{A} = \boldsymbol{B} \cdot diag(\boldsymbol{\pi}) \cdot \boldsymbol{\omega} \cdot diag(\boldsymbol{\pi})^{-1}$ denote the distribution of $\boldsymbol{Y}_1$ given values of $\boldsymbol{\alpha}_2$ (attribute profile at time point 2). Then, the identifiability is equivalent to the uniqueness of the decomposition of the following tensor (Kruskal, 1977):

$$\boldsymbol{T} = \sum_{\boldsymbol{\alpha}_2} \pi_{\boldsymbol{\alpha}_2} \boldsymbol{A}_{\boldsymbol{\alpha}_2} \odot \boldsymbol{B}_{\boldsymbol{\alpha}_2}^{\mathbb{J}_1} \odot \boldsymbol{B}_{\boldsymbol{\alpha}_2}^{\mathbb{J}_2} = \sum_{\boldsymbol{\alpha}_2} \tilde{\boldsymbol{A}}_{\boldsymbol{\alpha}_2} \odot \boldsymbol{B}_{\boldsymbol{\alpha}_2}^{\mathbb{J}_1} \odot \boldsymbol{B}_{\boldsymbol{\alpha}_2}^{\mathbb{J}_2}, \tag{C2}$$

where $\boldsymbol{A}_{\boldsymbol{\alpha}_2}$, $\boldsymbol{B}_{\boldsymbol{\alpha}_2}^{\mathbb{J}_1}$, $\boldsymbol{B}_{\boldsymbol{\alpha}_2}^{\mathbb{J}_2}$ are the $\boldsymbol{\alpha}_2$-th column of $\boldsymbol{A}$, $\boldsymbol{B}^{\mathbb{J}_1}$, $\boldsymbol{B}^{\mathbb{J}_2}$, and $\tilde{\boldsymbol{A}}_{\boldsymbol{\alpha}_2} = \pi_{\boldsymbol{\alpha}_2} \boldsymbol{A}_{\boldsymbol{\alpha}_2}$.
Next, we give the definition of Kruskal rank and state the theorem in Kruskal (1977) for our setting.

Definition 5. For a matrix $\boldsymbol{M}$, the Kruskal rank of $\boldsymbol{M}$, i.e., $rank_K(\boldsymbol{M})$, is the largest number I such that every set of I columns in $\boldsymbol{M}$ are linearly independent.

Remark 9. Compared with the rank of a matrix $\boldsymbol{M}$, we have $rank_K(\boldsymbol{M}) \leq rank(\boldsymbol{M})$. If $\boldsymbol{M}$ has full column rank, then the equality holds.

Theorem 7. *(Kruskal (1977)) If*

$$rank_K(\tilde{\boldsymbol{A}}) + rank_K(\boldsymbol{B}^{\mathbb{J}_1}) + rank_K(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2 \cdot 2^K + 2, \tag{C3}$$

then the tensor decomposition of $\boldsymbol{T}$ is unique up to simultaneous permutation and rescaling of the rows.

Since $\boldsymbol{\pi}$ has all positive entries, then we have $rank_K(\tilde{\boldsymbol{A}}) = rank_K(\boldsymbol{A})$. Moreover, $\boldsymbol{A}$, $\boldsymbol{B}^{\mathbb{J}_1}$ and $\boldsymbol{B}^{\mathbb{J}_2}$ are all stochastic matrices with column sum 1, so the decomposition of the tensor $\boldsymbol{T}$ is unique up to state label swapping if (C3) in Theorem 7 is satisfied.
Bonhomme et al. (2016) established strict identifiability of HMMs for $T > 2$. We next establish sufficient conditions for the identifiability of the restricted HMM with $T = 2$. Since $\boldsymbol{A} = \boldsymbol{B} \cdot diag(\boldsymbol{\pi}) \cdot \boldsymbol{\omega} \cdot diag(\boldsymbol{\pi})^{-1}$, the rank conditions on $\boldsymbol{B}^{\mathbb{J}_1}$ and $\boldsymbol{\omega}$ imply that $\boldsymbol{A}$ also has full column rank 2^K. Therefore, given $rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K$ and $rank_K(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2$, the HMM with $T = 2$ is identified by Theorem 7. This completes the proof of part (b) of Theorem 1.
Following the similar idea as the proof for Theorem 2 in Appendix A, we need to show that $rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K$, $rank(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2$ and $rank(\boldsymbol{\omega}) = 2^K$ hold almost everywhere in

