A Graphical Diagnostic Classification Model

Hyeon-Ah Kang, Jingchen Liu, Zhiliang Ying

Columbia University

Abstract

A framework is presented to model instances and degrees of local item dependence within the context of diagnostic classification models (DCMs). The study considers an undirected graphical model to describe dependent structure of test items and draws inference based on pseudo-likelihood. The new modeling framework explicitly addresses item interactions beyond those explained by latent classes and thus is more flexible and robust against the violation of local independence. It also facilitates concise interpretation of item relations by regulating complexity of a network underlying the test items. The viability and effectiveness are demonstrated via simulation and a real data example. Results from the simulation study suggest that the proposed methods adequately recover the model parameters in the presence of locally dependent items and lead to a substantial improvement in estimation accuracy compared to the standard DCM approach. The analysis of real data demonstrates that the graphical DCM provides a useful summary of item interactions in regards to the existence and extent of local dependence.

1 Introduction

The cognitive diagnostic analysis makes use of statistical probability models to characterize response patterns on a test. In psychometric context these models are often referred to as diagnostic classification models (DCMs; e.g., de la Torre, 2011; Haertel, 1989; Henson, Templin, & Willse, 2009; Junker & Sijtsma, 2001; Templin & Henson, 2006; von Davier, 2005). The primary function of the DCM is to classify individuals with respect to an array of attributes and summarize response profiles with a small number of latent classes. The underlying attribute is typically denoted by a binary value, indicating whether or not an examinee possesses the attribute being examined. In practice the presence or absence of particular attributes serves as useful diagnostic information
that can reveal examinee’s degree of possession of knowledge, skill, syndrome, or pathological symptom. For this reason, the DCMs have been used in a variety of social sectors, including education, psychology, and clinical settings.

One of the fundamental assumptions made in the DCMs is local item independence given a latent class. This assumption connotes that the latent class of interest provides all the necessary information about a given examinee, and after controlling for the latent variable, items are not expected to have any remaining association. While the assumption of conditional independence greatly simplifies the representation of the joint response distribution and consequently parameter estimation, it appears a strong assumption that may not always be justified. In real applications the violation of local item independence is often expected in test items involving common stimuli, similar contents or wording, order effects, speededness, etc. (Yen, 1984, 1993). Studies have shown that when local item dependence is present, parameter estimation and statistical properties of a test can be negatively impacted (see, e.g., C.-T. Chen & Wang, 2007; Kingston & Dorans, 1984; Sireci, Thissen, & Wainer, 1991; Tuerlinckx & De Boeck, 2001; Wainer & Thissen, 1996; Yen, 1984). The dependence structure, if not properly addressed, can result in biased estimation of item parameters as well as overestimation of measurement precision and reliability.

According to W. Chen and Thissen (1997), local item dependence can be classified into two kinds: underlying dependence and surface dependence. The first type assumes that local item dependence occurs as a result of a unmodeled latent variable. It posits that there exists an extra latent parameter that is common to a set of locally dependent items and yet irrelevant to the rest of items in a test. The bifactor models (Cai, 2010; Gibbons & Hedeker, 1992) and the random-effects testlet response models (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Wang, 2007; X. Wang, Bradlow, & Wainer, 2002; W. Wang & Wilson, 2005) are devised to address this type of local dependence. The bifactor models consider a secondary group factor to account for item association that cannot be explained by a primary latent factor. The testlet response models introduce a random effect parameter to model local dependence within the same testlet or item bundle. Unlike the bifactor models in which the covariance among the items (after controlling the primary factor) is attributed to within-person variance, the testlet response models ascribe local dependence to an interaction effect between a person and an item cluster. Note that both the modeling approaches evaluate local dependence in a confirmatory fashion by requiring prior
knowledge about the dependence structure of test items. When no such information is available a priori, they may lend little utility.

The other type, surface local dependence, arises when items are highly similar in contents or locations. When it takes place, it is assumed that endorsing an item is influenced by other items in the test. A salient example is a set of questions that are phrased slightly differently, while in point of fact, the items function as if they are a single question, measuring the same aspect of the construct. Consider a pair of perfectly locally dependent items for instance. In this event the item that comes later has a response identical to the preceding one without regard to the latent trait variable. In other words, the underlying cognitive process implied by the model does not play any role in determining the performance of the second item. The currently available approaches for identifying the existence of surface local dependence are to use general-purpose goodness-of-fit indices, such as Pearson’s $\chi^2$ or the likelihood ratio test statistic (Bishop, Fienberg, & Holland, 1975, p. 57; Chen & Thissen, 1997). Yen’s $Q_3$ statistic (Yen, 1984, 1993), score test statistics (Glas & Suárez Falcón, 2003; Y. Liu & Thissen, 2012), $M_3$ statistic (Maydeu-Olivares & Joe, 2005, 2006), etc. These indices, however, are not specifically aimed at detecting surface local dependence. Rather, they are designed to assess overall tenability of local independence or item misfit concerning the model assumptions (e.g., dimensionality, local independence, item characteristic function, distribution of latent variables). Even the indices designed for detecting local item dependence can only identify violations of a weak form of local independence, which deals with bivariate relationships of the items (see Junker, 1993; McDonald, 1979, 1982; Stout, 2002). Aside from the self-imposed limitation, the nature of pairwise comparison in these indices presents problems when only a few observations are available for compiling contingency tables.

