Statistical analysis of Q-matrix based diagnostic classification models

Article:

Chen, Yunxiao, Liu, Jingchen, Xu, Gongjun and Ying, Zhiliang (2015) Statistical analysis of Q-matrix based diagnostic classification models. Journal of the American Statistical Association, 110 (510). 850 - 866. ISSN 0162-1459

https://doi.org/10.1080/01621459.2014.934827

Statistical Analysis of Q-matrix Based Diagnostic Classification Models

Yunxiao Chen, Jingchen Liu, Gongjun Xu, and Zhiliang Ying

Abstract

Diagnostic classification models have recently gained prominence in educational assessment, psychiatric evaluation, and many other disciplines. Central to the model specification is the so-called Q-matrix that provides a qualitative specification of the item-attribute relationship. In this paper, we develop theories on the identifiability for the Q-matrix under the DINA and the DINO models. We further propose an estimation procedure for the Q-matrix through the regularized maximum likelihood. The applicability of this procedure is not limited to the DINA or the DINO model and it can be applied to essentially all Q-matrix based diagnostic classification models. Simulation studies show that the proposed method admits high probability recovering the true Q-matrix. Furthermore, two case studies are presented. The first case is a data set on fraction subtraction (educational application) and the second case is a subsample of the National Epidemiological Survey on Alcohol and Related Conditions concerning the social anxiety disorder (psychiatric application).

KEY WORDS: diagnostic classification models, identifiability, Q-matrix

1 Introduction

Cognitive diagnosis has recently gained prominence in educational assessment, psychiatric evaluation, and many other disciplines (Rupp and Templin, 2008b; Rupp et al., 2010). A cognitive diagnostic test, consisting of a set of items, provides each subject with a profile detailing the concepts and skills (often called “attributes”) that he/she masters. For instance, teachers identify students’ mastery of different skills (attributes) based on their solutions (responses) to exam questions (items); psychiatrists/psychologists learn patients’ presence/absence of disorders (attributes) based on their responses to diagnostic questions (items). Various diagnostic classification models (DCM) have been developed in the literature. A short list includes the conjunctive DINA and NIDA models (Junker and Sijtsma, 2001; Tatsuoka, 2002; de la Torre and Douglas, 2004; de la Torre, 2011), the reparameterized unified/fusion model (RUM) (DiBello et al., 1995), the compensatory DINO and NIDO models (Templin and Henson, 2006), the rule space method (Tatsuoka, 1985, 2009), the attribute hierarchy method (Leighton et al., 2004), and Generalized DINA models (de la Torre, 2011); see also Henson et al. (2009); Rupp et al. (2010) for more developments and approaches to cognitive diagnosis. The general diagnostic model (von Davier, 2005, 2008; von Davier and Yamamoto, 2004) provides a framework for the development of diagnostic models.

A common feature of these models is that the probabilistic distribution of subjects’ responses to items is governed by their latent attribute profiles. Upon observing the responses, one can make inferences on the latent attribute profiles. The key component in the model specification is the relationship between the observed item responses and the latent attribute profiles. A central quantity in this specification is the so-called Q-matrix. Suppose that there are J items measuring K attributes. Then, the Q-matrix is a J by K matrix with zero-one entries each of which indicates whether an item is associated to an attribute. In the statistical analysis of diagnostic classification models, it is customary to work with a prespecified Q-matrix; for instance, an exam maker specifies the set of skills tested by each exam problem (Tatsuoka, 1990). However, such a specification is usually subjective and may not be accurate. The misspeficiation of the Q-matrix could possibly lead to serious lack of fit and further inaccurate inferences on the latent attribute profiles.

In this paper, we consider an objective construction of the Q-matrix, that is, estimating it based on the data. This estimation problem becomes easy or even trivial if the item responses and the attribute profiles are both observed. However, subjects' attribute profiles are not directly observed and their information can only be extracted from item responses. The estimation of the Q-matrix should be solely based on the dependence structure among item responses. Due to the latent nature of the attribute profiles, when and whether the Q-matrix and other models parameters can be estimated consistently by the observed data under various models specifications is a challenging problem. Furthermore, theoretical results on the identifiability usually do not imply practically feasible estimation procedures. The construction of an implementable estimation procedure is the second objective of this paper.

Following the above discussion, the main contribution of this paper is two-fold. First, we provide identifiability results for the Q-matrix. As we will specify in the subsequent sections, the Q-matrix estimation is equivalent to a latent variable selection problem. Nontrivial conditions are necessary to guarantee the consistent identification of Q-matrix. We present the results for both the DINA and the DINO models that are two important diagnostic classification models. The theoretical results provide the possibility of estimating the Q-matrix, in particular, the consistency of the maximum likelihood estimator (MLE). However, due to the discrete nature of the Q-matrix, MLE requires a substantial computational overhead and it is practically infeasible. The second contribution of this paper is the proposal of a computationally affordable estimator. Formulating Q-matrix estimation into a latent variable selection problem, we propose an estimation procedure via the regularized maximum likelihood. This regularized estimator can be computed by means of a combination of the expectation-maximization algorithm and the coordinate descent algorithm. We emphasize that the applicability of this estimator is not limited to the DINA or the DINO model for which the theoretical results are developed. It can be applied to a large class of diagnostic classification models.

Statistical inference of Q-matrix has been largely an unexplored area in the cognitive assessment literature. Nevertheless, there are a few works related to the current one. Identifiability of the Q-matrix for the DINA model under a specific situation is discussed by Liu et al. (2013). The results require a complete knowledge of the guessing parameter. The theoretical results in the current paper are a natural extension of Liu et al. (2013) to generally all DINA models and further to the DINO model. Furthermore, various diagnosis tools and testing procedures have been developed in the literature (de la Torre and Douglas, 2004; Liu et al., 2007; Rupp and Templin, 2008a; de la Torre, 2008), none of which, however, addresses the estimation problem. In addition to the estimation of the Q-matrix, we discuss the estimation of other model parameters. Although there have been results on estimation (Junker, 1999; Rupp and Templin, 2008b; de la Torre, 2009; Rupp et al., 2010), formal statistical analysis, including rigorous results on identifiability and asymptotic properties, has not been developed.

The rest of the paper is organized as follows. We present the theoretical results for the Q-matrix and other model parameters under DINA and DINO models in Section 2. Section 3 presents a computationally affordable estimation procedure based on regularized maximum likelihood. Simulation studies and real data illustrations are presented in Sections 4 and 5. Detailed proofs are provided in the supplemental material.

2 The identifiability results

2.1 Diagnostic classification models

We consider that there are N subjects, each of whom responds to J items. To simplify the discussion, we assume that the responses are all binary. The analysis of other types of responses can be easily adapted. Diagnostic classification models assume that subject's responses to items are governed by his/her latent (unobserved) attribute profile that is a K-dimensional vector, each entry of which takes values in $\{0,1\}$, that is, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$ and $\alpha_k \in \{0,1\}$. In the context of educational testing, α_k indicates the mastery of skill k. Let $\mathbf{R} = (R^1, ..., R^J)$ denote the vector of responses to the J items. Both $\boldsymbol{\alpha}$ and $\mathbf{R}$ are subject-specific and we will later use subscript to indicate different subjects, that is, $\boldsymbol{\alpha}_i$ and $\mathbf{R}_i$ are the latent attribute profile and response vector of subject i for $i = 1, ..., N$.

The Q-matrix provides a link between the responses to items and the attributes. In particular, $Q = (q_{jk})_{J\times K}$ is a $J \times K$ matrix with binary entries. For each j and k, $q_{jk} = 1$ means that the response to item j is associated to the presence of attribute k and $q_{jk} = 0$ otherwise. The precise relationship depends on the model parameterization.

We use $\boldsymbol{\theta}$ as a generic notation for the unknown item parameters additional to the Q-matrix. Given a specific subject's profile $\boldsymbol{\alpha}$, the response R^j to item j follows a Bernoulli distribution

$$P(R^j|Q,\boldsymbol{\alpha},\boldsymbol{\theta}) = (c_{j,\boldsymbol{\alpha}})^{R^j}(1-c_{j,\boldsymbol{\alpha}})^{1-R^j}, \tag{1}$$

where $c_{j,\boldsymbol{\alpha}}$ is the probability for subjects with attribute profile $\boldsymbol{\alpha}$ to provide a positive response to item j, i.e.,

$$c_{j,\boldsymbol{\alpha}} = P(R^j = 1|Q,\boldsymbol{\alpha},\boldsymbol{\theta}).$$

The specific form of $c_{j,\boldsymbol{\alpha}}$ additionally depends on the Q-matrix, the item parameter vector $\boldsymbol{\theta}$, and the model parameterization. Conditional on $\boldsymbol{\alpha}$, $(R^1, ..., R^J)$ are jointly independent. We further assume that the attribute profiles are i.i.d. following distribution

$$p_{\boldsymbol{\alpha}} = P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}).$$

Let $\mathbf{p} = (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^K)$. In what follows, we present a few examples.

Example 1 (DINA model, Junker and Sijtsma (2001)) *For each item j and attribute vector $\boldsymbol{\alpha}$, we define the ideal response*

$$\xi^j_{DINA}(\boldsymbol{\alpha}, Q) = \prod_{k=1}^{K} (\alpha_k)^{q_{jk}} = I(\alpha_k \geq q_{jk} \text{ for all } k) \tag{2}$$

that is, whether $\boldsymbol{\alpha}$ has all the attributes required by item j. For each item, there are two additional parameters s_j and g_j that are known as the slipping and guessing parameters. The response probability $c_{j,\boldsymbol{\alpha}}$ takes the form

$$c_{j,\boldsymbol{\alpha}} = (1 - s_j)^{\xi^j_{DINA}(\boldsymbol{\alpha},Q)} g_j^{1-\xi^j_{DINA}(\boldsymbol{\alpha},Q)}. \tag{3}$$

If $\xi^j_{DINA}(\boldsymbol{\alpha}, Q) = 1$ (the subject is capable of solving a problem), then the positive response probability is $1 - s_j$; otherwise, the probability is g_j. The item parameter vector is $\boldsymbol{\theta} = \{s_j, g_j : \ j = 1, \cdots, J\}$.

The DINA model assumes a conjunctive (non-compensatory) relationship among attributes. It is necessary to possess all the attributes indicated by the Q-matrix to be capable of providing a positive response. In addition, having additional unnecessary attributes does not compensate for the lack of necessary attributes. The DINA model is popular in the educational testing applications and is often employed for modeling exam problem solving processes.

Example 2 (NIDA model) *The NIDA model admits the following form*

$$c_{j,\boldsymbol{\alpha}} = \prod_{k=1}^{K} [(1 - s_k)^{\alpha_k} g_k^{1-\alpha_k}]^{q_{jk}}.$$

The problem solving involves multiple skills indicated by the Q-matrix. For each skill, the student has a certain probability of implementing it: $1 - s_j$ *for mastery and* g_j *for non-mastery. The problem is solved correctly if all required skills have been implemented correctly by the student, which leads to the above positive response probability.*

The following reduced RUM model is also a conjunctive model, and it generalizes the DINA and the NIDA models by allowing the item parameters to vary among attributes.

Example 3 (Reduced NC-RUM model) *Under the reduced noncompensatory reparameterized unified model (NC-RUM), we have*

$$c_{j,\boldsymbol{\alpha}} = \pi_j \prod_{k=1}^{K} (r_{jk})^{q_{jk}(1-\alpha_k)}, \tag{4}$$

where π_j *is the correct response probability for subjects who possess all required attributes and* $r_{j,k}$*,* $0 < r_{j,k} < 1$*, is the penalty parameter for not possessing the kth attribute. The corresponding item parameters are* $\boldsymbol{\theta} = \{\pi_j, r_{j,k} : j = 1, \cdots, J, k = 1, \cdots, K\}$.

In contrast to the DINA, NIDA, and Reduced NC-RUM models, the following DINO and C-RUM models assume compensatory (non-conjunctive) relationship among attributes, that is, one only needs to possess one of the required attributes to be capable of providing a positive response.

Example 4 (DINO model) *The ideal response of the DINO model is given by*

$$\xi^j_{DINO}(\boldsymbol{\alpha}, Q) = 1 - \prod_{k=1}^{K} (1 - \alpha_k)^{q_{jk}} = I(\alpha_k \geq q_{jk} \textit{ for at least one } k). \tag{5}$$

Similar to the DINA model, the positive response probability is

$$c_{j,\boldsymbol{\alpha}} = (1 - s_j)^{\xi^j_{DINO}(\boldsymbol{\alpha},Q)} g_j^{1-\xi^j_{DINO}(\boldsymbol{\alpha},Q)}.$$

The DINO model is the dual model of the DINA model. The DINO model is often employed in the application of psychiatric assessment, for which the positive response to a diagnostic question (item) could be due to the presence of one disorder (attributes) among several.

Example 5 (C-RUM model) *The GLM-type parametrization with a logistic link function is used for the compensatory reparameterized unified model (C-RUM), that is*

$$c_{j,\boldsymbol{\alpha}} = \frac{\exp(\beta^j_0 + \sum_{k=1}^K \beta^j_k q_{jk}\alpha_k)}{1 + \exp(\beta^j_0 + \sum_{k=1}^K \beta^j_k q_{jk}\alpha_k)}. \tag{6}$$

The corresponding item parameter vector is $\boldsymbol{\theta} = \{\beta^j_k : j = 1, \cdots, J, k = 0, \cdots, K\}$. The C-RUM model is a compensatory model and one can recognize (6) as a structure in multidimensional item response theory model or in factor analysis.

2.2 Some concepts of identifiability

We consider two matrices Q and Q' that are identical if we appropriately rearrange the orders of their columns. Each column in the Q-matrix corresponds to an attribute. Reordering the columns corresponds to relabeling the attributes and it does not change the model. Upon estimating the Q-matrix, the data does not contain information about the specific meaning of each attribute. Therefore, one cannot differentiate Q and Q' solely based on data if there are identical up to a column permutation. For this sake, we present the following equivalent relation.