$$\boldsymbol{\Omega}'_{\boldsymbol{\omega},\boldsymbol{B}} = \{(\boldsymbol{\omega}, \boldsymbol{B}) : \ \boldsymbol{\omega} \in \boldsymbol{\Omega}(\boldsymbol{\omega}), \ \boldsymbol{B} \in \boldsymbol{\Omega}(\boldsymbol{B}), \ rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K \text{ and } rank_K(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2\},$$

which implies that the restricted HMM with $T = 2$ is generically identified.

Let $f'(\boldsymbol{B},\boldsymbol{\omega})=\sum_M([\boldsymbol{B}\cdot\boldsymbol{\omega}]_M)^2:\ \boldsymbol{\Omega}'_{\boldsymbol{\omega},\boldsymbol{B}}\to\mathbb{R}$, then the zero set of $f'(\boldsymbol{B},\boldsymbol{\omega})$ is

$$\begin{aligned}\boldsymbol{Z}_{f'}&=\{(\boldsymbol{\omega},\boldsymbol{B}):(\boldsymbol{\omega},\boldsymbol{B})\in\boldsymbol{\Omega}'_{\boldsymbol{\omega},\boldsymbol{B}}\text{ and }f'(\boldsymbol{B},\boldsymbol{\omega})=0\}\\&=\{(\boldsymbol{\omega},\boldsymbol{B}):(\boldsymbol{\omega},\boldsymbol{B})\in\boldsymbol{\Omega}'_{\boldsymbol{\omega},\boldsymbol{B}}\text{ and }[\boldsymbol{B}\cdot\boldsymbol{\omega}]_M=0\text{ for all possible }M\}\\&=\{(\boldsymbol{\omega},\boldsymbol{B}):\ \boldsymbol{\omega}\in\boldsymbol{\Omega}(\boldsymbol{\omega}),\ \boldsymbol{B}\in\boldsymbol{\Omega}(\boldsymbol{B}),\ rank(\boldsymbol{B}^{\mathbb{J}_1})=2^K,\ rank_K(\boldsymbol{B}^{\mathbb{J}_2})\geq 2\text{ and }rank(\boldsymbol{\omega})<2^K\}.\end{aligned}\tag{C4}$$

As shown in Appendix B, $rank(\boldsymbol{B}^{\mathbb{J}_1})=2^K$ and $rank_K(\boldsymbol{B}^{\mathbb{J}_2})\geq 2$ imply $rank(\boldsymbol{B})=rank(\boldsymbol{B}^{\mathbb{J}_1}\odot\boldsymbol{B}^{\mathbb{J}_2})=2^K$. By Proposition 2, we know that $f'(\boldsymbol{B},\boldsymbol{\omega})$ is not a zero function. Then, by Lemma 1, the zero set $\boldsymbol{Z}'_f$ has measure zero within $\boldsymbol{\Omega}'_{\boldsymbol{\omega},\boldsymbol{B}}$. So the restricted HMM with $T=2$ is generically identified. This completes the proof of part(b) of Theorem 2.

Appendix D: Proof of Part (b) of Theorems 5 and 6 ($T=2$ case)

We first introduce the following two propositions.

Proposition 4. *$rank(\boldsymbol{B}^{\mathbb{J}_1})=2^K$ if and only if $\boldsymbol{Q}^{\mathbb{J}_1}=\boldsymbol{I}_K$ and $g_j\neq 1-s_j$ for $j\in\{1,\ldots,K\}$.*