The purpose of this article is to provide a parametric approach for identifying the presence of surface local dependence within the framework of DCMs. It is useful to note that the aforementioned parametric models, though developed in the item response theory context, can be readily extended to DCM applications by appending extra latent parameters. The major distinction of the current work from the previous efforts is that, while the prior approaches relate local dependence partially or exclusively to unmodeled latent traits, the present study attempts to address local dependence arising from an interplay among the items. Specifically, this study assumes that local dependence occurs as a consequence of an interaction among items but still and all the latent class of interest
plays a dominant role in determining the item performances. It is also noteworthy that the existing parametric models are mainly designed for confirmatory analysis in which item subsets that are suspected of local dependence need to be determined beforehand. The present modeling involves no such pre-identification and hence can be applied to more general contexts, for example, when dependence structure underlying test items is unknown or when one wishes to explore any lurking violations of the local independence assumption.

For modeling conditional (in)dependence of items, this study considers an undirected graphical model, also known as Markov random field. A graphical model is a powerful tool for summarizing stochastic dependency of random variables in multivariate data and has been used in a wide variety of scientific domains, including spatial statistics, image analysis, natural language processing, among others. In this study, a sparse Markov network is considered for characterizing relationships among categorical variables by means of a sparse graph. The sparsity constraint is imposed such that the latent class of concern accounts for most variation in observed response profile and only a fraction of variation is ascribed to interdependency of items. Such assumption is easily fulfilled in most psychometric settings as it implies lack of test validity otherwise. The proposed modeling framework has two distinct merits. First, it is arguably more flexible than standard DCMs as it allows for nesting of items that cannot be explained by latent classes intended. In addition, it is parsimonious enough to offer a succinct interpretation of the model parameters in virtue of a sparse graph. In applied settings, the new modeling approach can serve as useful means for exploring the possible existence of local item dependence or analyzing data with such dependency.

The remainder of this article is laid out as follows. Section 2 presents a specific parameterization of the suggested model, namely the graphical diagnostic classification model (GDCM), followed by a brief introduction of the DCM and Markov network. A parameter estimation scheme and corresponding computational methods are given in Sections 3 and 4, respectively. Section 5 presents a simulation study to verify the performance of the proposed methods in terms of parameter recovery. The practicability of the GDCM is examined in Section 6 using real data from a well-known personality questionnaire. The overall model fit issue is also addressed in this section. Lastly, some discussions are provided in Section 7 on the suggested model and future research directions.
2 Graphical Diagnostic Classification Model

The present section elaborates on the GDCM framework after briefly introducing the DCM and the Markov network. Suppose a test consists of \( J \) items measuring \( K \) attributes, \( \alpha_1, \ldots, \alpha_K \). The cognitive status for each attribute is indicated by a binary value: 1 implying mastery of an attribute and 0 implying nonmastery of an attribute. Let \( X_j \) be a response variable for item \( j \), and \( x_j \) be a realization of \( X_j \). An examinee’s latent cognitive profile is denoted by a \( 2^K \)-dimensional binary vector, \( \alpha = (1, \alpha_1, \alpha_2, \ldots, \alpha_K, \alpha_1\alpha_2, \ldots, \alpha_1\alpha_2\cdots\alpha_K)^\top \). Likewise, statistical properties of item \( j \) are indicated by a vector of regression coefficients, \( \beta_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jK}, \beta_{j12}, \ldots, \beta_{j12\cdots K})^\top \), where \( \beta_{j0} \) is an intercept parameter; \( \beta_{j1}, \ldots, \beta_{jk} \) denote the main effects of the \( K \) attributes; \( \beta_{j12} \) shows an interaction effect of \( \alpha_1 \) and \( \alpha_2 \); and \( \beta_{j12\cdots K} \) shows an interaction effect of \( \alpha_1, \ldots, \alpha_K \).

The item response function of the DCM then admits the generalized linear form

\[
\text{logit Pr}(X_j = 1 | \alpha; \beta_j) = \beta_j^\top \alpha.
\] (1)

The current study assumes that the loading structure of test items on the latent attributes is properly implied by \( \beta_j \), for instance, via a \( J \)-by-\( K \) matrix \( Q = (q_j : j = 1, \ldots, J) \) (Tatsuoka, 1985). Statistical methods for estimating \( Q \) have been discussed in many papers; see, e.g., Y. Chen, Liu, Xu, and Ying (2015), Y. Chen, Liu, and Ying (2015), Chiu (2013), de la Torre (2008), and J. Liu, Xu, and Ying (2012, 2013).

To illustrate (1) via a specific model, let us consider the DINA model (Haertel, 1989; Junker & Sijtsma, 2001) as an example, one of the most popular and widely-studied models in the diagnostic modeling. The cognitive process underlying the DINA model is conjunctive in that, for responding an item correctly, an examinee needs to master all required attributes by the item. This results in two latent classes for each item: one with those who have mastered all desired attributes and the other with those who lack at least one of the measured attributes. Suppose item \( j \) requires \( 1, 2, \ldots, K_j \) attributes that, without loss of generality, can be recognized to the first \( K_j \) positions of a row vector \( q_j \). The item parameter \( \beta_j \) is then comprised of two nonzero values, \( \beta_{j0} \) and \( \beta_{j12\cdots K_j} \), and 0 otherwise. The intercept parameter \( \beta_{j0} \) characterizes the positive response probability for the baseline (nonmastery) group, and the interaction parameter \( \beta_{j12\cdots K_j} \) differentiates the probability
functions between the mastery and nonmastery groups. Expressly, an examinee with perfect mastery of all required attributes for item $j$ answers the item correctly with the probability

$$\Pr(X_j = 1 | \alpha; \beta_j) = 1 - s_j = \frac{\exp(\beta_{j0} + \beta_{j12} \ldots K_j)}{1 + \exp(\beta_{j0} + \beta_{j12} \ldots K_j)},$$