Definition 1 *We write $Q \sim Q'$ if and only if Q and Q' have identical column vectors that could be arranged in different orders; otherwise, we write $Q \nsim Q'$.*

Definition 2 *We say that Q is identifiable if there exits an estimator $\hat{Q}$ such that*

$$\lim_{N\to\infty} P(\hat{Q} \sim Q) = 1.$$

Given a response vector $\mathbf{R} = (R^1, \cdots, R^j)^\top$, the likelihood function of a diagnostic classification model can be written as

$$L(\boldsymbol{\theta}, \mathbf{p}, Q) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} P(R^j = 1|\boldsymbol{\theta}, \boldsymbol{\alpha}, Q)^{R^j} (1 - P(R^j = 1|\boldsymbol{\theta}, \boldsymbol{\alpha}, Q))^{1-R^j}.$$

Definition 3 (Definition 11.2.2 in Casella and Berger (2001)) *For a given Q, we say that the model parameters $\boldsymbol{\theta}$ and $\mathbf{p}$ are identifiable if distinct values of $(\boldsymbol{\theta}, \mathbf{p})$ yield different distributions of $\mathbf{R}$, i.e., there is no $(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{p}}) \neq (\boldsymbol{\theta}, \mathbf{p})$ such that $L(\boldsymbol{\theta}, \mathbf{p}, Q) \equiv L(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{p}}, Q)$ for all $\mathbf{R} \in \{0,1\}^J$.*

Let $\hat{Q}$ be a consistent estimator. Notice that the Q-matrix is a discrete parameter. The uncertainty of $\hat{Q}$ in estimating Q is not captured by its standard deviation or confidence interval type of statistics. It is more natural to consider the probability $P(\hat{Q} \nsim Q)$ that is usually very difficult to compute. Nonetheless, it is believed that $P(\hat{Q} \nsim Q)$ decays exponentially fast as the sample size (total number of subjects) approaches infinity. We do not pursue along this direction in this paper. The parameters $\boldsymbol{\theta}$ and $\mathbf{p}$ are both continuous parameters. As long as they are identifiable, the analysis falls into routine inference framework. That is, the maximum likelihood is asymptotically normal centered around the true value and its covariance matrix is the inverse of the Fisher information matrix. In what follows, we present some technical conditions that will be referred to in the subsequent sections.

A1 $\boldsymbol{\alpha}_1$,...,$\boldsymbol{\alpha}_N$ are independently and identically distributed random vectors following distribution $P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}) = p_{\boldsymbol{\alpha}}$. The population is fully diversified meaning that $p_{\boldsymbol{\alpha}} > 0$ for all $\boldsymbol{\alpha}$.

A2 All items have discriminating power meaning that $1 - s_j > g_j$ for all j.

A3 The true matrix Q_0 is complete meaning that $\{\mathbf{e}_i : i = 1, ..., k\} \subset R_Q$, where R_Q is the set of row vectors of Q and $\mathbf{e}_i$ is a row vector such that the i-th element is one and the rest are zero.

A4 Each attribute is required by at least two items, that is, $\sum_{j=1}^{J} q_{jk} \geq 2$ for all k.

The completeness of the Q-matrix requires that for each attribute there exists at least one item requiring only that attribute. If Q is complete, then we can rearrange row and column orders (corresponding to reordering the items and attributes) such that it takes the following form

$$Q = \begin{pmatrix} \mathcal{I}_K \\ \cdots \end{pmatrix}, \tag{7}$$

where matrix $\mathcal{I}_K$ is the $K \times K$ identity matrix. Completeness is an important assumption throughout the subsequent discussion. Without loss of generality, we assume that the rows and columns of the Q-matrix have been rearranged such that it takes the above form.

Remark 1 *One of the main objectives of cognitive diagnosis is to identify subjects' attribute profiles. It has been established that completeness is a sufficient and necessary condition for a set of items to consistently identify all types of attribute profiles for the DINA model when the slipping and the guessing parameters are both zero. It is usually recommended to use a complete Q-matrix. More discussions regarding this issue can be found in Chiu et al. (2009) and Liu et al. (2013).*

2.3 Identifiability of Q-matrix for the DINA and the DINO model

We consider the models in Examples 1 and 4 and start the discussion by citing the main result of Liu et al. (2013).

Theorem 1 (Theorem 4.2, Liu et al. (2013)) *For the DINA model, if the guessing parameters g_j's are known, under Conditions A1, A2, and A3, the Q-matrix is identifiable.*

The first result in this paper generalizes Theorem 1 to the DINO model with a known slipping parameter. In addition, we provide sufficient and necessary conditions for the identifiability of the slipping and guessing parameters.

Theorem 2 *For the DINO model with known slipping parameters, under Conditions A1, A2, and A3, the Q-matrix is identifiable; the guessing parameters g_j and the attribute population $\mathbf{p}$ are identifiable if and only if Condition A4 holds.*

Furthermore, under the setting of Theorem 1, the slipping parameters s_j and the attribute population parameter $\mathbf{p}$ are identifiable if and only if Condition A4 holds.

Theorems 1 and 2 require the knowledge of the guessing parameter (the DINA model) or the slipping parameter (the DINO model). They are applicable under certain situations. In the educational testing context, some testing problems are difficult to guess, for instance, the guessing probability of "$879 \times 234 =$?" is basically zero; for multiple choice problems, if all the choices look "equally correct," then the guessing probability may be set to one over the number of choices.

We further extend the results to the situation when neither the slipping nor the guessing parameters is known, for which additional conditions are required.

A5 Each attribute of the Q-matrix is associated to at least three items, that is, $\sum_{j=1}^{J} q_{jk} \geq 3$ for all k.

A6 Q has two complete submatrices, that is, for each attribute, there exists at least two items requiring only that attribute. If so, we can appropriately arrange the columns and rows such that

$$Q = \begin{pmatrix} \mathcal{I}_K \\ \mathcal{I}_K \\ Q_1 \end{pmatrix}. \tag{8}$$

Theorem 3 *Under the DINA and DINO models with* $(\mathbf{s}, \mathbf{g}, \mathbf{p})$ *unknown, if Conditions A1, 2, 5, and 6 hold, then* Q *is identifiable, i.e., one can construct an estimator* $\hat{Q}$ *such that for all* $(\mathbf{s}, \mathbf{g}, \mathbf{p})$

$$\lim_{N \to \infty} P(\hat{Q} \sim Q) = 1.$$

Theorem 4 *Suppose that Conditions A1, 2, 5, and 6 hold. Then* $\mathbf{s}$*,* $\mathbf{g}$*, and* $\mathbf{p}$ *are all identifiable.*

Theorems 3 and 4 state the identifiability results of Q and other model parameters. They are nontrivial generalizations of Theorems 1 and 2. As we mentioned in the previous section, given that $\mathbf{s}$, $\mathbf{g}$, and $\mathbf{p}$ are identifiable, their estimation falls into routine analysis. The asymptotic distribution of the maximum likelihood estimator and generalized estimating equation estimators are all asymptotically multivariate normal centered around the true values and their variances can be estimated either by the Fisher information inverse or by the sandwich variance estimators.

The identifiability results for Q only state the existence of a consistent estimator. We present the following corollary that the maximum likelihood estimator is consistent under the conditions required by the above theorems. The maximum likelihood estimator (MLE) takes the following form

$$\hat{Q}_{MLE} = \arg \sup_{Q} \sup_{\mathbf{s},\mathbf{g},\mathbf{p}} L_N(\mathbf{s}, \mathbf{g}, \mathbf{p}, Q), \tag{9}$$

where

$$L_N(\mathbf{s},\mathbf{g},\mathbf{p},Q)=\prod_{i=1}^{N}\sum_{\boldsymbol{\alpha}\in\{0,1\}^K}p_{\boldsymbol{\alpha}}\prod_{j=1}^{J}(c_{j,\boldsymbol{\alpha}})^{R_i^j}(1-c_{j,\boldsymbol{\alpha}})^{1-R_i^j}.$$

Corollary 1 *Under the conditions of Theorem 3, $\hat{Q}_{MLE}$ is consistent. Moreover, the maximum likelihood estimator of* $\mathbf{s},\mathbf{g},\mathbf{p}$

$$(\hat{\mathbf{s}},\hat{\mathbf{g}},\hat{\mathbf{p}})=\arg\sup_{\mathbf{s},\mathbf{g},\mathbf{p}}L_N(\mathbf{s},\mathbf{g},\mathbf{p},\hat{Q}_{MLE}) \tag{10}$$

are asymptotically normal with mean centered at the true parameters and variance being the inverse Fisher information matrix.

Proof of Corollary 1. Based on the results and proofs of Theorems 3 and 4, this corollary is straightforward to develop by means of Taylor expansion of the likelihood. We therefore omit the details. ■

To compute the maximum likelihood estimator $\hat{Q}_{MLE}$, one needs to evaluate the profile likelihood, $\sup_{\mathbf{s},\mathbf{g},\mathbf{p}}L_N(\mathbf{s},\mathbf{g},\mathbf{p},Q)$, for all possible J by K matrices with binary entries. The computation of $\hat{Q}_{MLE}$ induces a substantial overhead and is practically impossible to carry out. In the following section, we present a computationally feasible estimator via the regularized maximum likelihood estimator.

Remark 2 *The identifiability results are developed under the situation when there is no information about Q at all. In practice, partial information about the Q-matrix is usually available. For instance, a submatrix for some items (rows) is known and the rest needs to be estimated. This happens when new items are to be calibrated based on existing ones. Sometimes, a submatrix is known for some attributes (columns) and that corresponding to other attributes needs to be learned. This happens when some attributes are concrete and easily recognizable in a given item and the others are subtle and not obvious. Under such circumstances, the Q-matrix is easier to estimate and the identifiability conditions are weaker than those in Theorem 3. We do not pursue the partial information situation in this paper.*

Remark 3 *The equivalent relation "$\sim$" defines the finest equivalent classes, up to which Q can be estimated based on the data without assist of prior knowledge. In this sense, Theorem 4 provides the strongest type of identifiability and in turn it also requires some restrictive conditions. For instance, Condition A6 sometimes is difficult to satisfy in practice and it usually leads to some over simplified items especially when the number of attributes K is large. In that case, the Q-matrix can only be identified up to some weaker equivalence classes. We leave this investigation for future study.*

3 Q-matrix estimation via a regularized likelihood

3.1 Alternative representation of diagnostic classification models via generalized linear models

We first formulate the Q-matrix estimation as a latent variable selection problem and then construct a computationally feasible estimator via the regularized maximum likelihood, for which there is a large body of literature (Tibshirani, 1996, 1997; Fan and Li, 2001). The applicability of this estimator is not limited to the DINA or the DINO models and it can be applied to basically all Q-matrix based diagnostic classification models in use. A short list of such models includes DINA-type models (such as the DINA and HO-DINA models), RUM-type models (like the NC-RUM, reduced NC-RUM, and C-RUM), and the saturated models, the log linear cognitive diagnosis models (LCDM) and generalized DINA (Henson et al., 2009; Rupp et al., 2010; de la Torre, 2011).

In the model specification, the key element is mapping a latent attribute $\boldsymbol{\alpha}$ to a positive response probability, $c_{j,\boldsymbol{\alpha}}$, that additionally depends on the Q-matrix and other model parameters. To motivate the general alternative representation with the DINA model, we consider the following equivalent representation of the DINA model (c.f. (3))

$$\begin{aligned} c_{j,\boldsymbol{\alpha}} &= P(R^j = 1|\boldsymbol{\alpha}, \boldsymbol{\beta}^j) \\ &= \operatorname{logit}^{-1}\Big\{\beta_0^j + \sum_{k=1}^{K} \beta_k^j \alpha_k + \sum_{1\le k_1<k_2\le K} \beta_{k_1k_2}^j \alpha_{k_1}\alpha_{k_2} + \sum_{1\le k_1<k_2<k_3\le K} \beta_{k_1k_2k_3}^j \alpha_{k_1}\alpha_{k_2}\alpha_{k_3} \\ &\qquad + \cdots + \beta_{12\ldots K}^j \prod_{k=1}^{K} \alpha_k\Big\}, \end{aligned} \tag{11}$$

where $\operatorname{logit}(p) = \log \frac{p}{1-p}$ for $p \in (0,1)$. On the right-hand side, inside the logit-inverse function is a function of $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$ with all the interactions. Notice that the response to item j is determined by the underlying attribute $\boldsymbol{\alpha}$. Thus, the above generalized linear representation of $c_{j,\boldsymbol{\alpha}}$ is a saturated model, that is, all diagnostic classification models admitting a K-dimensional attribute profile is a special case of (11).

In what follows, we explain the adaptation of (11) to the DINA model and further to a Q-matrix. The response distribution to each item under the DINA model could be either Bernoulli $(1 - s_j)$ or Bernoulli (g_j) depending on the ideal responses ξ^j. Suppose that item j requires attributes 1, 2,..., and K_j, that is, $q_{jk} = 1$ for all $1 \le k \le K_j$ and 0 otherwise. Then, the positive response probability (3) can be written as

$$c_{j,\boldsymbol{\alpha}} = \operatorname{logit}^{-1}\Big\{\beta_0^j + \beta_{12\ldots K_j}^j \prod_{k=1}^{K_j} \alpha_k\Big\}.$$

Thus, if $\alpha_k = 1$ for all $1 \le k \le K_j$, then $c_{j,\boldsymbol{\alpha}} = 1 - s_j = e^{\beta_0^j + \beta_{12\ldots K_j}^j}/(1 + e^{\beta_0^j+\beta_{12\ldots K_j}^j})$; otherwise, $c_{j,\boldsymbol{\alpha}} = g_j = e^{\beta_0^j}/(1+e^{\beta_0^j})$. Generally speaking, if an item requires attributes $k_1, ..., k_j$, the coefficients β_0^j and $\beta_{k_1\ldots k_j}^j$ are non-zero and all other coefficients are zero. Therefore, each row vector of the Q-matrix, corresponding to the attribute requirement of one item, maps to two non-zero β-coefficients. One of these two coefficients is the intercept β_0^j and the other one is the coefficient for the product of all the required attributes suggested by the Q-matrix.

Therefore, each Q-matrix corresponds to a non-zero pattern of the regression coefficients in (11). Estimating the Q-matrix is equivalent to identifying the non-zero regression coefficients. There is a vast literature on variable and model selection, most of which are developed for linear and generalized linear models. Technically speaking, (11) is a generalized linear mixed model with $\alpha_1, ..., \alpha_K$ and their interactions being the random covariates and β being the regression coefficients. We would employ variable selection methods for the Q-matrix estimation.