Proof. According to Eq. (B5), we know that $\boldsymbol{Q}^{\mathbb{J}_1}=\boldsymbol{I}_K$ and $g_j\neq 1-s_j$ for $j\in\{1,\ldots,K\}$ imply $rank(\boldsymbol{B}^{\mathbb{J}_1})=2^K$. On the other hand, $rank(\boldsymbol{B}^{\mathbb{J}_1})=2^K$ implies that the columns of $\boldsymbol{B}^{\mathbb{J}_1}$ are distinct. The $\boldsymbol{\alpha}^\top\boldsymbol{v}=c$ column of $\boldsymbol{B}^{\mathbb{J}_1}$ is

$$\boldsymbol{B}^{\mathbb{J}_1}_c=\bigotimes_{j=1}^{K}\left[(1-g_j)^{1-\eta_{jc}}s_j^{\eta_{jc}},\ g_j^{1-\eta_{jc}}(1-s_j)^{\eta_{jc}}\right],$$

where $\eta_{jc}=\mathcal{I}\left(\boldsymbol{\alpha}_t^\top\boldsymbol{q}_j\geq\boldsymbol{q}_j^\top\boldsymbol{q}_j,\ \boldsymbol{\alpha}_t^\top\boldsymbol{v}=c\right)$, $c=0,\ldots,2^K-1$. Therefore, for $k=1,\ldots,K$, a full rank $\boldsymbol{B}^{\mathbb{J}_1}$ implies that $\boldsymbol{B}^{\mathbb{J}_1}_0\neq\boldsymbol{B}^{\mathbb{J}_1}_{\boldsymbol{e}_k^\top\boldsymbol{v}}$ only if there exists at least one row $\boldsymbol{q}_j$ in $\boldsymbol{Q}^{\mathbb{J}_1}$ satisfying $\boldsymbol{q}_j=\boldsymbol{e}_k$ given $g_j\neq 1-s_j$. Therefore, we must have $\boldsymbol{Q}^{\mathbb{J}_1}=\boldsymbol{I}_K$ after a permutation of rows in $\boldsymbol{Q}^{\mathbb{J}_1}$ and $g_j\neq 1-s_j$ for all $j\in\{1,\ldots,K\}$. □

Proposition 5. *$rank_K(\boldsymbol{B}^{\mathbb{J}_2})\geq 2$ if and only if $\boldsymbol{Q}^{\mathbb{J}_2}$ contains at least one $\boldsymbol{I}_K$ after a row permutation.*

Proof. Similar to the proof in Appendix B, we can prove that $rank_K(\boldsymbol{B}^{\mathbb{J}_2})=2^K\geq 2$ given $\boldsymbol{Q}^{\mathbb{J}_2}$ contains at least one $\boldsymbol{I}_K$ after a row permutation.
Given $rank_K(\boldsymbol{B}^{\mathbb{J}_2})\geq 2$, we know that every two columns in $\boldsymbol{B}^{\mathbb{J}_2}$ are linearly independent according to Definition 5. Assume that for some $k\in\{1,\ldots,K\}$, there does not exist a row in $\boldsymbol{Q}^{\mathbb{J}_2}$ satisfying $\boldsymbol{q}_j=\boldsymbol{e}_k$, then we would have $\boldsymbol{B}^{\mathbb{J}_2}_0=\boldsymbol{B}^{\mathbb{J}_2}_{\boldsymbol{e}_k^\top\boldsymbol{v}}$, which contradicts with the condition $rank_K(\boldsymbol{B}^{\mathbb{J}_2})\geq 2$. Therefore, $\boldsymbol{Q}^{\mathbb{J}_2}$ must contain at least one $\boldsymbol{I}_K$ after a row permutation. □

In Proposition 3, we already showed that if emission matrix $\boldsymbol{B}$ is identified, then parameters $\boldsymbol{s}$, $\boldsymbol{g}$ and $\boldsymbol{Q}$ in the restricted HMM can also be identified. Then, by Propositions 4 and 5, conditions (b) in Theorems 1 and 2 are all satisfied, which proves Theorems 5 and 6 with $T=2$.