where $s_j = \Pr(X_j = 0 | \alpha_1 \alpha_2 \ldots \alpha_K = 1)$ is a slipping parameter indicating the probability of giving an incorrect answer when all the necessary attributes have been mastered. An examinee lacking at least one desired attribute has the correct response probability

$$\Pr(X_j = 1 | \alpha; \beta_j) = g_j = \frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})},$$

where $g_j = \Pr(X_j = 1 | \alpha_1 \alpha_2 \ldots \alpha_K = 0)$ is a guessing parameter implying the probability of responding to the item correctly when at least one required attribute has been missed. The two item parameters, $s_j$ and $g_j$, characterize the random noise entering the examinee’s test-taking process. Each parameter has a one-to-one correspondence with the parameters in the generalized linear form such that $\beta_{j12} \ldots K_j = \logit(1 - s_j) - \beta_{j0}$ and $\beta_{j0} = \logit(g_j)$.

Rewriting (1) as a function of $x_j$, we have

$$\Pr(X_j = x_j | \alpha; \beta_j) = \exp(x_j \beta_j^\top \alpha) \propto \exp(x_j \beta_j^\top \alpha).$$

If the local independence assumption holds, the conditional probability distribution of responses to $J$ items can be written as

$$f(X_1 = x_1, \ldots, X_J = x_J | \alpha; B) = \prod_{j=1}^{J} \frac{\exp(x_j \beta_j^\top \alpha)}{1 + \exp(\beta_j^\top \alpha)} \propto \exp(x^\top B \alpha),$$

where $B = (\beta_j^\top : j = 1, \ldots, J)$ is a $J$-by-$2^K$ matrix containing all regression vectors, and $x = (x_1, \ldots, x_J)^\top$. If the items are locally dependent, on the other hand, the conditional distribution of the item scores can no longer be written as a product of the individual probability functions. This is because there remains correlation among the items, implying the existence of unmodeled
confounding factors that predict the correct response probability other than $\alpha$. This problem constitutes a substantial challenge to the evaluation of the likelihood function and consequently to the estimation of the person and item parameters.

To resolve this issue, we introduce a Markov network and explicitly model conditional dependence relationships among the items. The Markov network is specified by an undirected graph $G = (V, E)$ with a vertex set $V = \{1, \ldots, J\}$, of which each element is associated with a random variable, and an edge set $E \subset V \times V$, which characterizes pairwise relationships of the variables. In the present setup, the random variables of interest are the item response scores, $X_1, \ldots, X_J$. The graph models a conditional (in)dependence relationship between two random variables via an edge. That is, if $X_j$ and $X_{j'}$ are conditionally independent given all the other variables, there is no edge between the vertices $j$ and $j'$ (i.e., $(j, j') \notin E$); if, on the other hand, the two variables are conditionally dependent, their associated vertices are linked by an edge $(j, j') \in E$. The graph is undirected in the sense that $(j, j') \in E$ if and only if $(j', j) \in E$. The relationships of all random variables may be implied by a symmetric design matrix $\Phi = (\phi_{jj'} : j, j' = 1, \ldots, J)$, where $\phi_{jj'} = \phi_{j'j} = 0$ if $(j, j') \notin E$, and $\phi_{jj'} = \phi_{j'j} \neq 0$ if $(j, j') \in E$. The joint probability distribution of the random variables then has the form

$$f(X_1 = x_1, \ldots, X_J = x_J) = \frac{1}{z(\Phi)} \exp \left( \frac{1}{2} x^\top \Phi x \right) \propto \exp \left( \frac{1}{2} x^\top \Phi x \right),$$  

(2)

where $z(\Phi) = \sum_{x \in \{0, 1\}^J} \exp \left( \frac{1}{2} x^\top \Phi x \right)$ is a normalizing constant that ensures the distribution sum to 1. The diagonal entry of the design matrix, $\phi_{jj}$, shows a main effect of the variable $X_j$; the off-diagonal entry, $\phi_{jj'}$, quantifies an interaction effect between the variables $X_j$ and $X_{j'}$. By convention, a vertex is not a neighbor of itself. Hence, the current study fixes all vertex-wise parameters $\phi_{jj}$s at 0 in like manner and assumes that any nonzero values of the diagonal entries are traceable to the DCM of concern.

Incorporating the Markov network into the DCM, we obtain the graphical DCM where the joint probability function of $X = (X_1, \ldots, X_J)$ is given by

$$f(X = x \mid \alpha, B, \Phi) \propto \exp \left( x^\top B \alpha + \frac{1}{2} x^\top \Phi x \right).$$  

(3)
Note that the GDCM itself is a Markov network. If we reparameterize $\Phi^* = \Phi^*(\alpha, B, \Phi)$, the GDCM can be rewritten as

$$f(X = x | \alpha, B, \Phi^*) \propto \exp\left(\frac{1}{2}x^\top \Phi^* x\right),$$

where $\Phi^*$ has $\phi_{jj'}$ if $j \neq j'$ and $2\beta_j \alpha$ if $j = j'$, admitting a Markov network. It is clear from this notion that the diagonal elements of the design matrix account for the effects of the DCM (i.e., the interactions between the latent class and the items), while the off-diagonal entries describe the remaining item dependencies that cannot be explained by the latent class.