Notice that the current setup is different from the regular regression setting in that the covariates α_i's are not directly observed. Therefore the variables to be selected are all latent. The results in the previous section establish sufficient conditions under which the latent variables can be consistently selected. The validity of the methods proposed in this section stands on those theoretical results. We propose the usage of the regularized maximum likelihood estimator. In doing so, we first present the general form of diagnostic classification models. For each item j, the positive response probability given the latent attribute profile admits the following generalized linear form

$$c_{j,\boldsymbol{\alpha}} = g^{-1}([\boldsymbol{\beta}^j]^\top h(\boldsymbol{\alpha})) \tag{12}$$

where $\boldsymbol{\beta}^j$ is a 2^K-dimensional parameter (column) vector and $h(\boldsymbol{\alpha})$ is a 2^K-dimensional covariate (column) vector including all the necessary interaction terms. For instance, in representation (11), $h(\boldsymbol{\alpha})$ is the vector containing 1, α_1, α_2,..., α_K, and their interactions of all orders $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, ... For different diagnostic classification models, we may choose different $h(\boldsymbol{\alpha})$ so that their coefficients correspond directly to a Q-matrix. Examples will be given in the sequel. The likelihood function upon observing $\boldsymbol{\alpha}_i$ for each subject is

$$L(\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^J; \mathbf{R}_i, \boldsymbol{\alpha}_i, i = 1, ..., N) = \prod_{i,j} (c_{j,\boldsymbol{\alpha}_i})^{R_i^j} (1 - c_{j,\boldsymbol{\alpha}_i})^{1-R_i^j} \tag{13}$$

where $c_{j,\boldsymbol{\alpha}}$ is given by (12). Notice that $\boldsymbol{\alpha}_i$ are i.i.d. following distribution $p_{\boldsymbol{\alpha}}$. Then, the observed data likelihood is

$$L(\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^J; \mathbf{R}_i, i = 1, ..., N) = \prod_{i=1}^{N} \sum_{\boldsymbol{\alpha}_i} \Big[p_{\boldsymbol{\alpha}_i} \prod_{j=1}^{J} (c_{j,\boldsymbol{\alpha}_i})^{R_i^j} (1 - c_{j,\boldsymbol{\alpha}_i})^{R_i^j} \Big]. \qquad (14)$$

To simplify the notation, we use L to denote both the observed and the complete data likelihood (with different arguments) when there is no ambiguity. A regularized maximum likelihood estimator of the β-coefficients is given by

$$(\hat{\boldsymbol{\beta}}^1, ..., \hat{\boldsymbol{\beta}}^J) = \arg \max_{\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^J} \log[L(\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^J; \mathbf{R}_i, i = 1, ..., N)] - N \sum_{j=1}^{J} p_{\lambda_j}(\boldsymbol{\beta}^j) \qquad (15)$$

where p_{λ_j} is some penalty function and λ_j is the regularization parameter. In this paper, we choose p_λ to be either the L_1 penalty or the SCAD penalty (Fan and Li, 2001). In particular, to apply the L_1 penalty, we let

$$p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{k=1}^{k} |\beta_k|$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)$; to apply the SCAD penalty, we let

$$p_\lambda(\boldsymbol{\beta}) = \sum_{k=1}^{K} p_\lambda^S(\beta_k).$$

The function $p_\lambda^S(x)$ is defined as $p_\lambda^S(0) = 0$ and

$$\frac{dp_\lambda^S}{dx}(x) = \lambda \left\{ I(x \leq \lambda) + \frac{\max(0, a\lambda - x)}{(a-1)\lambda} \right\}$$

for $x > 0$; for $x < 0$, the function is $p_\lambda^S(x) = p_\lambda^S(-x)$. There is an additional "a" parameter that is chosen to be $a = 3.7$ as suggested by Fan and Li (2001).

On the consistency of the regularized estimator. A natural issue is whether the consistency results developed in the previous section can be applied to the regularized estimator. The consistency results for the regularized estimator can be established by means of the techniques developed in the literature (Yu and Zhao, 2006; Fan and Lv, 2011; Fan and Li, 2001). Therefore, we only provide an outline and omit the details. First of all, the parameter dimension is fixed and the sample size becomes large. The regularization parameter is chosen such that $\lambda_j \to 0$ and $\sqrt{N}\lambda_j \to \infty$ as $N \to \infty$. For the DINA (or DINO) model, let Q_1 and Q_2 be two matrices. If $Q_1 \nsim Q_2$, the consistency results in the previous section ensure that the two families of distributions under different Q's are separated. Thus, with probability tending to one, the true matrix Q is the global maximizer of the profiled likelihood. Since $\lambda_j = o(1)$ and the penalty term is of order $o(N)$, the results in the previous section suggests that the maximized regularized likelihood has to be obtained within ϵ distance from the true value, that is, the consistency results localize the regularized estimator to a small neighborhood of their true values. The oracle properties of the L_1 regularized estimator and SCAD regularized estimator are developed for maximizing the penalized likelihood function locally around the true model parameters (Yu and Zhao, 2006; Fan and Lv, 2011; Fan and Li, 2001). Thus, combining the global results (Q-matrix identifiability) and the local results (oracle condition for the local penalized likelihood maximizer), we obtain that the regularized estimators admit the oracle property in estimating the Q-matrix under the identifiability conditions in the previous section. We mention that for the L_1 regularized estimator irrepresentable condition is needed concerning the Fisher information matrix to ensure the oracle condition (Yu and Zhao, 2006).

For other DCM's, such as NIDA, reduced NC-RUM, and C-RUM, whose representation will be presented immediately, the families of response distributions may be nested among different Q's. Then, the consistency results of the regularized estimator could be developed similarly as those of generalized linear models or generic likelihood functions given that Q is identifiable and the regularization parameter λ_j is chosen carefully such that $\lambda_j \to 0$ and $\sqrt{N}\lambda_j \to \infty$ as $N \to \infty$. Further discussion on the choice of λ_j will be provided later in the discussion section.

3.2 Reparameterization for other diagnostic classification models

We present a few more examples mentioned previously. For each of them, we present the link function g, $h(\boldsymbol{\alpha})$, and the non-zero pattern of the β-coefficients corresponding to each Q-matrix.

DINO model. For the DINO model, we write the positive response probability as

$$\begin{aligned} c_{j,\boldsymbol{\alpha}} &= \operatorname{logit}^{-1}\Big\{\beta_0^j + \sum_{k=1}^{K}\beta_k^j(1-\alpha_k) + \sum_{1\leq k_1<k_2\leq K}\beta_{k_1k_2}^j(1-\alpha_{k_1})(1-\alpha_{k_2}) \\ &\qquad + \sum_{1\leq k_1<k_2<k_3\leq K}\beta_{k_1k_2k_3}^j(1-\alpha_{k_1})(1-\alpha_{k_2})(1-\alpha_{k_3}) \\ &\qquad + \cdots + \beta_{12\ldots K}^j\prod_{k=1}^{K}(1-\alpha_k)\Big\}. \end{aligned}$$

Similar to the DINA model, each row of the Q-matrix, corresponding to one item, maps to two non-zero coefficients. One is the β_0^j and the other one corresponds the interactions of all the required attributes by the Q-matrix.

NIDA model. The positive response probability can be written as

$$\log c_{j,\boldsymbol{\alpha}} = \beta_0^j + \sum_{k=1}^{K}\beta_k^j\alpha_k.$$

Then, the corresponding Q-matrix entries are given by $q_{jk} = I(\beta_k^j \neq 0)$. Unlike the DINA and the DINO model, the number of non-zero coefficients for each item is unknown.

Reduced NC-RUM. This model is very similar to the NIDA model. The positive probability can written as

$$\log c_{j,\boldsymbol{\alpha}} = \beta_0^j + \sum_{k=1}^{K} \beta_k^j (1 - \alpha_k)$$

and $q_{jk} = \mathbf{1}(\beta_k^j \neq 0)$.

C-RUM. The positive probability can written as

$$\text{logit}(c_{j,\boldsymbol{\alpha}}) = \beta_0^j + \sum_{k=1}^{K} \beta_k^j \alpha_k$$

and $q_{jk} = \mathbf{1}(\beta_k^j \neq 0)$.

As a summary, all the diagnostic classification models in the literature admit the generalized linear form as in (12). Furthermore, each Q-matrix corresponds a non-zero pattern of the regression coefficients and the regularized estimator has a wide applicability.

3.3 Computation via EM algorithm

The advantage of the regularized maximum likelihood estimation for the Q-matrix lies in computation. As mentioned previously, the computation of $\hat{Q}_{MLE}$ in (9) requires evaluation of the profiled likelihood for all possible Q-matrices and there are $2^{J \times K}$ such matrices. This is computationally impossible even for some practically small J and K. The computation of (15) can be done by combining the expectation-maximization (EM) algorithm and the coordinate descent algorithm. In particular, we view $\boldsymbol{\alpha}$ as the missing data following the prior distribution $p_{\boldsymbol{\alpha}}$. The EM algorithm consists of two steps. The E-step computes function

$$\begin{aligned}&H(\boldsymbol{\beta}^1_*, ..., \boldsymbol{\beta}^J_* | \boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^J, p_{\boldsymbol{\alpha}})\\ = \quad & E[\log L(\boldsymbol{\beta}^1_*, ..., \boldsymbol{\beta}^J_*; \mathbf{R}_i, \boldsymbol{\alpha}_i, i = 1, ..., N) | \mathbf{R}_i, i = 1, ..., N, \boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^J, p_{\boldsymbol{\alpha}}]\end{aligned}$$

where the above expectation is taken with respect to $\boldsymbol{\alpha}_i$, $i = 1, ..., N$, under the posterior distribution $P(\ \cdot\ |\mathbf{R}_i, i = 1, ..., N, \boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^J, p_{\boldsymbol{\alpha}})$. The E-step is a closed form computation. First, the complete data log-likelihood function is additive

$$\log L(\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^J; \mathbf{R}_i, \boldsymbol{\alpha}_i, i = 1, ..., N) = \sum_{i=1}^{N}\sum_{j=1}^{J}[R_i^j \log c_{j,\boldsymbol{\alpha}_i} + (1 - R_i^j)\log(1 - c_{j,\boldsymbol{\alpha}_i})].$$

Furthermore, under the posterior distribution $\boldsymbol{\alpha}_1$,..., $\boldsymbol{\alpha}_N$ are jointly independent. Therefore, one only needs to evaluate

$$E[R_i^j \log c_{j,\boldsymbol{\alpha}_i} + (1 - R_i^j)\log(1 - c_{j,\boldsymbol{\alpha}_i}) | \mathbf{R}_i, i = 1, ..., N, \boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^J, p_{\boldsymbol{\alpha}}]$$

for each $i = 1, ..., N$ and $j = 1, ..., J$. Notice that $\boldsymbol{\alpha}$ is a discrete random variable taking values in $\{0, 1\}^K$. Therefore, the posterior distribution of each $\boldsymbol{\alpha}_i$ can be computed exactly and the complexity of the above conditional expectation is 2^K that is manageable for K as large as 10 that is a very high dimension for diagnostic classification models in practice. Therefore the overall computational complexity of the E-step is $O(NJ2^K)$.

The M-step consists of maximizing the H-function with the penalty term

$$\max_{\boldsymbol{\beta}^1_*, ..., \boldsymbol{\beta}^J_*} H(\boldsymbol{\beta}^1_*, ..., \boldsymbol{\beta}^J_* | \boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^J, p_{\boldsymbol{\alpha}}) - N\sum_{j=1}^{J} p_{\lambda_j}(\boldsymbol{\beta}^j_*).$$

Before applying the coordinate descent algorithm, we further reduce the dimension.

The objective function can be written as

$$\sum_{j=1}^{J}\Big\{\sum_{i=1}^{N}E[R_i^j\log c_{j,\boldsymbol{\alpha}_i}+(1-R_i^j)\log(1-c_{j,\boldsymbol{\alpha}_i})|\mathbf{R}_i,i=1,...,N,\boldsymbol{\beta}^1,...,\boldsymbol{\beta}^J,p_{\boldsymbol{\alpha}}]-p_{\lambda_j}(\boldsymbol{\beta}_*^j)\Big\}.$$

For each j, the term

$$\sum_{i=1}^{N}E[R_i^j\log c_{j,\boldsymbol{\alpha}_i}+(1-R_i^j)\log(1-c_{j,\boldsymbol{\alpha}_i})|\mathbf{R}_i,i=1,...,N,\boldsymbol{\beta}^1,...,\boldsymbol{\beta}^J,p_{\boldsymbol{\alpha}}]-p_{\lambda_j}(\boldsymbol{\beta}_*^j)$$

consists only of $\boldsymbol{\beta}_*^j$. Thus, the M-step can be done by maximizing each $\boldsymbol{\beta}_*^j$ independently. Each $\boldsymbol{\beta}_*^j$ has 2^K coordinate and we apply the coordinate descent algorithm (developed for generalized linear models) to maximize the above function for each j. For details about this algorithm, see Friedman et al. (2010). Furthermore, $p_{\boldsymbol{\alpha}}$ is updated by $\sum_{i=1}^{N}P(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}|\mathbf{R},\boldsymbol{\beta}^1,...,\boldsymbol{\beta}^J,p_{\boldsymbol{\alpha}})/N$.

The EM algorithm guarantees a monotone increasing objective function. However, there is no guarantee that the algorithm converges to the global maximum. We empirically found that the algorithm sometimes does stop at a local maximum, especially when λ is large. Therefore, we suggest applying the algorithm with different starting points and select the best.

3.4 Further discussions

It is suggested by the theories that the regularization parameter λ be chosen such that $\lambda\to 0$ and $\sqrt{N}\lambda\to\infty$ that is a wide range. For specific diagnostic classification models, we may have more specific choices of λ. For the DINA and the DINO model, each row of the Q-matrix, corresponding to the attribute requirement of one item, maps to two non-zero coefficients. Therefore, we may choose λ_j for each item differently such that the resulted coefficients $\boldsymbol{\beta}^j$ has exactly two non-zero elements.

The NIDA, NC-RUM, and C-RUM models do not admit a fixed number of coefficients for each item. To simplify the problem, instead of using item-specific regularization parameters, we choose a single regularization parameter for all items. Furthermore, we build a solution path for different λ. Thus, instead of providing one estimate of the Q-matrix, a set of estimated Q-matrices corresponding to different λ is obtained. We may further investigate these matrices for further validation based on our knowledge of the item-attribute relationship. In case one does not have enough knowledge, one may choose λ via standard information criteria. For instance, we may choose λ such that the resulted selection of latent variables admits the smallest BIC.

4 Simulation study

In this section, simulation studies are conducted to illustrate the performance of the proposed method. The DINO model is mathematically equivalent to the DINA model (Proposition 1) and thus we only provide results for the DINA model. The data from the DINA model are generated under different settings and then the estimated Q-matrix and the true Q-matrix are compared. Two simulation studies are conducted when the attributes $\alpha_1, ..., \alpha_K$ are independent and dependent. The results are presented assuming all the model parameters are unknown including the Q-matrix, attribute distribution, slipping and guessing parameters.

4.1 Study 1: independent attributes

Attribute profiles are generated from the uniform distribution

$$p_{\boldsymbol{\alpha}} = 2^{-K}.$$

We consider the cases that $K = 3$ and 4 and $J = 18$ items. The following Q-matrices are adopted

	$N = 500$		$N = 1000$		$N = 2000$		$N = 4000$	
	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$
$K = 3$	38	62	81	19	98	2	100	0
$K = 4$	20	80	48	52	77	23	99	1

Table 1: Numbers of correctly estimated Q-matrices among 100 simulations with sample size 500, 1000, 2000, and 4000 for the L_1 penalty.

	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
$K = 3$	98.1%	99.6%	100.0%	100.0%
$K = 4$	97.7%	98.9%	99.6%	100.0%

Table 2: Proportion of entries correctly specified by $\hat{Q}$ for the L_1 regularized estimator averaging over all independent replications.

$$Q_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad Q_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

These two matrices are chosen such that the identifiability conditions are satisfied. The slipping and guessing parameters are set to be 0.2, but treated as unknown when estimating Q. All other conditions are also satisfied. For each Q, we consider sample sizes $N = 500$, 1000, 2000, and 4000. For each particular Q and N, 100 independent data sets are generated to evaluate the performance.