Appendix E: Gibbs Sampling Step in Algorithm 1

The full conditional distributions of the parameters are shown as follows. For the attribute profiles $\boldsymbol{\alpha}_{it}$, at time point $t = 1$, given $\boldsymbol{\alpha}_{i2} = \boldsymbol{\alpha}_{c2}$,

$$P(\boldsymbol{\alpha}_{i1} = \boldsymbol{\alpha}_{c'1}|\boldsymbol{Y}_i, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\omega}, \boldsymbol{\alpha}_{i2} = \boldsymbol{\alpha}_{c2}) \propto p(\boldsymbol{Y}_{i1}|\boldsymbol{\alpha}_{i1} = \boldsymbol{\alpha}_{c'1}, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}) \cdot \pi_{c'} \cdot \omega_{c|c'}. \quad \text{(E1)}$$

For $1 < t < T$, given $\boldsymbol{\alpha}_{i,t-1} = \boldsymbol{\alpha}_{c,t-1}$ and $\boldsymbol{\alpha}_{i,t+1} = \boldsymbol{\alpha}_{c',t+1}$,

$$\begin{aligned} &P(\boldsymbol{\alpha}_{it} = \boldsymbol{\alpha}_{lt}|\boldsymbol{Y}_i, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\omega}, \boldsymbol{\alpha}_{c,t-1}, \boldsymbol{\alpha}_{c',t+1}) \\ &\quad \propto p(\boldsymbol{Y}_i|\boldsymbol{\alpha}_l, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}) p(\boldsymbol{\alpha}_{lt}|\boldsymbol{\alpha}_{c,t-1}, \boldsymbol{\omega}) p(\boldsymbol{\alpha}_{c',t+1}|\boldsymbol{\alpha}_{lt}, \boldsymbol{\omega}) \\ &\quad \propto p(\boldsymbol{Y}_{it}|\boldsymbol{\alpha}_{lt}, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}) \cdot \omega_{l|c} \cdot \omega_{c'|l}. \end{aligned} \quad \text{(E2)}$$

At time point $t = T$, given $\boldsymbol{\alpha}_{i,T-1} = \boldsymbol{\alpha}_{c,T-1}$,

$$P(\boldsymbol{\alpha}_{iT} = \boldsymbol{\alpha}_{lT}|\boldsymbol{Y}_i, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\omega}, \boldsymbol{\alpha}_{c,T-1}) \propto p(\boldsymbol{Y}_{iT}|\boldsymbol{\alpha}_{lT}, \boldsymbol{Q}, \boldsymbol{s}, \boldsymbol{g}) \cdot \omega_{l|c}. \quad \text{(E3)}$$

For other parameters, we have

$$p(\boldsymbol{\pi}_1|\boldsymbol{\alpha}_1) \propto \left(\prod_{i=1}^{N} \prod_{l=0}^{2^K-1} \pi_{1,l}^{\mathcal{I}(\boldsymbol{\alpha}_{i1}=\boldsymbol{\alpha}_l)} \right) p(\boldsymbol{\pi}_1) \propto \prod_{l=0}^{2^K-1} \pi_{1,l}^{\tilde{N}_{0,l}+\delta_{0,l}-1}, \quad \text{(E4)}$$

$$\begin{aligned} p(\boldsymbol{\omega}|\boldsymbol{\alpha}) &\propto \left(\prod_{l=0}^{2^K-1} \prod_{c=0}^{2^K-1} \prod_{i=1}^{N} \prod_{t=2}^{T} P(\boldsymbol{\alpha}_{it} = \boldsymbol{\alpha}_c|\boldsymbol{\alpha}_{i,t-1} = \boldsymbol{\alpha}_l, \boldsymbol{\omega}_l) \right) p(\boldsymbol{\omega}) \\ &\propto \prod_{l=0}^{2^K-1} \left(\prod_{c=0}^{2^K-1} \omega_{c|l}^{\tilde{N}_{c|l}} \right) p(\boldsymbol{\omega}_l) \propto \prod_{l=0}^{2^K-1} \prod_{c=0}^{2^K-1} \omega_{c|l}^{\tilde{N}_{c|l}+\delta_{c|l}-1}, \end{aligned} \quad \text{(E5)}$$

$$p(\boldsymbol{Q}|\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\omega}) \propto p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{Q}) \cdot \mathcal{I}(\boldsymbol{Q} \in \mathcal{Q}). \quad \text{(E6)}$$

Details about some of the prior and posterior distributions of parameters shown above could be found in Chen, Culpepper, Wang, and Douglas (2018).