Define a prior distribution of $\alpha$ such that the joint distribution of $(X, \alpha)$ has the form

$$f(X = x, \alpha | B, \Phi) \propto \pi_\alpha \exp\left(x^\top B \alpha + \frac{1}{2}x^\top \Phi x\right),$$

wherein the normalizing constant is given by

$$z(B, \Phi) = \sum_{x \in \{0,1\}^J} \sum_{\alpha \in \{0,1\}^K} \pi_\alpha \exp\left(x^\top B \alpha + \frac{1}{2}x^\top \Phi x\right).$$

It follows that the complete-data likelihood of the parameters $(B, \Phi)$ is expressed as

$$f(x, \alpha | B, \Phi) = \frac{\pi_\alpha}{z(B, \Phi)} \exp\left(x^\top B \alpha + \frac{1}{2}x^\top \Phi x\right). \quad (4)$$

Since the latent class is unobservable, inference about the item parameters is made via a marginal distribution $f(x | B, \Phi)$ by integrating out $\alpha$ from the joint probability distribution $f(x, \alpha | B, \Phi)$ with respect to $\pi_\alpha$. Evaluating $f(x, \alpha | B, \Phi)$, however, is computationally challenging because the normalizing constant requires computation of $2^J \times 2^K$ terms, making it almost infeasible to evaluate the joint probability function for even realistic sizes of $J$ and $K$. A viable alternative is to approximate the joint likelihood function through a surrogate version, for example, most notably the pseudo-likelihood (Besag, 1975) or the composite likelihood (Cox & Reid, 2004; Lindsay, 1988). The current study employs the pseudo-likelihood approach that has been standard practice for estimating the Markov network. The ensuing sections discuss the relevant parameter estimation scheme and the computational methods.
3 Parameter Estimation

Let $\theta = (B, \Phi, \pi)$ denote the parameters of interest where $\pi = (\pi_\alpha : \alpha \in \{0, 1\}^K)$. The current study does not impose any particular form on $\pi_\alpha$; instead, it regards $\pi$ as one of the model parameters to be estimated together with $B$ and $\Phi$. Let $\theta^{(t)} = (B^{(t)}, \Phi^{(t)}, \pi^{(t)})$ be the provisional estimate of $\theta$ obtained at the $t$-th iteration. The likelihood function of the unknown parameters is obtained by integrating out the nuisance parameter from the joint probability distribution $f(x, \alpha | \theta)$. The corresponding objective function is expressed as

$$Q(\theta | \theta^{(t)}) = E \left[ \log f(x, \alpha | \theta) \mid x, \theta^{(t)} \right],$$

where $x$ is the observation from a single subject. The expectation of the complete-data log-likelihood with respect to the conditional distribution of $\alpha$ given $x$ is obtained as

$$E \left[ \log \mathcal{L}(\theta | x, \alpha) \mid x, \theta^{(t)} \right] = \sum_\alpha f(\alpha | x, \theta^{(t)}) \log f(x, \alpha | \theta), \quad (5)$$

where $\mathcal{L}(\theta | x, \alpha)$ is the complete-data likelihood for a single subject. The support set in the summation over $\alpha$ runs through $\{0, 1\}^K$ and is omitted for notational simplicity. As alluded to above, evaluating $f(x, \alpha | \theta)$ is computationally demanding due to the intractable normalizing constant. Hence, we consider the pseudo-likelihood function for optimization:

$$\mathcal{L}(\theta | x, \alpha) \triangleq \prod_{j=1}^J f(x_j | x_{-j}, \alpha, \theta) = \prod_{j=1}^J f(x_j | x_{-j}, \alpha, \theta) f(\alpha | x_{-j}, \theta), \quad (6)$$

where

$$f(x_j | x_{-j}, \alpha, \theta) = \frac{\exp \left( x_j \beta_j^\top \alpha + \sum_{j' \neq j} \phi_{jj'} x_j x_{j'} \right)}{1 + \exp \left( \beta_j^\top \alpha + \sum_{j' \neq j} \phi_{jj'} x_j x_{j'} \right)},$$

and

$$f(\alpha | x_{-j}, \theta) = \frac{\pi_\alpha \exp \left( \sum_{j' \neq j} x_{j'} \beta_{j'}^\top \alpha \right) \left( 1 + \exp \left( \beta_j^\top \alpha + \sum_{j' \neq j} \phi_{jj'} x_{j'} \right) \right)}{\sum_\alpha \pi_\alpha \exp \left( \sum_{j' \neq j} x_{j'} \beta_{j'}^\top \alpha \right) \left( 1 + \exp \left( \beta_j^\top \alpha + \sum_{j' \neq j} \phi_{jj'} x_{j'} \right) \right)}.$$
Taking all observations from a sample into account, the objective function can be expressed as

$$Q(\theta | \theta^{(t)}) = \sum_{i=1}^{N} \sum_{\alpha} f(\alpha | x_i, \theta^{(t)}) \sum_{j=1}^{J} \left[ \log f(x_{ij} | x_{i,j-}, \alpha, B, \Phi) + \log f(\alpha | x_{i,j-}, B, \pi_{\alpha}) \right],$$

where $x_{i,j-} = (x_{i1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{ij})$. The optimization problem is then given by

$$\left( \hat{B}, \hat{\Phi}, \hat{\pi} \right) = \arg \max_{(B, \Phi, \pi)} Q(\theta | \theta^{(t)}). \quad (7)$$

It can be shown that, after algebraic simplification, the $Q$-function has the form

$$Q(\theta | \theta^{(t)}) = \sum_{i=1}^{N} \sum_{\alpha} f(\alpha | x_i, \theta^{(t)}) \sum_{j=1}^{J} \left[ \log \pi_{\alpha} x_i^\top B \alpha + \frac{1}{2} x_i^\top \Phi x_i 
- \log \left\{ \sum_{\alpha} \pi_{\alpha} \exp \left( \sum_{j' \neq j} x_{ij'} \beta_{j'}^\top \alpha \right) \left( 1 + \exp \left( \beta_j^\top \alpha + \sum_{j' \neq j} \phi_{jj'} x_{ij'} \right) \right) \right\} \right].$$