	Q_1					Q_2			
Sample size	500	1000	2000	4000		500	1000	2000	4000
$\hat{Q}_{1:15} = Q_{1:15}$	100	100	100	100	$\hat{Q}_{1:14} = Q_{1:14}$	98	100	100	100
$\hat{Q}_{1:15} \neq Q_{1:15}$	0	0	0	0	$\hat{Q}_{1:14} \neq Q_{1:14}$	2	0	0	0
$\hat{Q}_{16:18} = Q_{16:18}$	38	81	98	100	$\hat{Q}_{15:18} = Q_{15:18}$	20	48	77	99
$\hat{Q}_{16:18} \neq Q_{16:18}$	62	19	2	0	$\hat{Q}_{15:18} \neq Q_{15:18}$	80	52	23	1

Table 3: Numbers of correctly estimated $Q_{1:15}$ and $Q_{16:18}$ for Q_1 and numbers of correctly estimated $Q_{1:14}$ and $Q_{15:18}$ for Q_2 among 100 simulations with solutions for the L_1 regularized estimator

L_1 regularized estimator. The simulation results of the L_1 regularized estimator are summarized in Tables 1, 2, and 3. According to Table 1, for both $K = 3$ and 4, our method estimates the Q-matrix almost without error when the sample size is as large as 4000. In addition, the higher the dimension is the more difficult the problem is. Furthermore, for the cases when the estimator misses the Q-matrix, $\hat{Q}$ differs from the true by only one or two rows. We look closer into the estimators in Table 2 that reports the proportion of entries correctly specified by $\hat{Q}$

$$CR(\hat{Q}) = \max_{Q' \sim Q} \left\{ \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \mathbf{1}_{\{\hat{q}_{j,k} = q'_{j,k}\}} \right\}.$$

We empirically found that the row vectors of Q_1 and Q_2 that require three attributes or four attributes (rows 15 to 18 in Q_1 and rows 16 to 18 in Q_2) are much more difficult to estimate than others. This phenomenon is reflected by Table 3, in which the notation $Q_{I_1:I_2}$ represents the submatrix of Q containing row I_1 to row I_2. In fact, for all simulations in this study, most misspecifications are due to the misspecification of the submatrices of Q_1 and Q_2 that the corresponding items require three attributes or more.

SCAD estimator. Under the same setting, we investigate the SCAD estimator. The results are summarized in Tables 4 and 5. The SCAD estimator performs better than the L_1 regularized estimator upon comparing Table 1 and Table 4.

	$N = 500$		$N = 1000$		$N = 2000$		$N = 4000$	
	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$
$K = 3$	98	2	100	0	100	0	100	0
$K = 4$	30	70	96	4	100	0	100	0

Table 4: Numbers of correctly estimated Q-matrices among 100 simulations with sample size 500, 1000, 2000, and 4000 for the SCAD estimator.

	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
$K = 3$	99.9%	100%	100.0%	100.0%
$K = 4$	97.6%	99.9%	100.0%	100.0%

Table 5: Proportion of entries correctly specified by $\hat{Q}$ ($CR(\hat{Q})$) for the SCAD estimator averaging over all independent replications.

4.2 Study 2: dependent attributes

For each subject, we generate $\theta = (\theta_1, \cdots, \theta_K)$ that is a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where the covariance matrix Σ has unit variance and has a common correlation ρ, that is,

$$\Sigma = (1 - \rho)I_K + \rho \mathbf{1}\mathbf{1}^\top$$

where $\mathbf{1}$ is the vector of ones and I_K is the K by K identity matrix. We consider the situations that $\rho = 0.05$, 0.15 and 0.25. Then the attribute profile $\boldsymbol{\alpha}$ is given by

$$\alpha_k = \begin{cases} 1 & \text{if } \theta_k \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

We consider $K = 3$ and Q_1 be the Q-matrix. Table 6 shows the probability distribution $p_{\boldsymbol{\alpha}}$. The slipping and the guessing parameters remain 0.2. The rest of the setting is the same as that of Study 1.

L_1 **regularized estimator.** The simulation results of the L_1 regularized estimator are summarized in Tables 7 and 8. Based on Table 7, the estimation accuracy is improved when the sample size increases. We also observe that the proposed algorithm performs better when ρ increases. A heuristic interpretation is as follows. The row

Class	(0,0,0)	(1,0,0)	(0,1,0)	(1,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,1)
$\rho = 0.05$	0.137	0.121	0.121	0.121	0.121	0.121	0.121	0.137
$\rho = 0.15$	0.161	0.113	0.113	0.113	0.113	0.113	0.113	0.161
$\rho = 0.25$	0.185	0.105	0.105	0.105	0.105	0.105	0.105	0.185

Table 6: The distribution of the latent attributes of the three-dimensional DINA model for $\rho = 0.05, 0.15$ and 0.25

	$N = 500$		$N = 1000$		$N = 2000$		$N = 4000$	
	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$
$\rho = 0.05$	54	46	87	13	99	1	100	0
$\rho = 0.15$	67	33	93	7	100	0	100	0
$\rho = 0.25$	76	24	95	5	100	0	100	0

Table 7: Numbers of correctly estimated Q-matrices among 100 simulations for sample sizes 500, 1000, 2000, and 4000 for the L_1 regularized estimator.

vector of Q tends to be more difficult to estimate when the numbers of subjects who are capable and who are not capable to answer are not balanced. The row vector $(1, 1, 1)$ is the most difficult to estimate because only subjects with attribute profile $(1, 1, 1)$ are able to solve them and all other subjects are not. According to Table 6, as ρ increases, the proportion of subjects with attribute profile $(1, 1, 1)$ increases, which explains the improvement of the performance. In fact, similar to the situation that α_i's are independent, for most simulations in which the $\hat{Q}$ misses the true, $\hat{Q}$ differs from the true at the row vectors whose true value is $(1, 1, 1)$.

SCAD estimator. Under the same simulation setting, the results of the SCAD estimator are summarized in Tables 9 and 10. Its performance is empirically better than that of the L_1 regularized estimator. When the sample size is as small as 500, it has a very high probability estimating all the entries of the Q-matrix correctly.

	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
$\rho = 0.05$	98.5%	99.7%	100.0%	100.0%
$\rho = 0.15$	99.2%	99.8%	100.0%	100.0%
$\rho = 0.25$	99.4%	99.9%	100.0%	100.0%

Table 8: Proportion of entries correctly specified by $\hat{Q}$ for the L_1 regularized estimator averaging over all independent replications.

	$N = 500$		$N = 1000$		$N = 2000$		$N = 4000$	
	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$
$\rho = 0.05$	97	3	100	0	100	0	100	0
$\rho = 0.15$	98	2	100	0	100	0	100	0
$\rho = 0.25$	99	1	100	0	100	0	100	0

Table 9: Numbers of correctly estimated Q-matrices among 100 simulations for sample sizes 500, 1000, 2000, and 4000 under the SCAD penalty

	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
$\rho = 0.05$	99.7%	100.0%	100.0%	100.0%
$\rho = 0.15$	100.0%	100.0%	100.0%	100.0%
$\rho = 0.25$	100.0%	100.0%	100.0%	100.0%

Table 10: Proportion of entries correctly specified by $\hat{Q}$ ($CR(\hat{Q})$) under the SCAD penalty averaging over all independent replications.

Remark 4 *Once an estimate of the Q-matrix has been obtained, other model parameters such as the slipping and the guessing parameters and the attribute population can be estimated via the maximum likelihood estimator* (10). *Simulation studies show that these parameters can be estimated accurately given that the Q-matrix is recovered with a high chance. As the main focus of this paper is on the Q-matrix, we do not report detailed simulation results for these parameters.*

5 Real data analysis

5.1 Example 1: fraction subtraction data

The data set contains 536 middle school students' responses to 17 fraction subtraction problems. The responses are binary: correct or incorrect solution to the problem. The data were originally described by Tatsuoka (1990) and later by Tatsuoka (2002); de la Torre and Douglas (2004) and many other studies of diagnostic classification models. In these works, the DINA model is fitted with a Q-matrix pre-specified. We fit the DINA model to the data and estimate the Q-matrix for $K = 3$ and 4. Then we validate the estimated Q-matrix by our knowledge of the cognitive processes of

		$K = 3$					$K = 4$					
ID	Content	$\hat{Q}$			$\hat{s}$	$\hat{g}$	$\hat{Q}$				$\hat{s}$	$\hat{g}$
1	$\frac{5}{3} - \frac{3}{4}$	1	0	0	0.12	0.03	1	0	0	0	0.12	0.03
2	$\frac{3}{4} - \frac{3}{8}$	1	0	0	0.05	0.04	1	0	0	0	0.04	0.03
3	$\frac{5}{6} - \frac{1}{9}$	1	0	0	0.13	0.00	1	0	0	0	0.12	0.00
4	$3\frac{1}{2} - 2\frac{3}{2}$	0	1	0	0.13	0.17	0	1	1	0	0.13	0.21
5	$1\frac{1}{8} - \frac{1}{8}$	0	0	1	0.07	0.28	0	0	1	0	0.07	0.24
6	$3\frac{4}{5} - 3\frac{2}{5}$	0	0	1	0.04	0.20	0	0	1	0	0.04	0.13
7	$4\frac{5}{7} - 1\frac{4}{7}$	0	0	1	0.08	0.20	0	0	1	1	0.05	0.27
8	$4\frac{3}{5} - 3\frac{4}{10}$	1	0	1	0.18	0.31	1	0	1	0	0.19	0.31
9	$3 - 2\frac{1}{5}$	1	0	1	0.32	0.06	0	1	1	1	0.23	0.11
10	$2 - \frac{1}{3}$	1	0	1	0.23	0.07	0	1	1	1	0.15	0.14
11	$4\frac{4}{12} - 2\frac{7}{12}$	0	1	1	0.23	0.03	0	1	1	0	0.24	0.03
12	$4\frac{1}{3} - 2\frac{4}{3}$	0	1	1	0.07	0.07	0	1	0	0	0.09	0.06
13	$7\frac{3}{5} - \frac{4}{5}$	0	1	1	0.13	0.05	0	1	1	0	0.15	0.04
14	$4\frac{1}{10} - 2\frac{8}{10}$	0	1	1	0.15	0.13	0	1	1	0	0.16	0.12
15	$4 - 1\frac{4}{3}$	0	1	1	0.37	0.02	0	1	0	1	0.32	0.02
16	$4\frac{1}{3} - 1\frac{5}{3}$	0	1	1	0.18	0.01	0	1	1	0	0.20	0.01
17	$3\frac{3}{8} - 2\frac{5}{6}$	1	1	1	0.33	0.01	1	1	1	1	0.31	0.01

Table 11: The estimated Q-matrix based on L_1 regularization and the corresponding slipping and guessing parameters for the three and four dimensional DINA model for the fraction subtraction data

problem solving.

Table 11 presents the estimated Q-matrix along with the slipping and the guessing parameters for $K = 3$ based on L_1 regularization. The slipping and the guessing parameter are estimated by (10). According to our knowledge of the cognitive processes, the items are clustered according to $\hat{Q}$ reasonably. Roughly speaking, the three attributes can be interpreted as "finding common denominator", "writing integer as fraction", and "subtraction of two fraction numbers when there are integers involved" respectively.

We further fit a four dimensional DINA model and the results are also summarized in Table 11. The first attribute can be interpreted as "finding common denominator", the second as "borrowing from the whole number part", and the third and fourth attributes can be interpreted as "subtraction of two fraction numbers when there are integers involved". However, it seems difficult to interpret the third and the fourth

ID	Content	A1	A2	A3	A4	A5	A6	A7	A8
1	$\frac{5}{3} - \frac{3}{4}$	0	0	0	1	0	1	1	0
2	$\frac{3}{4} - \frac{3}{8}$	0	0	0	1	0	0	1	0
3	$\frac{5}{6} - \frac{1}{9}$	0	0	0	1	0	0	1	0
4	$3\frac{1}{2} - 2\frac{3}{2}$	0	1	1	0	1	0	1	0
5	$1\frac{1}{8} - \frac{1}{8}$	0	0	0	0	0	0	1	1
6	$3\frac{4}{5} - 3\frac{2}{5}$	0	1	0	0	0	0	1	0
7	$4\frac{5}{7} - 1\frac{4}{7}$	0	1	0	0	1	1	1	0
8	$4\frac{3}{5} - 3\frac{4}{10}$	0	1	0	1	0	0	1	1
9	$3 - 2\frac{1}{5}$	1	1	0	0	0	0	1	0
10	$2 - \frac{1}{3}$	0	1	0	0	1	0	1	0
11	$4\frac{4}{12} - 2\frac{7}{12}$	0	1	0	0	1	0	1	1
12	$4\frac{1}{3} - 2\frac{4}{3}$	0	1	0	0	1	0	1	0
13	$7\frac{3}{5} - \frac{4}{5}$	0	1	0	0	1	0	1	0
14	$4\frac{1}{10} - 2\frac{8}{10}$	0	1	0	0	1	1	1	0
15	$4 - 1\frac{4}{3}$	1	1	1	0	1	0	1	0
16	$4\frac{1}{3} - 1\frac{5}{3}$	0	1	1	0	1	0	1	0
17	$3\frac{3}{8} - 2\frac{5}{6}$	0	1	0	1	1	0	1	0

Table 12: The Q-matrix specified in de la Torre and Douglas (2004)

attributes separately.

This data set has been studied intensively. A Q-matrix (with a little bit variation from study to study) is also prespecified based on understandings of each test problem. Table 12 presents the Q-matrix used in de la Torre and Douglas (2004) that contains eight attributes ($K = 8$). Each attribute corresponds one type of manipulation of fractions:

A1: Convert a whole number to a fraction

A2: Separate a whole number from a fraction

A3: Simplify before subtracting

A4: Find a common denominator

A5: Borrow from whole number part

A6: Column borrow to subtract the second numerator from the first

A7: Subtract numerators

A8: Reduce answers to simplest form

We believe that K is overspecified given that there are only 17 items. Nevertheless, we are able to find some approximate matching between this prespecified matrix and ours. Attributes one in Table 11 roughly corresponds to attribute four in Table 12, attribute two in Table 11 to attribute five in Table 12, and attribute three in Table 11 to attribute two in Table 12.

We also estimate the Q-matrix using SCAD. The estimated Q-matrices are given in Tables 13 for $K = 3$ and 4. The estimates are not as interpretable as those by the L_1 penalty, although SCAD has better performance in the simulation study. We believe that this is mostly due to the lack of fit of the DINA model. This is an illustration of the difficulties in the analysis of cognitive diagnosis. Most models impose restrictive parametric assumptions such that the lack of fit may affect the quality of the inferences. Thus, the performance in simulations does not extrapolate to real data analysis. We also emphasize that the estimated Q-matrix only serves as a guide of the item-attribute association and strongly recommend that researchers verify or even modify the estimates based on their understanding of the items.