Appendix F: Proof of Theorems 3 and 4

Proof. To prove part (a) of Theorem 3, we will apply Theorem 1 which requires $rank(\boldsymbol{B}) = 2^K$. In Appendix B, we decompose matrix $\boldsymbol{B}$ into two parts: $\boldsymbol{B} = \boldsymbol{B}^{\mathbb{J}_1} \odot \boldsymbol{B}^{\mathbb{J}_2}$. Since emission probabilities in matrix $\boldsymbol{B}$ are all positive due to the CDF $\Psi(\cdot)$, then it is sufficient to show that $rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K$. With condition (B1), we have $\boldsymbol{D} = (\boldsymbol{1}_K, \boldsymbol{I}_K, \boldsymbol{0})$, which implies a DINA model with K skills, K items and $\boldsymbol{Q} = \boldsymbol{I}_K$. Then, similar to Eq. (B5), we can rewrite the first part of the emission matrix $\boldsymbol{B}^{\mathbb{J}_1}$ as

$$\boldsymbol{B}^{\mathbb{J}_1} = \bigotimes_{j=1}^{K} \begin{bmatrix} 1-g_j & s_j \\ g_j & 1-s_j \end{bmatrix} = \bigotimes_{j=1}^{K} \begin{bmatrix} 1-\Psi(\beta_{j,0}) & 1-\Psi(\beta_{j,0}+\beta_{j,j}) \\ \Psi(\beta_{j,0}) & \Psi(\beta_{j,0}+\beta_{j,j}) \end{bmatrix} := \bigotimes_{j=1}^{K} \boldsymbol{B}_j^{\mathbb{J}_1}. \quad \text{(F1)}$$

Under condition (B1), we have $\Psi(\beta_{j,0}) \neq \Psi(\beta_{j,0}+\beta_{j,j})$, which implies $rank(\boldsymbol{B}^{\mathbb{J}_1}) = \prod_{j=1}^{K} rank(\boldsymbol{B}_j^{\mathbb{J}_1}) = 2^K$, so part (a) of Theorem 3 holds based on part (a) of Theorem 1. Also, part (a) of Theorem 4 can be proved similarly using the argument above and part (a) of Theorem 2.

To prove part (b) of Theorem 3, we will again apply Theorem 1 which requires $rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K$ and $rank_K(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2$. According to the proof shown above, we have $rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K$ under condition (B1). Under condition (B2), we have $rank_K(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2$, so part (b) of Theorem 3 holds based on part (b) of Theorem 1. Also, by the similar argument and part (b) of Theorem 2, we can prove that part (b) of Theorem 4 holds under conditions (B1)-(B2). □

Appendix G: Proof of Remark 1

We start from representing the marginal distribution of $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)^\top$ as a three-way array by decomposing $\boldsymbol{Y}_1$ into two parts as shown in Eq. (B3):

$$\begin{aligned}\boldsymbol{T}(\boldsymbol{y}_1^{\mathbb{J}_1}, \boldsymbol{y}_1^{\mathbb{J}_2}, \boldsymbol{y}_2) &= P(\boldsymbol{Y}_1^{\mathbb{J}_1} = \boldsymbol{y}_1^{\mathbb{J}_1}, \boldsymbol{Y}_1^{\mathbb{J}_2} = \boldsymbol{y}_1^{\mathbb{J}_2}, \boldsymbol{Y}_2 = \boldsymbol{y}_2 \mid \boldsymbol{\pi}_1, \boldsymbol{\omega}, \boldsymbol{B}) \\ &= \sum_{\boldsymbol{\alpha}_1} \pi_{1,\boldsymbol{\alpha}_1} P(\boldsymbol{Y}_1^{\mathbb{J}_1} = \boldsymbol{y}_1^{\mathbb{J}_1}, \boldsymbol{Y}_1^{\mathbb{J}_2} = \boldsymbol{y}_1^{\mathbb{J}_2}, \boldsymbol{Y}_2 = \boldsymbol{y}_2 \mid \boldsymbol{\omega}, \boldsymbol{B}, \boldsymbol{\alpha}_1) \\ &= \sum_{\boldsymbol{\alpha}_1} \pi_{1,\boldsymbol{\alpha}_1} P(\boldsymbol{Y}_1^{\mathbb{J}_1} = \boldsymbol{y}_1^{\mathbb{J}_1} \mid \boldsymbol{\omega}, \boldsymbol{B}, \boldsymbol{\alpha}_1) P(\boldsymbol{Y}_1^{\mathbb{J}_2} = \boldsymbol{y}_1^{\mathbb{J}_2} \mid \boldsymbol{\omega}, \boldsymbol{B}, \boldsymbol{\alpha}_1) P(\boldsymbol{Y}_2 = \boldsymbol{y}_2 \mid \boldsymbol{\omega}, \boldsymbol{B}, \boldsymbol{\alpha}_1).\end{aligned} \tag{G1}$$