The posterior probability for an individual being classified into $\alpha$ given the response profile $x_i$ is obtained by

$$f(\alpha | x_i, \theta^{(t)}) = f(\alpha | x_i, B^{(t)}, \pi_{\alpha}^{(t)}) = \frac{\pi_{\alpha}^{(t)} \exp(x_i^\top B^{(t)} \alpha)}{\sum_{\alpha} \pi_{\alpha}^{(t)} \exp(x_i^\top B^{(t)} \alpha)}.$$

4 Computation

The present study employs coordinate descent algorithm to solve the optimization problem (7). The algorithm minimizes the multivariate function $-Q(\theta | \theta^{(t)})$ with respect to each coordinate direction at a time and solves a sequence of univariate optimization subproblems in a loop. The procedure is simple to implement and computationally advantageous when the gradient of a function is expensive to evaluate. When it comes to estimating the design matrix, the log pseudo-likelihood becomes a concave function, and thus, standard convex programming can be used to estimate $\Phi$. This study employs the Newton’s algorithm to minimize the quadratic approximation of $\log f(x_{ij} | x_{i,j-}, \alpha, B, \Phi)$, similarly to solving the lasso problem (e.g., Friedman, Hastie, Höfling, & Tibshirani, 2007; Friedman, Hastie, & Tibshirani, 2010).

Let $P_{ij}^{(t)}(\alpha)$ denote the conditional probability of answering item $j$ correctly for examinee $i$ with
latent class $\alpha$ given the current item parameter estimates and the rest score:

$$P_{ij}(\alpha) = \frac{\exp \left( \beta_j^{(t)} \alpha + \sum_{j' \neq j} \phi_{j'j} x_{ij'} \right)}{1 + \exp \left( \beta_j^{(t)} \alpha + \sum_{j' \neq j} \phi_{j'j} x_{ij'} \right)}.$$ 

The $\beta_j^{(t)}$ and $\phi_{j'j}$ are the provisional estimates of $\beta_j$ and $\phi_{j'j}$ obtained at the $t$-th iteration. The local quadratic approximation to $\log f(x_{ij} \mid x_{i-,j}, \alpha, B, \Phi)$ is obtained as

$$\log f(x_{ij} \mid x_{i-,j}, \alpha, B, \Phi) \approx -\frac{1}{2} \omega_{ij}^{(t)} \left( \beta_j^{(t)} \alpha + \sum_{j' \neq j} \phi_{j'j} x_{ij'} - y_{ij}^{(t)} \right)^2,$$

where

$$\omega_{ij}^{(t)} = P_{ij}(\alpha) \left( 1 - P_{ij}(\alpha) \right),$$

and

$$y_{ij}^{(t)} = \beta_j^{(t)} \alpha + \sum_{j' \neq j} \phi_{j'j} x_{ij'} + \left( x_{ij} - P_{ij}(\alpha) \right) \omega_{ij}^{(t)} \omega_{ij}^{(t)} - 1.$$

Recall that the key assumption of the GDCM is that the latent class accounts for most variation of response data and only a fraction of variance is attributed to interdependency of items. This assumption encourages sparsity and simplicity in the design matrix such that many of the coefficients in $\Phi$ are close to zero and a small subset of the coefficients have larger and nonzero values. Common practice for enforcing sparsity on the Markov network is to impose $L_1$ regularization on the log pseudo-likelihood function (Höfling & Tibshirani, 2009), and this approach is employed in the current study. Let $\lambda$ be a regularization parameter that controls the sparsity (or complexity) of $\Phi$. The objective function to be optimized with respect to $\Phi$ is then given by

$$\arg \max_{\Phi} \sum_{i=1}^{N} \sum_{\alpha} f(\alpha \mid x_i, \theta(t)) \sum_{j=1}^{J} -\frac{1}{2} \omega_{ij}^{(t)} \left( \beta_j^{(t)} \alpha + \sum_{j' \neq j} \phi_{j'j} x_{ij'} - y_{ij}^{(t)} \right)^2 - N \sum_{1 \leq j < j' \leq J} \lambda |\phi_{jj'}|$$

subject to symmetry $\phi_{jj'} = \phi_{j'j} (1 \leq j < j' \leq J)$. The $L_1$ penalty in (9) performs both the variable selection and regularization on the parameters in that it induces a solution with many zero
coefficients and discourages the coefficients from having extreme values in $\Phi$. The same value of $\lambda$ for all $\phi_{jj'}$'s implies that the same degree of sparsity is enforced across all pairs of the items. In practice it is possible to impose different levels of penalties on the respective $\phi_{jj'}$'s—for example, when the prior knowledge about the item relationships is available due to the testlet design. Also note that no penalty is imposed on the diagonal elements in the above expression. This is because it has been assumed that $\phi_{jj} = 0$ for all $j = 1, \ldots, J$ and that any nonzero value can be inferred from the DCM. Even if $\phi_{jj}$ were not assumed to be zero, the diagonal entries of the design matrix should not be penalized because they impart the effects of the DCM of concern. In this case constraints on the parameters are needed to ensure identifiability of the model.