5.2 Example 2: Social anxiety disorder data

The social anxiety disorder data is a subset of the National Epidemiological Survey on Alcohol and Related Conditions (NESARC) (Grant et al., 2003). We consider participants' binary responses (Yes/No) to thirteen diagnostic questions for social anxiety disorder. The questions are designed by the Diagnostic and Statistical Manual of Mental Disorders, 4th ed and are displayed in Table 14. Incomplete cases are removed from the data set. The sample size is 5226. To understand the latent

		$K=3$					$K=4$					
ID	Content	$\hat{Q}$			$\hat{s}$	$\hat{g}$	$\hat{Q}$				$\hat{s}$	$\hat{g}$
1	$\frac{5}{3}-\frac{3}{4}$	1	0	0	0.13	0.03	1	1	1	1	0.12	0.24
2	$\frac{3}{4}-\frac{3}{8}$	1	0	0	0.06	0.04	1	0	0	0	0.05	0.04
3	$\frac{5}{6}-\frac{1}{9}$	1	0	0	0.14	0.00	1	0	0	0	0.14	0.01
4	$3\frac{1}{2}-2\frac{3}{2}$	1	1	1	0.13	0.26	1	1	1	1	0.14	0.25
5	$1\frac{1}{8}-\frac{1}{8}$	0	0	1	0.05	0.52	1	1	1	1	0.04	0.54
6	$3\frac{4}{5}-3\frac{2}{5}$	0	0	1	0.04	0.49	0	1	1	0	0.03	0.50
7	$4\frac{5}{7}-1\frac{4}{7}$	1	1	1	0.05	0.50	1	1	0	1	0.06	0.49
8	$4\frac{3}{5}-3\frac{4}{10}$	1	1	1	0.17	0.39	1	1	1	1	0.17	0.38
9	$3-2\frac{1}{5}$	1	1	1	0.24	0.12	1	1	1	1	0.25	0.11
10	$2-\frac{1}{3}$	1	1	1	0.16	0.14	1	1	1	1	0.17	0.14
11	$4\frac{4}{12}-2\frac{7}{12}$	0	0	1	0.25	0.03	0	1	1	1	0.23	0.03
12	$4\frac{1}{3}-2\frac{4}{3}$	0	0	1	0.10	0.07	0	1	0	0	0.12	0.07
13	$7\frac{3}{5}-\frac{4}{5}$	0	0	1	0.16	0.04	0	1	0	0	0.17	0.04
14	$4\frac{1}{10}-2\frac{8}{10}$	0	0	1	0.16	0.11	0	1	0	1	0.15	0.10
15	$4-1\frac{4}{3}$	1	1	1	0.32	0.02	1	1	1	1	0.34	0.02
16	$4\frac{1}{3}-1\frac{5}{3}$	0	0	1	0.21	0.01	0	1	0	0	0.22	0.01
17	$3\frac{3}{8}-2\frac{5}{6}$	1	1	1	0.33	0.00	1	1	1	1	0.35	0.00

Table 13: The estimated Q-matrix based on SCAD regularization and the corresponding slipping and guessing parameters for the three and four dimensional DINA model for the fraction subtraction data

structure of social phobia, we fit the compensatory DINO model for $K = 2$, 3, and 4.

We first consider the L_1 penalty and fit the two-dimensional DINO model. The estimates $\hat{Q}$, $\hat{\mathbf{s}}$, and $\hat{\mathbf{g}}$ are summarized as Case $K = 2$ of Table 15. In addition, the correlation between the two attributes is 0.47. We further explore the latent structure by considering the three-dimensional DINO model. For the result, $\hat{Q}$, $\hat{\mathbf{s}}$, and $\hat{\mathbf{g}}$ are summarized as Case $K = 3$ of Table 15. A similar latent structure as $\hat{Q}$ in Case $K = 3$ of Table 15 is considered in an item response theory model based confirmatory factor analysis (Iza et al., 2014), where the item-attribute structure is prespecified. In their study, the three (continuous) factors are interpreted as “public performance”, “close scrutiny”, and “interaction”, which correspond roughly to those in Case $K = 3$ of Table 15. Finally, the four-dimensional DINO model is considered. The results

ID	Have you ever had a strong fear or avoidance of
1	speaking in front of other people?
2	taking part/ speaking in class?
3	taking part/ speaking at a meeting?
4	performing in front of other people?
5	being interviewed?
6	writing when someone watches?
7	taking an important exam?
8	speaking to an authority figure?
9	eating/drinking in front of other people?
10	having conversations with people you don't know well?
11	going to parties/social gatherings?
12	dating?
13	being in a small group situation?

Table 14: The content of 13 items for the social anxiety disorder data.

are summarized as Case $K = 4$ in Table 15. According to the corresponding $\hat{Q}$, the third group (items 9 - 13) based on three-dimensional model splits into two attributes. Furthermore, item 6 "writing when someone watches" becomes associated with attribute three. Furthermore, we estimate the Q-matrix via SCAD. The estimates are summarized in Tables 16 that are similar to those of the L_1 penalty.

We observe that the estimated slipping parameters are relatively large for some items (such as items 6, 9, 12 and 13) and their guessing parameters are small. These are the low prevalence items that are unlikely to be present even among the abnormal populations. On the other hand, if someone responds positively to that item, he/she is very likely to possess the corresponding disorder (attribute). This is reflected by the small guessing parameters.

6 Concluding remarks

This paper considers the estimation of Q-matrix that is a key quantity in the specification of diagnostic classification models. The results are two-fold. First, we present theoretical identifiability results of the Q-matrix for two stylized diagnostic classifica-

	$K=2$				$K=3$					$K=4$					
ID	$\hat{Q}$		$\hat{s}$	$\hat{g}$	$\hat{Q}$			$\hat{s}$	$\hat{g}$	$\hat{Q}$				$\hat{s}$	$\hat{g}$
1	1	0	0.05	0.54	1	0	0	0.05	0.49	1	0	0	0	0.05	0.49
2	1	0	0.09	0.27	1	0	0	0.11	0.21	1	0	0	0	0.11	0.21
3	1	0	0.13	0.15	1	0	0	0.16	0.09	1	0	0	0	0.16	0.09
4	1	0	0.12	0.30	1	0	0	0.15	0.25	1	0	0	0	0.14	0.25
5	1	0	0.46	0.07	0	1	0	0.29	0.09	0	1	0	0	0.29	0.08
6	0	1	0.66	0.07	0	1	0	0.68	0.06	0	0	1	0	0.56	0.08
7	1	0	0.42	0.22	0	1	0	0.26	0.21	0	1	0	0	0.27	0.20
8	0	1	0.34	0.16	0	1	0	0.30	0.09	0	1	0	0	0.30	0.08
9	0	1	0.68	0.02	0	0	1	0.68	0.02	0	0	1	0	0.58	0.02
10	0	1	0.13	0.21	0	0	1	0.13	0.20	0	0	1	1	0.14	0.16
11	0	1	0.20	0.12	0	0	1	0.17	0.10	0	0	0	1	0.21	0.08
12	0	1	0.59	0.05	0	0	1	0.59	0.05	0	0	1	0	0.47	0.06
13	0	1	0.67	0.01	0	0	1	0.68	0.01	0	0	1	0	0.57	0.01

Table 15: The estimated Q-matrix based on L_1 regularization and the slipping and guessing parameters for the two, three and four dimensional DINO model for the social anxiety disorder data.

tion models, the DINA model and the DINO model. A set of sufficient conditions is provided, under which it is theoretically possible to reconstruct the matrix based on only the dependence of the response patterns. The development of the theory is by means of the maximum likelihood estimation (MLE). Unfortunately, MLE, though consistent (under conditions), is not practically implementable due to the unaffordable computational overhead. Thus, the second objective is to present a computationally feasible estimator for Q. We formulate the Q-matrix estimation to a latent variable selection problem and employ the regularized maximum likelihood as the main tool. The L_1 penalty and the SCAD penalty are considered. For the optimization, we combine the expectation-maximization algorithm and the coordinate decent algorithm. Both are well studied numerical methods for optimization. The estimation procedure is applicable to most diagnostic classification models and is not limited to the DINA or the DINO model.

The performances of the two penalty functions are compared via simulation studies, in which SCAD penalty yields better results. However, in the analysis of the

	$K = 2$				$K = 3$					$K = 4$					
ID	$\hat{\mathbf{Q}}$		$\hat{s}$	$\hat{g}$	$\hat{Q}$			$\hat{s}$	$\hat{g}$	$\hat{Q}$				$\hat{s}$	$\hat{g}$
1	1	0	0.05	0.54	1	0	0	0.05	0.49	1	0	0	0	0.05	0.49
2	1	0	0.09	0.27	1	0	0	0.11	0.21	1	0	0	0	0.11	0.21
3	1	0	0.13	0.14	1	0	0	0.16	0.09	1	0	0	0	0.16	0.09
4	1	0	0.12	0.30	1	0	0	0.15	0.25	1	0	0	0	0.15	0.25
5	1	0	0.46	0.07	0	1	0	0.29	0.09	0	1	0	0	0.30	0.08
6	0	1	0.65	0.07	0	1	0	0.68	0.06	0	0	1	0	0.55	0.07
7	1	1	0.43	0.20	0	1	0	0.26	0.21	0	1	0	0	0.27	0.20
8	0	1	0.33	0.16	0	1	0	0.30	0.09	0	1	0	0	0.31	0.08
9	0	1	0.68	0.02	0	0	1	0.68	0.02	0	0	0	1	0.68	0.02
10	0	1	0.13	0.21	0	0	1	0.13	0.20	0	0	0	1	0.11	0.19
11	0	1	0.20	0.13	0	0	1	0.17	0.10	0	0	0	1	0.16	0.09
12	0	1	0.59	0.05	0	0	1	0.59	0.05	0	0	1	0	0.44	0.05
13	0	1	0.67	0.01	0	0	1	0.68	0.01	0	0	1	0	0.55	0.01

Table 16: The estimated Q-matrix based on SCAD regularization and the slipping and guessing parameters for the two, three and four dimensional DINO model for the social anxiety disorder data.

fraction subtraction data, SCAD yields results that are difficult to interpret, while the L_1 penalty produces more interpretable Q-matrices. We believe that this is mostly due to the lack of fit of the DINA model. This data set in part illustrates the complications in the real data analysis for diagnostic classification models. Although the theory and estimation procedure do not require a prior knowledge of Q, we strongly recommend researchers should try to combine their knowledge in the subject matter and our inference tools. That is, our estimated Q-matrix serves as a guideline for the item-attribute association. Further refinement (such as choosing the regularization parameter or the penalty function) should rely on understanding of the items.

Throughout the discussion, the number of attributes (K) is assumed to be known. A natural extension is to estimate K simultaneously with other parameters. This can be done by introducing an additional penalty function added to the likelihood function. We leave this topic for future study.

References

Casella, G. and Berger, R. L. (2001). *Statistical Inference*. Duxbury Press, Belmont, CA.

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

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45:343–362.

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

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

de la Torre, J. and Douglas, J. (2004). Higher order latent trait models for cognitive diagnosis. *Psychometrika*, 69:333–353.

DiBello, L., Stout, W., and Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In Nichols, P. D., Chipman, S. F., and Brennan, R. L., editors, *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

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

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57:5467–5484.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

Grant, B. F., Kaplan, K., Shepard, J., and Moore, T. (2003). Source and accuracy statement for wave 1 of the 2001–2002 national epidemiologic survey on alcohol and related conditions. *Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism*.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74:191–210.

Iza, M., Wall, M. M., Heimberg, R. G., Rodebaugh, T. L., Schneier, F. R., Liu, S.-M., and Blanco, C. (2014). Latent structure of social fears and social anxiety disorders. *Psychological Medicine*, 44:361–370.

Junker, B. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment. *Technical Report, available from http://www.stat.cmu.edu/ ∼ brian/nrc/cfa/documents/final.pdf.*

Junker, B. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25:258–272.

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on tatsuoka's rule-space approach. *Journal of Educational Measurement*, 41:205–237.

Liu, J., Xu, G., and Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli*, 19:1790–1817.

Liu, Y., Douglas, J., and Henson, R. (2007). Testing person fit in cognitive diagnosis. In *the annual meeting of the National Council on Measurement in Education (NCME)*, Chicago, IL, April.

Rupp, A. and Templin, J. (2008a). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement*, 68:78–98.

Rupp, A. and Templin, J. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspective*, 6:219–262.

Rupp, A., Templin, J., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications.* Guilford Press, New York, NY.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Applied Statistics (JRSS-C)*, 51:337–350.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, 12:55–73.

Tatsuoka, K. (2009). *Cognitive assessment: an introduction to the rule space method.* Routledge, New York, NY.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In Frederiksen, N., Glaser, R., Lesgold, A., and Shafto, M., editors, *Diagnostic monitoring of skill and knowledge acquisition.* Hillsdale, NJ: Erlbaum.

Templin, J. and Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11:287–305.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in medicine*, 16:385–395.

von Davier, M. (2005). *A General Diagnostic Model Applied to Language Testing Data*. Educational Testing Service Research Report No. RR-05-16, Princeton, NJ.

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

von Davier, M. and Yamamoto, K. (2004). A class of models for cognitive diagnosis. In *4th Spearman Conference*, Philadelphia, PA.

Yu, B. and Zhao, P. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

A The duality between the DINA and the DINO model and some technical constructions

We establish the duality between the DINA and the DINO model.

Proposition 1 *Consider a response vector* $\mathbf{R} = (R^1, ..., R^J)$ *following a DINA model with latent attribute* $\boldsymbol{\alpha}$ *and* $\mathbf{R}' = (R'^1, ..., R'^J)$ *following the DINO model with latent attribute* $\boldsymbol{\alpha}'$. *Their slipping and guessing parameters are denoted by* s_j, g_j, s'_j, *and* g'_j, *respectively. If* $1 - s_j = g'_j$, $g_j = 1 - s'_j$, *and* $\alpha_j = 1 - \alpha'_j$, *then* $\mathbf{R}$ *and* $\mathbf{R}'$ *are identically distributed.*

The above proposition is straightforward to verify through the ideal response indicators in (2) and (5). Thus, we omit the detailed proof. The above proposition suggests that the DINA and the DINO model are mathematically the same but with different parameterizations. Therefore, all the theoretical results we developed for the DINA model can be directly translated to the DINO model based on the above proposition. Therefore, the rest of the technical proofs are all for the DINA model. In the rest of this subsection, we present some technical construction for the subsequent proof.

T-matrix for the DINA model. For notational convenience, we will write

$$c = 1 - s$$

that is the correct response probability for capable students ("c" for correct). Then,

$$\mathbf{c} = \mathbf{1} - \mathbf{s}$$

is the corresponding parameter vector.

The T-matrix serves as a connection between the observed response distribution and the model structure. We first specify each row vector of the T-matrix for a general conjunctive diagnostic model.