As shown in Bonhomme et al. (2016), we let $\boldsymbol{A}^* = \boldsymbol{B} \cdot \boldsymbol{\omega}^\top$ denote the distribution of $\boldsymbol{Y}_2$ given values of $\boldsymbol{\alpha}_1$ (attribute profile at time point 1). Then, the identifiability is equivalent to the uniqueness of the decomposition of the following tensor (Kruskal, 1977):

$$\boldsymbol{T} = \sum_{\boldsymbol{\alpha}_1} \pi_{1,\boldsymbol{\alpha}_1} \boldsymbol{B}_{\boldsymbol{\alpha}_1}^{\mathbb{J}_1} \odot \boldsymbol{B}_{\boldsymbol{\alpha}_1}^{\mathbb{J}_2} \odot \boldsymbol{A}_{\boldsymbol{\alpha}_1}^* = \sum_{\boldsymbol{\alpha}_1} \tilde{\boldsymbol{B}}_{\boldsymbol{\alpha}_1}^{\mathbb{J}_1} \odot \boldsymbol{B}_{\boldsymbol{\alpha}_1}^{\mathbb{J}_2} \odot \boldsymbol{A}_{\boldsymbol{\alpha}_1}^*, \tag{G2}$$

where $\boldsymbol{A}_{\boldsymbol{\alpha}_1}^*$, $\boldsymbol{B}_{\boldsymbol{\alpha}_1}^{\mathbb{J}_1}$, $\boldsymbol{B}_{\boldsymbol{\alpha}_1}^{\mathbb{J}_2}$ are the $\boldsymbol{\alpha}_1$-th column of $\boldsymbol{A}^*$, $\boldsymbol{B}^{\mathbb{J}_1}$, $\boldsymbol{B}^{\mathbb{J}_2}$, and $\tilde{\boldsymbol{B}}_{\boldsymbol{\alpha}_1}^{\mathbb{J}_1} = \pi_{1,\boldsymbol{\alpha}_1} \boldsymbol{B}_{\boldsymbol{\alpha}_1}^{\mathbb{J}_1}$. Since $\boldsymbol{\pi}_1$ has all positive entries, then we have $rank_K(\tilde{\boldsymbol{B}}^{\mathbb{J}_1}) = rank_K(\boldsymbol{B}^{\mathbb{J}_1})$. Moreover, $\boldsymbol{A}^*$, $\boldsymbol{B}^{\mathbb{J}_1}$ and $\boldsymbol{B}^{\mathbb{J}_2}$ are all stochastic matrices with column sum 1, so by Theorem 7, the decomposition of the tensor $\boldsymbol{T}$ is unique up to state label swapping if $rank_K(\boldsymbol{A}^*) + rank_K(\boldsymbol{B}^{\mathbb{J}_1}) + rank_K(\tilde{\boldsymbol{B}}^{\mathbb{J}_2}) \geq 2 \cdot 2^K + 2$ holds.

We next establish sufficient conditions for the identifiability of the restricted HMM with $T = 2$. Since $\boldsymbol{A}^* = \boldsymbol{B} \cdot \boldsymbol{\omega}^\top$, the rank conditions on $\boldsymbol{B}^{\mathbb{J}_1}$ and $\boldsymbol{\omega}$ imply that $\boldsymbol{A}^*$ also has full column rank 2^K. Therefore, given $rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K$ and $rank_K(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2$, the HMM with $T = 2$ is strictly identified by Theorem 7.