Simple calculus (Donoho & Johnstone., 1994) suggests that the solution to the problem (9) has a closed-form expression. A coordinate-wise update for $\phi_{jj'}$ is obtained as

$$
\phi_{jj'}^{(t+1)} \leftarrow S \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{\alpha} f(\alpha \mid x_i, \theta^{(t)}_{ij} \omega_{ij}^{(t)} x_{ij'} r_{ij'}, \lambda) }{ \frac{1}{N} \sum_{i=1}^{N} \sum_{\alpha} f(\alpha \mid x_i, \theta^{(t)}_{ij} \omega_{ij}^{(t)} x_{ij'}^2) } , \right),
$$

where $S(z, \lambda)$ is the soft-thresholding operator with the value

$$
S(z, \lambda) = \text{sign}(z)(|z| - \lambda)_+ = \begin{cases} 
  z - \lambda & \text{if } z > 0 \text{ and } \lambda < |z| \\
  z + \lambda & \text{if } z < 0 \text{ and } \lambda < |z| \\
  0 & \text{if } \lambda \geq |z|,
\end{cases}
$$

and $r_{ij'} = y_{ij}^{(t)} - y_{ij',-j'}^{(t)}$ is the partial residual for fitting $\phi_{jj'}$.

Once the estimate $\Phi^{(t+1)}$ is obtained from the above procedure, $B^{(t+1)}$ and $\pi^{(t+1)}$ are obtained sequentially via the coordinate descent method. The $B$-matrix is updated with a maximizer of the function

$$
\sum_{i=1}^{N} \sum_{\alpha} f(\alpha \mid x_i, \theta^{(t)}_{ij} \sum_{j=1}^{J} \log f(x_{ij} \mid x_i, -j, \alpha, B, \Phi).
$$

The optimization problem is solved for each parameter separately, holding all other parameters fixed at the latest values. It should be pointed out that the maximization in (11) is enforced only
on the log pseudo-likelihood term because the log posterior term provides little information on $B$ while introducing much expensive computation. The following simulation study suggests that maximizing (11) does not induce much bias in $\beta$ estimation. The optimization problem for $\pi^{(t+1)}$ is solved by maximizing

$$Q(\theta | \theta^{(t)}) \approx \sum_{i=1}^{N} \sum_{\alpha} f(\alpha | x_i, \theta^{(t)}) \sum_{j=1}^{J} \log f(\alpha | x_i, -j, B, \pi_{\alpha}),$$

(12)

where $B$ is evaluated at the most current estimate $B^{(t+1)}$. The iteration continues until no improvement is made along all the coordinate directions and the estimated coefficients are stabilized.

In (10) the performance of soft-thresholding relies on the choice of a tuning parameter. The criteria for selecting $\lambda$ are commonly classified into two categories depending on the resulting statistical properties: consistency and efficiency. The general consensus from the prior research suggests that the Bayesian information criterion (BIC; Schwarz, 1978) leads to consistent variable selection and the Akaike information criterion (AIC; Akaike, 1974) and cross-validation (Craven & Wahba, 1979) methods are related to asymptotically loss-efficient selection of tuning parameters (Yang, 2005; Zhang, Li, & Tsai, 2010). In the present work we employ the BIC in the interest of consistent solution of the design matrix:

$$\lambda = \arg \min \text{ BIC}(M_\lambda),$$

(13)

where $M_\lambda$ is the model fitted under the specific tuning parameter $\lambda$. The value of the BIC is calculated as

$$\text{BIC}(M_\lambda) = -2 \log L_\lambda + |M_\lambda| \log N,$$

The number of parameters being estimated for the DCM is subject to the number of nonzero values in $Q$. The DINA model, for example, has two nonzero parameters for each $\beta_j$ and thus results in $2 \times J$ free parameters. The value of $|M_\lambda|$ in this case is $2J + \sum_{1 \leq j < j' \leq J} 1_{\hat{\phi}_{j,j'} \neq 0}$.
Figure 1: Graphical Representation of Simulated $\Phi$
5 Simulation Study

5.1 Design

In this section we provide a simulation study to validate the performance of the GDCM and the corresponding estimation method under known parameters. The simulation was carried out in a $2 \times 3 \times 3$ factorial design by varying the number of attributes ($K=3, 4$), graphical structure (null, pairs, triplets), and sample size ($N=500, 1000, 3000$). The test length was fixed at $J = 30$. The primary purpose of the simulation study is to evaluate parameter recovery of the GDCM under varying degrees of local item dependence and compare the results against those obtained from the standard DCM approach.

The three types of graphical structure considered are illustrated in Figure 1. The pair condition represents when items have pairwise relationships. The items are grouped in pairs such that there exist edges between items $\{1, 2\}$, $\{3, 4\}$, $\ldots$, and $\{29, 30\}$. This results in a $\Phi$-matrix that has nonzero values in $\phi_{j,j+1}$ for $j=1, 3, \ldots, 29$; and 0 otherwise. The triplet condition indicates a situation in which items are grouped in triples, $\{1, 2, 3\}$, $\{4, 5, 6\}$, $\ldots$, $\{28, 29, 30\}$. There exist edges between every pair of the items within a triplet. Accordingly, the $\Phi$-matrix has nonzero values in $\phi_{j,j+1}$, $\phi_{j,j+2}$, and $\phi_{j+1,j+2}$ for $j = 1, 4, \ldots, 28$; and 0 otherwise. Defining the sparsity as the proportion of independent edges in a graph, the three simulated graphs correspond to the sparsity level of 100%, 96.55%, and 93.10%, respectively. These settings amount to zero, 15, and 30 pairs of nonzero edges in $\Phi$. For each of the graphs, we set edge-wise parameters at a positive constant 1 if the graph has an edge between the variables $X_j$ and $X_j'$ and set all the node-wise parameters at 0.