For each item j, we have

$$P(R^j = 1 | Q, \mathbf{p}, \boldsymbol{\theta}) = \sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} c_{j,\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} P(R^j = 1 | Q, \boldsymbol{\alpha}, \boldsymbol{\theta}), \tag{16}$$

We create a row vector $B_{\boldsymbol{\theta},Q}(j)$ of length 2^K containing the probabilities $c_{j,\boldsymbol{\alpha}}$ for all $\boldsymbol{\alpha}$'s and arrange those elements in an appropriate order, then we write (16) in the form of a matrix product

$$\sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} c_{j,\boldsymbol{\alpha}} = B_{\boldsymbol{\theta},Q}(j)\mathbf{p},$$

where $\mathbf{p}$ is the column vector containing the probabilities $p_{\boldsymbol{\alpha}}$. For each pair of items, we may establish that the probability of responding positively to both items j_1 and j_2 is

$$P(R^{j_1} = 1, R^{j_2} = 1 | Q, \mathbf{p}, \boldsymbol{\theta}) = \sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} c_{j_1,\boldsymbol{\alpha}} c_{j_2,\boldsymbol{\alpha}} = B_{\boldsymbol{\theta},Q,}(j_1, j_2)\mathbf{p}.$$

where $B_{\boldsymbol{\theta},Q}(j_1, j_2)$ is defined as a row vector containing the probabilities $c_{j_1,\boldsymbol{\alpha}} c_{j_2,\boldsymbol{\alpha}}$ for each $\boldsymbol{\alpha}$. Note that each element of $B_{\boldsymbol{\theta},Q}(j_1, j_2)$ is the product of the corresponding elements of

$B_{\boldsymbol{\theta},Q}(j_1)$ and $B_{\boldsymbol{\theta},Q}(j_2)$. With a completely analogous construction, for items $j_1, \cdots, j_l$, we can write the probability of responding positively to all items as

$$P(R^{j_1} = 1, \ldots, R^{j_l} = 1 | Q, \mathbf{p}, \boldsymbol{\theta}) = B_{\boldsymbol{\theta},Q}(j_1, \ldots, j_l)\mathbf{p},$$

Note that $B_{\boldsymbol{\theta},Q}(j_1, \ldots, j_l)$ is the element-by-element product of $B_{\boldsymbol{\theta},Q}(j_1)$,. . . ,$B_{\boldsymbol{\theta},Q}(j_l)$.

The T-matrix for the DINA model has 2^K columns and 2^J rows. Each of the first $2^J - 1$ row vectors of the T-matrix is one of the vectors $B_{\boldsymbol{\theta},Q}(j_1, ..., j_l)$. The last row of the T-matrix is taken as $\mathbf{1}^\top$. The T-matrix can be written as

$$T_{\mathbf{c},\mathbf{g}}(Q) = \begin{pmatrix} B_{\boldsymbol{\theta},Q}(1) \\ \vdots \\ B_{\boldsymbol{\theta},Q}(J) \\ B_{\boldsymbol{\theta},Q}(1,2) \\ \vdots \\ B_{\boldsymbol{\theta},Q}(1, ..., J) \\ \mathbf{1}^\top \end{pmatrix}. \tag{17}$$

Response $\boldsymbol{\gamma}$-vector. We further define $\boldsymbol{\gamma}$ to be the vector containing the probabilities of the empirical distribution corresponding to those in $T_{\boldsymbol{\theta}}(Q)\mathbf{p}$, e.g., the first element of $\boldsymbol{\gamma}$ is $\frac{1}{N}\sum_{i=1}^N I(R_i^1 = 1)$ and the $(J+1)$-th element is $\frac{1}{N}\sum_{i=1}^N I(R_i^1 = 1 \textit{ and } R_i^2 = 1)$, i.e.,

$$\boldsymbol{\gamma} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^N I(R_i^1 = 1) \\ \vdots \\ \frac{1}{N}\sum_{i=1}^N I(R_i^J = 1) \\ \frac{1}{N}\sum_{i=1}^N I(R_i^1 = 1 \textit{ and } R_i^2 = 1) \\ \vdots \\ \frac{1}{N}\sum_{i=1}^N I(R_i^1 = 1,\ R_i^2 = 1, \cdots, \textit{ and } R_i^J = 1) \\ 1 \end{pmatrix}. \tag{18}$$

An objective function. Under the true Q-matrix Q, let $(\boldsymbol{\theta}, \mathbf{p})$ be the true model parameters. By the the law of large number, we have that

$$\boldsymbol{\gamma} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^N I(R_i^1 = 1) \\ \vdots \\ \frac{1}{N}\sum_{i=1}^N I(R_i^J = 1) \\ \frac{1}{N}\sum_{i=1}^N I(R_i^1 = 1 \textit{ and } R_i^2 = 1) \\ \vdots \end{pmatrix} \to \begin{pmatrix} P(R_i^1 = 1|Q, \boldsymbol{\theta}, \mathbf{p}) \\ \vdots \\ P(R_i^J = 1|Q, \boldsymbol{\theta}, \mathbf{p}) \\ P(R_i^1 = 1 \textit{ and } R_i^2 = 1|Q, \boldsymbol{\theta}, \mathbf{p}) \\ \vdots \end{pmatrix} = T_{\boldsymbol{\theta}}(Q)\mathbf{p}$$

almost surely as $N \to \infty$. For each Q, we define

$$S(Q) = \inf_{\mathbf{c},\mathbf{g},\mathbf{p}} |T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p} - \boldsymbol{\gamma}|^2, \tag{19}$$

where the minimization is subject to the natural constraints that $c_j, g_j, p_{\boldsymbol{\alpha}} \in (0,1)$ and $\sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = 1$. Here $|\cdot|$ means the Euclidian norm. Thanks to the law of large numbers, $S(Q) \to 0$ as $N \to \infty$. The estimator

$$\tilde{Q} = \text{argmin}_Q S(Q)$$

is consistent meaning that

$$P(\tilde{Q} \sim Q) \to 1$$

if and only if the vector $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p} \neq T_{\mathbf{c}',\mathbf{g}'}(Q')\mathbf{p}'$ for $Q' \nsim Q$ and all possible $\mathbf{c}'$, $\mathbf{g}'$ and $\mathbf{p}'$.

B Proof of Theorems

The following proposition provides a connection between the likelihood function and the T-matrix, which makes it possible to the T-matrix to show the model identifiability.

Proposition 2 *Under the DINA and DINO models, for two sets of parameters* $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}})$ *and* $(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$*,*

$$L(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}}, Q) = L(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}}, Q)$$

for all $\mathbf{R}$ *if and only if the following equation holds:*

$$T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}. \tag{20}$$

The following proposition provides a relationship between T-matrices of different model parameters.

Proposition 3 *There exists an invertible matrix* $D_{\mathbf{g}^*}$ *depending only on* $\mathbf{g}^* = (g_1^*, ..., g_J^*)$*, such that*

$$D_{\mathbf{g}^*} T_{\mathbf{c},\mathbf{g}}(Q) = T_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*}(Q).$$

Thus, (20) is equivalent to $T_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*}(Q)\bar{\mathbf{p}} = T_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*}(Q)\hat{\mathbf{p}}$ for some $\mathbf{g}^*$. This is a very important technique that will be used repeatedly in the subsequent development. We now cite a proposition.

Proposition 4 (Proposition 6.6 in Liu et al. (2013)) *For the DINA model, under Condition A1-3,* $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p}$ *is not in the column space of* $T_{\mathbf{c}',\mathbf{g}}(Q')$ *for all* $\mathbf{c}'$*, that is,* $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p} \neq T_{\mathbf{c}',\mathbf{g}}(Q')\mathbf{p}'$ *for all* $\mathbf{c}'$ *and* $\mathbf{p}'$*. In addition,* $T_{\mathbf{c},\mathbf{g}}(Q)$ *is of full column rank.*

The following proposition provides the first step result.

Proposition 5 *Under the DINA and DINO models, with* Q*,* $\mathbf{s}$*, and* $\mathbf{g}$ *being known, the population proportion parameter* $\mathbf{p}$ *is identifiable if and only if* Q *is complete.*

Proof of Proposition 5. When Q is complete, the matrix $T_{\mathbf{c},\mathbf{g}}(Q)$ has full column rank from Proposition 4. Thus, $\mathbf{p}$ is identifiable by Proposition 2.

Consider the case where the Q is incomplete. Without loss of generality, we assume $\mathbf{e}_1 = (1, 0, \cdots, 0)$ is not in the set of row vectors of Q. Then in the T-matrix $T_{\mathbf{c},\mathbf{g}}(Q)$, the columns corresponding to attribute profiles $\mathbf{0}$ and $\mathbf{e}_1$ are the same. Therefore, by

Proposition 2, we can always find two different set of estimates of $p_{\mathbf{0}}$ and $p_{\mathbf{e}_1}$ such that equation (20) holds and therefore $\mathbf{p} = (p_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \{0,1\}^K)$ is nonidentifiable. ■

Proof of Theorem 2. The identifiability of the Q-matrix for the DINO model is an application of Theorem 1 and Proposition 1. In what follows, we focus on the identifiability of the model parameters $\mathbf{c}$ and $\mathbf{p}$ under the DINA model.

We only need to show that when $\mathbf{g}$ is known, for two sets of parameters $(\hat{\mathbf{c}}, \mathbf{g}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \mathbf{g}, \bar{\mathbf{p}})$, $L(\hat{\mathbf{c}}, \mathbf{g}, \hat{\mathbf{p}}, Q) = L(\bar{\mathbf{c}}, \mathbf{g}, \bar{\mathbf{p}}, Q)$ holds if and only if A4 satisfied. By Propositions 2 and 3, two sets of parameters $(\hat{\mathbf{c}}, \mathbf{g}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \mathbf{g}, \bar{\mathbf{p}})$ yield identical likelihood if and only if

$$T_{\hat{\mathbf{c}}-\mathbf{g},\mathbf{0}}(Q)\hat{\mathbf{p}} = D_{\mathbf{g}} T_{\hat{\mathbf{c}},\mathbf{g}}(Q)\hat{\mathbf{p}} = D_{\mathbf{g}} T_{\bar{\mathbf{c}},\mathbf{g}}(Q)\bar{\mathbf{p}} = T_{\bar{\mathbf{c}}-\mathbf{g},\mathbf{0}}(Q)\bar{\mathbf{p}}. \tag{21}$$

Thus under the assumption that $c_j > g_j$, we only need to consider that $\mathbf{g} = \mathbf{0}$.

Sufficiency of A4. For notational convenience, we write $B_Q(j_1, ..., j_l) = B_{\mathbf{c},\mathbf{g},Q}(j_1, ..., j_l)$ when $\mathbf{c} = \mathbf{1}$ and $\mathbf{g} = \mathbf{0}$. For each item $j \in 1, \cdots, J$, condition A4 implies that there exist items $j_1, ..., j_l$ (different from j) such that

$$B_Q(j, j_1, ..., j_l) = B_Q(j_1, ..., j_l),$$

that is, the attributes required by item j are a subset of the attributes required by items $j_1, ..., j_l$.

Let a and a_* be the row vectors in $D_{\mathbf{g}}$ corresponding to item combinations $j_1, ..., j_l$ and $j, j_1, ..., j_l$; see (21) for the definition of $D_{\mathbf{g}}$. If $(\hat{\mathbf{c}}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \bar{\mathbf{p}})$ satisfy by (21), then

$$\frac{{a_*}^\top T_{\hat{\mathbf{c}},\mathbf{g}}(Q)\hat{\mathbf{p}}}{a^\top T_{\hat{\mathbf{c}},\mathbf{g}}(Q)\hat{\mathbf{p}}} = \frac{{a_*}^\top T_{\bar{\mathbf{c}},\mathbf{g}}(Q)\bar{\mathbf{p}}}{a^\top T_{\bar{\mathbf{c}},\mathbf{g}}(Q)\bar{\mathbf{p}}}.$$

On the other hand, we have that

$$\frac{{a_*}^\top T_{\hat{\mathbf{c}},\mathbf{g}}(Q)\hat{\mathbf{p}}}{a^\top T_{\hat{\mathbf{c}},\mathbf{g}}(Q)\hat{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\mathbf{g},\mathbf{0};Q}(j, j_1, ..., j_l)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\mathbf{g},\mathbf{0};Q}(j_1, ..., j_l)\hat{\mathbf{p}}} = \hat{c}_j - g_j,$$
$$\frac{{a_*}^\top T_{\bar{\mathbf{c}},\mathbf{g}}(Q)\bar{\mathbf{p}}}{a^\top T_{\bar{\mathbf{c}},\mathbf{g}}(Q)\bar{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\mathbf{g},\mathbf{0};Q}(j, j_1, ..., j_l)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\mathbf{g},\mathbf{0};Q}(j_1, ..., j_l)\bar{\mathbf{p}}} = \bar{c}_j - g_j.$$

Therefore, $\hat{c}_j = \bar{c}_j$ for all $j = 1, \cdots, J$, which gives the identifiability of the slipping parameter. According to Proposition 5, the completeness of the Q-matrix ensures that the identifiability of $\mathbf{p}$, therefore we have the sufficiency of A4.

Necessity of A4. We reach the conclusion by contradiction. (21) suggests that it is sufficient to show the necessity for $\mathbf{g} = \mathbf{0}$. Without loss of generality, suppose that the first attribute only appears once in the first column of the Q-matrix, i.e., the Q-matrix takes the following form:

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \mathbf{0} & Q_1 \end{pmatrix}. \tag{22}$$

We construct $\bar{\mathbf{c}}$ and $\bar{\mathbf{p}}$ different from $\hat{\mathbf{c}}$ and $\hat{\mathbf{p}}$ such that $T_{\hat{\mathbf{c}},\mathbf{0}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\mathbf{0}}(Q)\bar{\mathbf{p}}$. We write $\hat{\mathbf{c}} = (\hat{c}_1, \cdots, \hat{c}_J)$ and $\hat{\mathbf{p}} = \{\hat{p}_{(b,a)} : b \in \{0,1\}, a \in \{0,1\}^{K-1}\}$. For some x close to 1, define

$$\bar{\mathbf{c}} = (\bar{c}_1, \bar{c}_2, \cdots, \bar{c}_J) = (x\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_J)$$

and

$$\bar{\mathbf{p}} = \{\bar{p}_{(b,a)} : \bar{p}_{(1,a)} = \hat{p}_{(1,a)}/x \text{ and } \bar{p}_{(0,a)} = \hat{p}_{(0,a)} + \hat{p}_{(1,a)}(1-1/x), \text{ for all } a \in \{0,1\}^{K-1}\}.$$

Notice that the parameters related to the first item have been changed. Consider the rows in the T-matrix related to the first item. Keeping in mind that $\mathbf{g} = \mathbf{0}$, we have that

$$\hat{c}_1 \sum_{a\in\{0,1\}^{K-1}} \hat{p}_{(1,a)} + g_1 \sum_{a\in\{0,1\}^{K-1}} \hat{p}_{(0,a)} = \bar{c}_1 \sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(1,a)} + g_1 \sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(0,a)}. \qquad (23)$$

This corresponds to $P(R^1 = 1)$. Similar identities can be established for $P(R^1 = R^{j_1} = ... = R^{j_l} = 1)$. Therefore, we have constructed $(\bar{\mathbf{c}}, \bar{\mathbf{p}}) \neq (\hat{\mathbf{c}}, \hat{\mathbf{p}})$ such that $T_{\bar{\mathbf{c}},\mathbf{0}}(Q)\bar{\mathbf{p}} = T_{\hat{\mathbf{c}},0}(Q)\hat{\mathbf{p}}$. Thus, $\mathbf{c}$ and $\mathbf{p}$ are not identifiable if A4 does not hold. ■

Proof of Theorem 3. Consider the true Q and a candidate $Q' \nsim Q$. According to the discussion at the end of Section A, it is sufficient to show that it is impossible to have two sets of parameters $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ such that $\hat{c}_j > \hat{g}_j$, $\bar{c}_j > \bar{g}_j$, $\hat{p}_{\boldsymbol{\alpha}} > 0$, $\bar{p}_{\boldsymbol{\alpha}} > 0$, and

$$T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q')\bar{\mathbf{p}}. \qquad (24)$$

We prove this first assuming that there exist two such sets of parameters and then reach a contradiction. The true matrix Q is arranged as in (8) such that the first $2K$ rows form two identity matrices. We try to reach a contradiction under the following two cases.