Similar to the proof for Theorem 2 in Appendix A, in order to prove that the restricted HMM with $T = 2$ is generically identified, we need to show that $rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K$, $rank(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2$ and $rank(\boldsymbol{\omega}) = 2^K$ hold almost everywhere in

$$\boldsymbol{\Omega}'_{\boldsymbol{\omega},\boldsymbol{B}} = \{(\boldsymbol{\omega}, \boldsymbol{B}) : \ \boldsymbol{\omega} \in \boldsymbol{\Omega}(\boldsymbol{\omega}), \ \boldsymbol{B} \in \boldsymbol{\Omega}(\boldsymbol{B}), \ rank(\boldsymbol{B}^{\mathbb{J}_1}) = 2^K \text{ and } rank_K(\boldsymbol{B}^{\mathbb{J}_2}) \geq 2\},$$

which is already proved in Appendix C.

Appendix H: Problem set of Elementary Probability Theory

The two sets of questions from R package pks (Heller & Wickelmaier, 2013) are shown as follows.

6.1. The first set of questions

1. A box contains 30 marbles in the following colors: 8 red, 10 black, 12 yellow. What is the probability that a randomly drawn marble is yellow?
2. A bag contains 5-cent, 10-cent, and 20-cent coins. The probability of drawing a 5-cent coin is 0.35, that of drawing a 10-cent coin is 0.25, and that of drawing a 20-cent coin is 0.40. What is the probability that the coin randomly drawn is not a 5-cent coin?
3. A bag contains 5-cent, 10-cent, and 20-cent coins. The probability of drawing a 5-cent coin is 0.20, that of drawing a 10-cent coin is 0.45, and that of drawing a 20-cent coin is 0.35. What is the probability that the coin randomly drawn is a 5-cent coin or a 20-cent coin?
4. In a school, 40% of the pupils are boys and 80% of the pupils are right-handed. Suppose that gender and handedness are independent. What is the probability of randomly selecting a right-handed boy?
5. Given a standard deck containing 32 different cards, what is the probability of not drawing a heart?
6. A box contains 20 marbles in the following colors: 4 white, 14 green, 2 red. What is the probability that a randomly drawn marble is not white?
7. A box contains 10 marbles in the following colors: 2 yellow, 5 blue, 3 red. What is the probability that a randomly drawn marble is yellow or blue?
8. What is the probability of obtaining an even number by throwing a dice?
9. Given a standard deck containing 32 different cards, what is the probability of drawing a 4 in a black suit?
10. A box contains marbles that are red or yellow, small or large. The probability of drawing a red marble is 0.70, the probability of drawing a small marble is 0.40. Suppose that the color of the marbles is independent of their size. What is the probability of randomly drawing a small marble that is not red?
11. In a garage there are 50 cars, 20 are black and 10 are diesel powered. Suppose that the color of the cars is independent of the kind of fuel. What is the probability that a randomly selected car is not black and it is diesel powered?
12. A box contains 20 marbles, 10 marbles are red, 6 are yellow and 4 are black. 12 marbles are small and 8 are large. Suppose that the color of the marbles is independent of their size. What is the probability of randomly drawing a small marble that is yellow or red?

6.2. The Second Set of Questions

1. A box contains 30 marbles in the following colors: 10 red, 14 yellow, 6 green. What is the probability that a randomly drawn marble is green?
2. A bag contains 5-cent, 10-cent, and 20-cent coins. The probability of drawing a 5-cent coin is 0.25, that of drawing a 10-cent coin is 0.60, and that of drawing a 20-cent coin is 0.15. What is the probability that the coin randomly drawn is not a 5-cent coin?
3. A bag contains 5-cent, 10-cent, and 20-cent coins. The probability of drawing a 5-cent coin is 0.35, that of drawing a 10-cent coin is 0.20, and that of drawing a 20-cent coin is 0.45. What is the probability that the coin randomly drawn is a 5-cent coin or a 20-cent coin?