For simulating response data, the study considered the DINA model. Note that the proposed GDCM framework and the estimation method can accommodate any restricted DCMs in a generalized form and alternative DCMs can also be used as an underlying model. The item parameters for the DINA model were randomly sampled from uniform distributions with endpoints $(0.05, 0.2)$ and $(0, 0.2)$ for guessing ($g$) and slipping ($s$), respectively. Examinees’ attribute profiles were drawn from a multinominal distribution with the success probability 0.5 across all latent dimensions. For indicating item loadings on each latent dimension, a $J$-by-$K$ matrix $Q = (q_{j} : j = 1, \ldots, J)$ was employed. Each entry of $q_{j}$ indicates whether or not the $k$-th attribute is required for item success.
The Q-matrix was designed such that there exist at least three items measuring unique attribute or else the row vectors were randomly sampled from \( \{0, 1\}^K \setminus \mathbf{0} \). Every simulation condition was implemented with a total of 100 repetitions consisting of unique sets of item and person parameters. Given the parameters generated, Gibbs sampler was implemented to obtain binary item responses. Let \( x_1^{(t)}, \ldots, x_J^{(t)} \) denote a set of draws available from the \( t \)-th iteration for a simulee. A response in the next iteration, \( x_j^{(t+1)} \), is then drawn from the Bernoulli distribution with the success probability

\[
P_{j}^{(t+1)} = \frac{\exp \left( \beta_j^\top \alpha + \sum_{j' \neq j} \phi_{jj'} x_{j'}^{(t)} \right)}{1 + \exp \left( \beta_j^\top \alpha + \sum_{j' \neq j} \phi_{jj'} x_{j'}^{(t)} \right)}.
\]

To ensure that the sample observations come from the appropriate stationary distribution, the study collected samples from a chain after 300 burn-ins. The generated response data were then calibrated as presented in Sections 3 and 4, and the resulting item parameters were transformed to the DINA parameters as discussed in Section 2.

### 5.2 Evaluation Criteria

The accuracy of the estimated item parameters was evaluated via root mean square deviation (RMSD) and biasedness from the true values. The RMSD for the guessing parameter was calculated as the square root of the mean square error \( E[(\hat{g} - g^*)^2] \), where \( \hat{g} \) is the estimated value and \( g^* \) is the generating parameter value. The biasedness was computed as the absolute deviation of \( \hat{g} \) from \( g^* \) in order to avoid the cancellation between positive and negative biases. The RMSDs and absolute biases for the slipping parameter were calculated likewise. A similar criterion was used to evaluate the closeness of the design matrix. In addition to the RMSD as a distance measure, the false positive rate (FPR) and correct positive rate (CPR) were considered as follows:

\[
FPR = \frac{\left| \{(j, j') : j < j', \hat{\phi}_{jj'} \neq 0 \text{ and } \phi_{jj'}^* = 0 \} \right|}{\left| \{(j, j') : j < j', \phi_{jj'}^* = 0 \} \right|},
\]
and

\[
\text{CPR} = \frac{ \{ (j, j') : j < j', \hat{\phi}_{jj'} \neq 0 \text{ and } \phi_{jj'}^{*} \neq 0 \} }{ \{ (j, j') : j < j', \phi_{jj'}^{*} \neq 0 \} },
\]

where \( \hat{\phi}_{jj'} \) is the estimated value for the true parameter \( \phi_{jj'}^{*} \). If the underlying graph is well recovered, it is expected that the FPR is close to 0 and the CPR is close to 1. Finally, the recovery of \( \pi \) was evaluated by the pairwise Euclidean distance from the true value.

5.3 Results

The RMSDs of the item parameter estimates of the DINA model are summarized in Table 1. Each table entry is the averaged result across the replications. Table 1 suggests that the generating item parameters were overall well recovered via the GDCM framework. Across the simulation conditions established, the GDCM maintained the small RMSDs for both the item parameter types. Although the estimation accuracy tended to degenerate as the test set included locally dependent items, the general performance was found more accurate than that of the standard DCM. Speaking concretely, under the null case, the two modeling approaches seemed to show comparable performances in estimating the item parameters. As some of the test items were interrelated, however, the GDCM

| Graph | \( K = 3 \) | \( K = 4 \) |
|-------|-----------|-----------|
|       | GDCM      | DCM       | GDCM      | DCM       |
|       | Guess     | Slip      | Guess     | Slip      | Guess     | Slip      | Guess     | Slip      |
| Null  | 500       | .022      | .037      | .027      | .024      | .022      | .038      | .032      | .027      |
|       | 1000      | .015      | .022      | .019      | .017      | .015      | .024      | .022      | .019      |
|       | 3000      | .008      | .012      | .011      | .010      | .009      | .013      | .012      | .011      |
| Pair  | 500       | .036      | .042      | .062      | .056      | .038      | .045      | .070      | .057      |
|       | 1000      | .037      | .037      | .059      | .054      | .036      | .036      | .065      | .055      |
|       | 3000      | .036      | .037      | .056      | .053      | .036      | .033      | .062      | .053      |
| Triplet | 500   | .068      | .055      | .155      | .086      | .068      | .062      | .168      | .087      |
|       | 1000      | .062      | .047      | .153      | .085      | .063      | .051      | .166      | .086      |
|       | 3000      | .060      | .043      | .152      | .085      | .061      | .041      | .165      | .085      |