Case 1: either $Q'_{1:K}$ or $Q'_{K+1:2K}$ is incomplete. We only focus on the case when $Q'_{1:K}$ is not $\mathcal{I}_K$. We borrow an intermediate result in the proof of Proposition 6.4 in Liu et al. (2013): we can identify an item $1 \le h \le K$ and an item set $\mathcal{H} \subset \{1, \cdots, K\}$ ($h \notin \mathcal{H}$) such that under Q', $\mathcal{H}$ requires all attributes required by item h, that is, if someone is capable of solving all problems in $\mathcal{H}$ then he/she is able to solve problem h. We say someone "is able to" or "can" solve a problem or a set of problems if his/her ideal responses to the set of problems are all one.

For items $K+1, \cdots, 2K$, since $Q_{K+1:2K} = \mathcal{I}_K$, there exists an item set $\mathcal{B} \subset \{K+1, ..., 2K\}$ such that under Q it requires the same attributes as $\mathcal{H}$, that is, if a person is capable of solving all items in $\mathcal{B}$ if and only if they can solving all problems in $\mathcal{H}$. Since $Q_{1:K} = \mathcal{I}_K$, under Q, the attributes required by $\mathcal{H}$ and $\mathcal{B}$ are different from those of item h. Define

$$\tilde{\mathbf{g}} = (\bar{g}_1, \cdots, \bar{g}_K, \hat{g}_{K+1}, \cdots, \hat{g}_J).$$

Assumption (24) and Proposition 3 suggests $T_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}}}(Q')\bar{\mathbf{p}}$.

Under Q' if h requires strictly fewer attributes than $\mathcal{H}$, there are three types of attributes profiles: unable to answer h (denoted by $0_h0_{\mathcal{H}}$), unable to answer $\mathcal{H}$ but able to answer h

(denoted by $0_{\mathcal{H}}1_h$), and able to answer $\mathcal{H}$ (denoted by $1_{\mathcal{H}}$). We have

$$\begin{array}{rcccl}
 & 0_h 0_{\mathcal{H}} & 0_{\mathcal{H}} 1_h & 1_{\mathcal{H}} & \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H}) = (& 0 & 0 & \prod_{j\in\mathcal{H}}(\bar{c}_j-\bar{g}_j) &), \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(h) = (& 0 & (\bar{c}_h-\bar{g}_h) & (\bar{c}_h-\bar{g}_h) &), \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},h) = (& 0 & 0 & (\bar{c}_h-\bar{g}_h)\prod_{j\in\mathcal{H}}(\bar{c}_j-\bar{g}_j) &),
\end{array}$$

If h and $\mathcal{H}$ require the same attributes, $0_{\mathcal{H}}1_h$ case does not exist and the above equations do not have the $0_{\mathcal{H}}1_h$ column. Under both situations, we have

$$\bar{c}_h-\bar{g}_h = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},h)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H})\bar{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},h,K+1,\cdots,2K)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},K+1,\cdots,2K)\bar{\mathbf{p}}}. \tag{25}$$

Under Q, we have

$$\begin{array}{rccl}
 & \boldsymbol{\alpha}\neq\mathbf{1} & \boldsymbol{\alpha}=\mathbf{1} & \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(K+1,\cdots,2K) = (& 0 & \prod_{j=K+1}^{2K}(\hat{c}_j-\hat{g}_j) &), \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},K+1,\cdots,2K) = (& 0 & \prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j=K+1}^{2K}(\hat{c}_j-\hat{g}_j) &), \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},h,K+1,\cdots,2K) = (& 0 & (\hat{c}_h-\bar{g}_h)\prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j=K+1}^{2K}(\hat{c}_j-\hat{g}_j) &).
\end{array}$$

This gives

$$\hat{c}_h-\bar{g}_h = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},h,K+1,\cdots,2K)\hat{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},K+1,\cdots,2K)\hat{\mathbf{p}}}. \tag{26}$$

$T_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}}}(Q')\bar{\mathbf{p}}$ allows to equate the right-hand sides of (25) and (26) which yields

$$\hat{c}_h = \bar{c}_h. \tag{27}$$

Now under Q', with a similarly argument, we have

$$\bar{c}_h-\bar{g}_h = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},h,\mathcal{B})\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},\mathcal{B})\bar{\mathbf{p}}}. \tag{28}$$

Under Q, consider three types of attributes profiles: unable to answer $\mathcal{H}$ (denoted by $0_{\mathcal{H}}$), able to answer $\mathcal{H}$ but unable to answer h (denoted by $0_h1_{\mathcal{H}}$), and able to answer both $\mathcal{H}$ and h (denoted by $1_{\mathcal{H}}1_h$). We have

$$\begin{array}{rccc}
 & 0_{\mathcal{H}} & 0_h 1_{\mathcal{H}} & 1_{\mathcal{H}} 1_h \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},\mathcal{B}) = (& 0 & \prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j\in\mathcal{B}}(\hat{c}_j-\hat{g}_j) & \prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j\in\mathcal{B}}(\hat{c}_j-\hat{g}_j)), \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},h,\mathcal{B}) = (& 0 & (\hat{g}_h-\bar{g}_h)\prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j\in\mathcal{B}}(\hat{c}_j-\hat{g}_j) & (\hat{c}_h-\bar{g}_h)\prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j\in\mathcal{B}}(\hat{c}_j-\hat{g}_j)).
\end{array}$$

Since $\hat{g}_h-\bar{g}_h < \hat{c}_h-\bar{g}_h$ and $p_{\boldsymbol{\alpha}} > 0$ for all $\boldsymbol{\alpha}$, we have that

$$\hat{c}_h-\bar{g}_h \neq \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},h,\mathcal{B})\hat{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},\mathcal{B})\hat{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},h,\mathcal{B})\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},\mathcal{B})\bar{\mathbf{p}}}. \tag{29}$$

$T_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}}}(Q')\bar{\mathbf{p}}$ allows use to equate the right-hand sides of (28) and (29), which yields $\hat{c}_h > \bar{c}_h$. This contradicts (27).

Thus, under this case, we have that $T_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}}}(Q)\hat{\mathbf{p}} \neq T_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}}}(Q')\bar{\mathbf{p}}$ if $\hat{c}_j > \hat{g}_j$, $\bar{c}_j > \bar{g}_j$, $\hat{p}_{\boldsymbol{\alpha}} > 0$, $\bar{p}_{\boldsymbol{\alpha}} > 0$. Furthermore, if the conditions in the theorem are satisfied and $Q'_{1:K}$ or $Q'_{(K+1):2K}$ is incomplete, then we cannot find parameters $\bar{\mathbf{c}}$, $\bar{\mathbf{g}}$, and $\bar{\mathbf{p}}$ that yields the same response distribution as Q and thus Q can be differentiated from Q' by the maximum likelihood.

Case 2: both $Q'_{1:K}$ and $Q'_{K+1:2K}$ are complete, but $Q \nsim Q'$. In this case, we can always arrange the columns of Q' such that either $Q'_{1:K} = \mathcal{I}_K$. Redefine

$$\tilde{\mathbf{g}} = (\bar{c}_1, \cdots, \bar{c}_K, \hat{c}_{K+1}, \cdots, \hat{c}_{2K}, 0, \cdots, 0)$$

and assumption (24) suggests that $T_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}}}(Q')\bar{\mathbf{p}}$.

The row vectors of T-matrices corresponding to items 1,..., $2K$ are

$$B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, 2K) = \left(\prod\nolimits_{k=1}^{K}(\hat{g}_k - \bar{c}_k)\prod\nolimits_{k=K+1}^{2K}(\hat{g}_k - \hat{c}_k), \mathbf{0}^\top\right)$$

and

$$B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, 2K) = \left(\prod\nolimits_{k=1}^{K}(\bar{g}_k - \bar{c}_k)\prod\nolimits_{k=K+1}^{2K}(\bar{g}_k - \hat{c}_k), \mathbf{0}^\top\right)$$

where only the element corresponding to zero attribute is non-zero. Therefore, for any $j \geq 2K+1$, we have

$$\hat{g}_j = \frac{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, 2K,\, j)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, 2K)\hat{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, 2K,\, j)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, 2K)\bar{\mathbf{p}}} = \bar{g}_j.$$

Once again, we redefine $\tilde{\mathbf{g}} = (\bar{g}_1, \cdots, \bar{g}_K, 0, \cdots, 0, \hat{g}_{2K+1}, \cdots, \hat{g}_J)$. By Condition A5, we have for $K+1 \leq j \leq 2K$

$$\begin{aligned}\hat{c}_j &= \frac{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, K,\, j,\, (2K+1), \cdots, J)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, K, (2K+1), \cdots, J)\hat{\mathbf{p}}} \\ &= \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, K,\, j,\, (2K+1), \cdots, J)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, K, (2K+1), \cdots, J)\bar{\mathbf{p}}} = \bar{c}_j.\end{aligned}$$

Similarly take $\tilde{\mathbf{g}} = (0, \cdots, 0, \bar{g}_{K+1}, \cdots, \bar{g}_{2K}, \hat{g}_{2K+1}, \cdots, \hat{g}_J)$. We have $\hat{c}_j = \bar{c}_j$ for $1 \leq j \leq K$.

Now take $\tilde{\mathbf{g}} = (\bar{c}_1, \cdots, \bar{c}_K, 0, \cdots, 0)$, we have for $K+1 \leq j \leq 2K$

$$\hat{g}_j = \frac{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, K,\, j)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, K)\hat{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, K,\, j)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, K)\bar{\mathbf{p}}} = \bar{g}_j.$$

Similarly, for $\hat{g}_j = \bar{g}_j$ for $j = 1, ..., K$. Thus, we have $\hat{g}_j = \bar{g}_j$ for $j = 1, ..., J$. Therefore, assumption (24) becomes

$$T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\hat{\mathbf{g}}}(Q')\bar{\mathbf{p}}. \tag{30}$$

This contradicts Proposition 4. Thus, we have reached the conclusion that

$$T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} \neq T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q')\bar{\mathbf{p}}.$$

for all $\hat{c}_j > \hat{g}_j$, $\bar{c}_j > \bar{g}_j$, $\hat{p}_{\boldsymbol{\alpha}} > 0$, $\bar{p}_{\boldsymbol{\alpha}} > 0$ and $Q' \nsim Q$. Thus, by maximizing the profiled likelihood, Q can be consistently estimated. ■

Proof of Theorem 4. Suppose there are two sets of parameters $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ such that $L(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}}) = L(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$, equivalently, $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$. We show that $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}}) = (\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ if $\hat{c}_j > \hat{g}_j$, $\hat{p}_{\boldsymbol{\alpha}} > 0$, $\bar{c}_j > \bar{g}_j$, and $\bar{p}_{\boldsymbol{\alpha}} > 0$. Condition A5 allows us to consider the following three cases.

Case 1. There exit at least three items with Q-matrix row vector $\mathbf{e}_1$. Without loss of generality, we write the Q-matrix as (with reordering of the rows)

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \mathbf{0}^\top \\ 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \mathbf{0} & Q' \end{pmatrix}. \tag{31}$$

In what follows, we show that $\hat{c}_j = \bar{c}_j$ and $\hat{g}_j = \bar{g}_j$ for $j = 1, 2, 3$. By Proposition 3, $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$ suggests that $T_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}}}(Q)\bar{\mathbf{p}}$. Together with the fact that

$$\frac{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1,2,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1,2)\hat{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1)\hat{\mathbf{p}}} = \hat{c}_3 - \hat{g}_3, \tag{32}$$

we have that

$$\frac{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1)\bar{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1,2,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1,2)\bar{\mathbf{p}}}. \tag{33}$$

Expanding the above identity, we have

$$\begin{aligned} &\frac{(\bar{g}_1 - \hat{g}_1)(\bar{g}_3 - \hat{g}_3)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(0,a)} + (\bar{c}_1 - \hat{g}_1)(\bar{c}_3 - \hat{g}_3)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(1,a)}}{(\bar{g}_1 - \hat{g}_1)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(0,a)} + (\bar{c}_1 - \hat{g}_1)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(1,a)}} \\ = \; &\frac{\prod_{j=1}^3(\bar{g}_j - \hat{g}_j)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(0,a)} + \prod_{j=1}^3(\bar{c}_j - \hat{g}_j)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(1,a)}}{(\bar{g}_1 - \hat{g}_1)(\bar{g}_2 - \hat{g}_2)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(0,a)} + (\bar{c}_1 - \hat{g}_1)(\bar{c}_2 - \hat{g}_2)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(1,a)}}, \end{aligned} \tag{34}$$

which can be simplified to $(\bar{g}_1 - \hat{g}_1)(\bar{c}_1 - \hat{g}_1)(\bar{c}_2 - \bar{g}_2)(\bar{c}_3 - \bar{g}_3) = 0$. Then under the constraint that $\bar{c}_j > \bar{g}_j$, we have $\bar{g}_1 = \hat{g}_1$ or $\bar{c}_1 = \hat{g}_1$. A similar argument yields

$$\begin{cases} \bar{g}_2 = \hat{g}_2 \text{ or } \bar{c}_2 = \hat{g}_2 \\ \bar{g}_3 = \hat{g}_3 \text{ or } \bar{c}_3 = \hat{g}_3 \end{cases} \text{ and } \begin{cases} \hat{g}_1 = \bar{g}_1 \text{ or } \hat{c}_1 = \bar{g}_1 \\ \hat{g}_2 = \bar{g}_2 \text{ or } \hat{c}_2 = \bar{g}_2 \\ \hat{g}_3 = \bar{g}_3 \text{ or } \hat{c}_3 = \bar{g}_3 \end{cases}.$$

For $j = 1$, 2, or 3, if $\hat{g}_j \neq \bar{g}_j$ we have $\hat{c}_j = \bar{g}_j$ and $\bar{c}_j = \hat{g}_j$. This contradict the condition that $\hat{c}_j > \hat{g}_j$ and $\bar{c}_j > \bar{g}_j$. Thus we have $\hat{g}_j = \bar{g}_j$ for $j = 1, 2, 3$. Repeating the proof of Theorem 2, we have $\hat{c}_j = \bar{c}_j$ for $i = 1, 2, 3$.