4. In a school, 70% of the pupils are girls and 10% of the pupils are left-handed. Suppose that gender and handedness are independent. What is the probability of randomly selecting a left-handed girl?
5. Given a standard deck containing 32 different cards, what is the probability of not drawing a club?
6. A box contains 20 marbles in the following colors: 6 yellow, 10 red, 4 green. What is the probability that a randomly drawn marble is not yellow?
7. A box contains 10 marbles in the following colors: 5 blue, 3 red, 2 green. What is the probability that a randomly drawn marble is red or blue?
8. What is the probability of obtaining an odd number by throwing a dice?
9. Given a standard deck containing 32 different cards, what is the probability of drawing a 10 in a red suit?
10. A box contains marbles that are red or yellow, small or large. The probability of drawing a green marble is 0.40, the probability of drawing a large marble is 0.20. Suppose that the color of the marbles is independent of their size. What is the probability of randomly drawing a large marble that is not green?
11. In a garage there are 50 cars, 15 are white and 20 are diesel powered. Suppose that the color of the cars is independent of the kind of fuel. What is the probability that a randomly selected car is not white and it is diesel powered?
12. A box contains 20 marbles, 8 marbles are white, 4 are green and 8 are red. 15 marbles are small and 5 are large. Suppose that the color of the marbles is independent of their size. What is the probability of randomly drawing a large marble that is white or green?

References

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics, 37*(6A), 3099–3132.

Baras, J. S., & Finesso, L. (1992). Consistent estimation of the order of hidden Markov chains. In T. E. Duncan & B. Pasik-Duncan (Eds.), *Stochastic theory and adaptive control* (pp. 26–39). Berlin & Heidelberg: Springer.

Blasiak, S., & Rangwala, H. (2011). A hidden Markov model variant for sequence classification. In: Proceedings of the twenty-second international joint conference on artificial intelligence - volume two (1192–1197). AAAI Press.

Bonhomme, S., Jochmans, K., & Robin, J. M. (2016). Estimating multivariate latent-structure models. *The Annals of Statistics, 44*(2), 540–563.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455.

Chen, Y., Culpepper, S., Chen, Y., & Douglas, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika, 83*(1), 89–108.

Chen, Y., Culpepper, S. A., & Liang, F. (2020). A sparse latent class model for cognitive diagnosis, Psychometrika, 85, 121–153.

Chen, Y., Culpepper, S., Wang, S., & Douglas, J. (2018). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement, 42*(1), 5–23.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*(510), 850–866.

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

Chiu, C. Y., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633–665.

Cox, D. A., Little, J., & O'Shea, D. (2015). *Ideals, varieties, and algorithms*. New York: Springer.

Crouse, M. S., Nowak, R. D., & Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing, 46*(4), 886–902.

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. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

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

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301–321.

Heller, J., & Wickelmaier, F. (2013). Minimum discrepancy estimation in probabilistic knowledge structures. *Electronic Notes in Discrete Mathematics, 42*, 49–56.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210.
Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.
Khatri, C. G., & Rao, C. R. (1968). Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhya: The Indian Journal of Statistics, Series A, 30*(2), 167–180.
Kruskal, J. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications, 18*, 95–138.
Lathauwer, L. D., Moor, B. D., & Vandewalle, J. (2004). Computation of the canonical decomposition by means of a simultaneous generalized schur decomposition. *SIAM Journal on Matrix Analysis and Applications, 26*, 295–327.
Matsaglia, G., & Styan, G. P. H. (1974). Equalities and inequalities for ranks of matrices. *Linear and Multilinear Algebra, 2*(3), 269–292.
Paz, A. (1971). Stochastic Sequential Machines. In A. Paz (Ed.), *Introduction to probabilistic automata* (pp. 1–66). Academic Press.
Petrie, T. (1969). Probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics, 40*(1), 97–115.
Sipos, I. R., Ceffer, A., & Levendovszky, J. (2017). Parallel optimization of sparse portfolios with AR-HMMs. *Computational Economics, 49*, 563–578.
Von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307.
Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *Annals of Statistics, 45*(2), 675–707.

Manuscript Received: 22 OCT 2021
Published Online Date: 16 FEB 2023