Case 2. There exit two items with row vector $\mathbf{e}_1$. Without loss of generality, we write the Q-matrix as

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \mathbf{0}^\top \\ 1 & \mathbf{v}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \mathbf{0} & Q' \end{pmatrix}, \quad Q_{1:4} = \begin{pmatrix} 1 & 0 & \mathbf{0}^\top \\ 1 & 0 & \mathbf{0}^\top \\ 1 & 1 & {\mathbf{v}_*}^\top \\ 0 & 1 & \mathbf{0}^\top \end{pmatrix}, \tag{35}$$

where $\mathbf{v}$ is a non-zero vector. Without loss of generality we assume $\mathbf{v}^\top = (1, {\mathbf{v}_*}^\top)$. Consider the sub-matrix containing the first four items. i.e., $Q_{1:4}$ in (35). Similar to the proof of Case 1, for $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ such that $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$, we will show

$$\begin{cases} \hat{c}_j = \bar{c}_j & j = 1, 2, 4 \\ \hat{g}_j = \bar{g}_j & j = 1, 2, 3 \end{cases}. \tag{36}$$

A similar argument as in Case 1 yields

$$\frac{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(3)\hat{\mathbf{p}}} = \hat{c}_1 - \hat{g}_1 = \frac{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1,4,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(4,3)\hat{\mathbf{p}}}.$$

Together with the fact that $T_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}}}(Q)\bar{\mathbf{p}} = T_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0}}(Q)\hat{\mathbf{p}}$, we have

$$\frac{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(3)\bar{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1,4,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(4,3)\bar{\mathbf{p}}}.$$

This implies

$$\begin{aligned} & \frac{\tilde{g}_1\tilde{g}_4\tilde{g}_3\bar{\mathbf{p}}_{0,0} + \tilde{c}_1\tilde{g}_4\tilde{g}_3\bar{\mathbf{p}}_{1,0} + \tilde{g}_1\tilde{c}_4\tilde{g}_3\bar{\mathbf{p}}_{0,1} + \tilde{c}_1\tilde{c}_4\tilde{c}_3\bar{\mathbf{p}}_{1,1}}{\tilde{g}_4\tilde{g}_3\bar{\mathbf{p}}_{0,0} + \tilde{g}_4\tilde{g}_3\bar{\mathbf{p}}_{1,0} + \tilde{c}_4\tilde{g}_3\bar{\mathbf{p}}_{0,1} + \tilde{c}_4\tilde{c}_3\bar{\mathbf{p}}_{1,1}} \\ = & \frac{\tilde{g}_1\tilde{g}_3\bar{\mathbf{p}}_{0,0} + \tilde{c}_1\tilde{g}_3\bar{\mathbf{p}}_{1,0} + \tilde{g}_1\tilde{g}_3\bar{\mathbf{p}}_{0,1} + \tilde{c}_1\tilde{c}_3\bar{\mathbf{p}}_{1,1}}{\tilde{g}_3\bar{\mathbf{p}}_{0,0} + \tilde{g}_3\bar{\mathbf{p}}_{1,0} + \tilde{g}_3\bar{\mathbf{p}}_{0,1} + \tilde{c}_3\bar{\mathbf{p}}_{1,1}}, \end{aligned} \tag{37}$$

where $\tilde{g}_j = \bar{g}_j - \hat{g}_j$ for $j = 1, 3, 4$, $\tilde{c}_j = \bar{c}_j - \hat{g}_j$ for $j = 1, 4$,

$$\tilde{c}_3 = \frac{(\bar{c}_3 - \hat{g}_3)\sum_{\mathbf{v}_* \preceq a \in \{0,1\}^{K-2}} \bar{p}_{(1,1,a)} + (\bar{g}_3 - \hat{g}_3)\sum_{\mathbf{v}_* \not\preceq a \in \{0,1\}^{K-2}} \bar{p}_{(1,1,a)}}{\sum_{a \in \{0,1\}^{K-2}} \bar{p}_{(1,1,a)}},$$

and $\bar{\mathbf{p}}_{i,j} = \sum_{a \in \{0,1\}^{K-2}} \bar{p}_{(i,j,a)}$ for $i, j \in \{0, 1\}$. Here $\mathbf{v}_* \preceq a$ means that each element of $\mathbf{v}_*$ is less than or equals to the corresponding element of a, and $\mathbf{v}_* \not\preceq a$ means that $\mathbf{v}_* \preceq a$ does not hold.

Simplifying (37), we obtain $\bar{\mathbf{p}}_{0,0}\bar{\mathbf{p}}_{1,1}\tilde{g}_3\tilde{c}_3(\tilde{g}_1 - \tilde{c}_1) = \bar{\mathbf{p}}_{1,0}\bar{\mathbf{p}}_{0,1}\tilde{g}_3\tilde{g}_3(\tilde{g}_1 - \tilde{c}_1)$. Since $\tilde{g}_1 - \tilde{c}_1 \neq 0$, we have

$$\tilde{g}_3 = 0 \quad \text{or} \quad \bar{\mathbf{p}}_{0,0}\bar{\mathbf{p}}_{1,1}\tilde{c}_3 = \bar{\mathbf{p}}_{1,0}\bar{\mathbf{p}}_{0,1}\tilde{g}_3. \tag{38}$$

We show that $\tilde{g}_3$ has to be zero. Otherwise, we have

$$\bar{\mathbf{p}}_{0,0}\bar{\mathbf{p}}_{1,1}(\bar{c}_3^* - \hat{g}_3) = \bar{\mathbf{p}}_{1,0}\bar{\mathbf{p}}_{0,1}(\bar{g}_3 - \hat{g}_3), \tag{39}$$

where

$$\bar{c}_3^* = \tilde{c}_3 + \hat{g}_3 = \frac{\bar{c}_3 \sum_{v_* \preceq a \in \{0,1\}^{K-2}} \bar{p}_{(1,1,a)} + \bar{g}_3 \sum_{v_* \npreceq a \in \{0,1\}^{K-2}} \bar{p}_{(1,1,a)}}{\sum_{a \in \{0,1\}^{K-2}} \bar{p}_{(1,1,a)}}.$$

A similar argument gives that

$$\hat{\mathbf{p}}_{0,0}\hat{\mathbf{p}}_{1,1}(\hat{c}_3^* - \bar{g}_3) = \hat{\mathbf{p}}_{1,0}\hat{\mathbf{p}}_{0,1}(\hat{g}_3 - \bar{g}_3), \tag{40}$$

where

$$\hat{c}_3^* = \frac{\hat{c}_3 \sum_{v_* \preceq a \in \{0,1\}^{K-2}} \hat{p}_{(1,1,a)} + \hat{g}_3 \sum_{v_* \npreceq a \in \{0,1\}^{K-2}} \hat{p}_{(1,1,a)}}{\sum_{a \in \{0,1\}^{K-2}} \hat{p}_{(1,1,a)}}.$$

Equations (39) and (40) imply that $\hat{c}_3^* > \hat{g}_3 > \bar{c}_3^* > \bar{g}_3$ or $\bar{c}_3^* > \bar{g}_3 > \hat{c}_3^* > \hat{g}_3$, which conflicts with the equation that $B_{\hat{\mathbf{c}},\hat{\mathbf{g}};Q}(3)\hat{\mathbf{p}} = B_{\bar{\mathbf{c}},\bar{\mathbf{g}};Q}(3)\bar{\mathbf{p}}$, i.e.,

$$\hat{g}_3(\hat{\mathbf{p}}_{0,0} + \hat{\mathbf{p}}_{1,0} + \hat{\mathbf{p}}_{0,1}) + \hat{c}_3^*\hat{\mathbf{p}}_{1,1} = \bar{g}_3(\bar{\mathbf{p}}_{0,0} + \bar{\mathbf{p}}_{1,0} + \bar{\mathbf{p}}_{0,1}) + \bar{c}_3^*\bar{\mathbf{p}}_{1,1}.$$

To see this, notice that $\hat{\mathbf{p}}_{0,0} + \hat{\mathbf{p}}_{1,0} + \hat{\mathbf{p}}_{0,1} = 1 - \hat{\mathbf{p}}_{1,1}$, $\bar{\mathbf{p}}_{0,0} + \bar{\mathbf{p}}_{1,0} + \bar{\mathbf{p}}_{0,1} = 1 - \bar{\mathbf{p}}_{1,1}$, and $\hat{\mathbf{p}}_{1,1}, \bar{\mathbf{p}}_{1,1} \in (0,1)$. By simple algebra, the above identity cannot be achieved if either $\hat{c}_3^* > \hat{g}_3 > \bar{c}_3^* > \bar{g}_3$ or $\bar{c}_3^* > \bar{g}_3 > \hat{c}_3^* > \hat{g}_3$ is true. Therefore, we have $\tilde{g}_3 = \bar{g}_3 - \hat{g}_3 = 0$. Let $\underline{\mathbf{g}} = (0, 0, \hat{g}_3, 0, \cdots, 0)$. $T_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}}}(Q)\bar{\mathbf{p}} = T_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}}}(Q)\hat{\mathbf{p}}$ yields

$$\begin{aligned}
\bar{c}_1 &= \frac{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,4,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(4,3)\bar{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,4,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(4,3)\hat{\mathbf{p}}} = \hat{c}_1,\\
\bar{c}_2 &= \frac{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(2,4,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(4,3)\bar{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(2,4,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(4,3)\hat{\mathbf{p}}} = \hat{c}_2,\\
\bar{c}_4 &= \frac{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,4,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,3)\bar{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,4,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,3)\hat{\mathbf{p}}} = \hat{c}_4.
\end{aligned}$$

Consider items 1 and 2. Let $\underline{\mathbf{c}} = (\hat{c}_1, \hat{c}_2, 0, \cdots, 0)$. $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$ yields

$$\begin{aligned}
\bar{g}_1 &= \frac{B_{\bar{\mathbf{c}}-\underline{\mathbf{c}},\bar{\mathbf{g}}-\underline{\mathbf{c}};Q}(1,2)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\underline{\mathbf{c}},\bar{\mathbf{g}}-\underline{\mathbf{c}};Q}(2)\bar{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\underline{\mathbf{c}},\hat{\mathbf{g}}-\underline{\mathbf{c}};Q}(1,2)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\underline{\mathbf{c}},\hat{\mathbf{g}}-\underline{\mathbf{c}};Q}(2)\hat{\mathbf{p}}} = \hat{g}_1,\\
\bar{g}_2 &= \frac{B_{\bar{\mathbf{c}}-\underline{\mathbf{c}},\bar{\mathbf{g}}-\underline{\mathbf{c}};Q}(1,2)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\underline{\mathbf{c}},\bar{\mathbf{g}}-\underline{\mathbf{c}};Q}(1)\bar{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\underline{\mathbf{c}},\hat{\mathbf{g}}-\underline{\mathbf{c}};Q}(1,2)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\underline{\mathbf{c}},\hat{\mathbf{g}}-\underline{\mathbf{c}};Q}(1)\hat{\mathbf{p}}} = \hat{g}_2.
\end{aligned}$$

Therefore, (36) is true.

Now combining the results in Cases 1 and 2, we have that for the Q-matrix taking the form of (8), the following holds:

$$\begin{cases} \hat{c}_j = \bar{c}_j & j = 1, \cdots, 2K \\ \hat{g}_j = \bar{g}_j & j = 1, \cdots, J \end{cases}. \tag{41}$$

Let $\mathbf{g}^* = (\hat{c}_1, \cdots, \hat{c}_K, \hat{g}_{K+1}, \cdots, \hat{g}_J)$. For each $j \in \{(2K+1), \cdots, J\}$, let $\mathcal{A}_j$ be the set of items $\{(K+1), \cdots, J\}\backslash\{j\}$, i.e., the set of all items from $K+1$ to J except the jth one. For the sub-matrix $Q_{K+1:J}$, condition A5 implies that each attribute appears at least twice.

Therefore, we have

$$\hat{c}_j - \hat{g}_j = \frac{B_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{A}_j, j)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{A}_j)\hat{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{A}_j, j)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{A}_j)\bar{\mathbf{p}}} = \bar{c}_j - \hat{g}_j.$$

This gives $\hat{c}_j = \bar{c}_j$ for $j = 2K+1, \cdots, J$. Together with (41), $\hat{c}_j = \bar{c}_j$ for all $j = 1, \cdots, J$. This further yields $\hat{\mathbf{p}} = \bar{\mathbf{p}}$ due to the full column rank of the matrix $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)$.

Therefore, for two sets of parameters $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ such that $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$, we have $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}}) = (\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$. This finishes the proof of Theorem 4. ■

C Proof of Propositions

Proof of Proposition 2. Notice that the column vector $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p}$ contains the probabilities $P(R_{j_1} = 1, ..., R_{j_l} = 1)$ for all possible distinct combinations j_1,...,j_l. Thus, $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p}$ completely characterizes the distribution of $\mathbf{R}$. Two sets of parameters $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$ if and only if they correspond to the same distribution of $\mathbf{R}$. This concludes the proof. ■

Proof of the Proposition 3. In what follows, we construct a D matrix satisfying the condition in the proposition. We show that there exists a matrix D only depending on g^* so that $DT_{\mathbf{c},\mathbf{g}}(Q) = T_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*}(Q)$. Note that each row of $DT_{\mathbf{c},\mathbf{g}}(Q)$ is just a row linear transform of $T_{\mathbf{c},\mathbf{g}}(Q)$. Then, it is sufficient to show that each row vector of $T_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*}(Q)$ is a linear transform of rows of $T_{\mathbf{c},\mathbf{g}}(Q)$ with coefficients only depending on g^*. We prove this by induction.

First, note that

$$B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j) = B_{\mathbf{c},\mathbf{g};Q}(j) - g^*_j \mathbf{1}^\top$$

where $\mathbf{1}^\top$ is a row vector with all elements being 1. Then all row vectors of $T_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*}(Q)$ of the form $B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*,Q}(j)$ are inside the row space of $T_{\mathbf{c},\mathbf{g}}(Q)$ with coefficients only depending on g^*. Suppose that all the vectors of the form

$$B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_1, ..., j_l)$$

for all $1 \leq l \leq \iota$ can be written linear combinations of the row vectors of $T_{\mathbf{c},\mathbf{g}}(Q)$ with coefficients only depending on g^*. Then, we consider

$$B_{\mathbf{c},\mathbf{g};Q}(j_1, ..., j_{\iota+1}) = \Upsilon_{h=1}^{\iota+1} \left(B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_h) + g^*_{j_h}\mathbf{1}^\top \right),$$

where "Υ" refers to element by element multiplication. The left hand side is just a row vector of $T_{\mathbf{c},\mathbf{g}}(Q)$. We expand the right hand side of the above display. Note that the last term is precisely

$$B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_1, ..., j_{\iota+1}) = \Upsilon_{h=1}^{\iota+1} B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_h).$$

The rest terms are all of the form $B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_1, ..., j_l)$ for $1 \leq l \leq \iota$ multiplied by coefficients only depending on g^*. Therefore, according to the induction assumption, we have that $B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_1, ..., j_{\iota+1})$ can be written as linear combinations of rows of $T_{\mathbf{c},\mathbf{g}}(Q)$ with coefficients only depending on g^*. ■