Note. RMSD= root mean square deviation. \( K = \) number of latent attributes being measured. GDCM= graphical DCM. \( N = \) sample size. The null graph includes no edge, the pair graph includes 15 edges, and the triplet graph includes 30 edges.
Table 2: Absolute Biases of Guessing and Slipping Parameter Estimates

| Graph | $N$  | GDCM Guess | GDCM Slip | DCM Guess | DCM Slip | GDCM Guess | GDCM Slip | DCM Guess | DCM Slip |
|-------|------|------------|-----------|-----------|----------|------------|-----------|-----------|----------|
|      | 500  | .017       | .027      | .020      | .019     | .018       | .028      | .024      | .020     |
|      | 1000 | .012       | .016      | .015      | .013     | .012       | .018      | .016      | .014     |
|      | 3000 | .007       | .009      | .009      | .007     | .007       | .010      | .009      | .008     |
| Pair |      |            |           |           |          |            |           |           |          |
|      | 500  | .028       | .032      | .050      | .047     | .029       | .033      | .058      | .047     |
|      | 1000 | .029       | .029      | .047      | .046     | .028       | .028      | .054      | .047     |
|      | 3000 | .030       | .030      | .045      | .046     | .029       | .026      | .051      | .046     |
| Triplet |      |            |           |           |          |            |           |           |          |
|      | 500  | .054       | .043      | .132      | .075     | .053       | .048      | .144      | .076     |
|      | 1000 | .051       | .037      | .131      | .075     | .050       | .038      | .143      | .076     |
|      | 3000 | .051       | .034      | .131      | .075     | .049       | .032      | .143      | .076     |

consistently demonstrated the smaller estimation errors in retrieving the true parameters. The improvement of the GDCM over the DCM estimates was more salient in the guessing parameters than in the slipping parameters. The pattern of the improvement became pronounced as there existed more dependent items in the test. In Table 1 some anticipated observations may deserve some comments. The table shows that the larger the sample size, the more precisely the parameters were recovered. The more complex the underlying dimensions, it became more difficult to estimate the true parameters.

Presented in Table 2 are absolute biases of the item parameter estimates. The trends in the results were largely consistent with those reported for Table 1. In the null case of no local dependency, the two modeling approaches showed marginal differences with respect to estimation bias. As some of the test items were interdependent, the two approaches exhibited increasing biases. The GDCM notwithstanding constantly showed the smaller degree of bias across all conditions evaluated. The direction of the bias in the estimated parameter values was by and large in concordance with each other. In both the modeling approaches, the guessing parameter estimates typically showed positive biases, whereas the slipping parameter estimates mostly showed negative biases. These patterns suggest that the positive association between the items (i.e., $\phi_{jj'} = 1$) results in increase in guessing and decrease in slipping, and thus, decrease in item discriminating power.

Table 3 reports results of retrieving the graphical model parameters. The results suggest that true design matrices were quite well recovered by the GDCM. The observed RMSDs were small;
the CPRs were reasonably high while the FPRs were well under control. We note that the overall level of the RMSD observed in Table 3 is greater than that of Table 1 because, in contrast to the guessing and slipping parameters whose values are bounded by \([0, 1]\), the value of \(\phi_{jj'}\) is defined on a continuum with no bounds. Table 3 indicates that, despite the wider support, the suggested estimation method can recover the true \(\phi_{jj'}\) values with the distance less than 0.2 units. Also presented in Table 3 are the FPRs calculated as the proportion of edges that were erroneously flagged for conditional dependence. The table shows that the overall false alarm rates remained at an acceptable level, ranging from 2.84% to 5.82% on average. The CPR in Table 3 denotes the proportion of nonzero edges that were correctly identified for a given design matrix. Note that a direct comparison of FPRs or CPRs across the different \(\Phi\) cases lends little meaning because the number of (non)zero edges varies in each simulated condition. On the whole, it can be concluded that the GDCM served well the purpose of identifying the presence of local item dependence. Applying the cutoff 0.8 (Cohen, 1992), the framework demonstrated moderate recovery when the sample size was 500 and showed excellent recovery as the sample size was greater than 1000. In Table 3 the sample size seemed to be the most influential factor affecting the recovery of \(\Phi\). The trends in the results indicate that the larger the sample, the smaller the RMSDs and FPRs, and the greater the CPRs.

Table 4 presents the average pairwise distance of the estimated \(\pi_s\) from the true values. It suggests that the proposed algorithm recovered the underlying prior distributions with minimal errors. Across the simulation conditions, the smallest estimation error was associated with the smaller \(K\), the largest \(N\), and the simplest \(\Phi\). This pattern is in line with the expectation that \(\pi\) is

| Graph | \(N\) | \(K = 3\)       | \(K = 4\)       |
|-------|-----|-----------------|-----------------|
|       | RMSD | FPR  | CPR   | RMSD | FPR  | CPR   |
| Pair  | 500  | .132 | .058  | .759 | .131 | .056  | .733  |
|       | 1000 | .120 | .056  | .873 | .116 | .053  | .860  |
|       | 3000 | .115 | .054  | .948 | .109 | .051  | .941  |
| Triplet | 500 | .169 | .040  | .731 | .161 | .040  | .762  |
|        | 1000 | .145 | .034  | .844 | .140 | .036  | .856  |
|        | 3000 | .132 | .028  | .901 | .128 | .034  | .909  |

**Note.** FPR=false positive rate. CPR=correct positive rate.
more precisely recovered as data structure becomes less complicated and as more observations are available. Table 4 also indicates that the complexity of the design matrix had an appreciable impact on identifying the prior distributions. Compared to the null and pair cases, the design matrix with the triplets of locally dependent items resulted in substantially larger distances between the true and estimated $\pi$s. This implies that, while the existence of a small subset of dependent items does not significantly degrade the prediction of $\pi$, the larger number of interrelated items or the greater complexity in item network can have a severe impact on $\pi$ recovery, and in turn, examinee classification using the prior information.

In brief, the simulation study supports several conclusions. First, the GDCM and the pseudo-likelihood estimator can recover the model parameters with reasonable accuracy in the presence of local item dependence. The suggested methods are found particularly effective in regulating the biasedness of item parameter estimates. Second, the estimator presents a reliable projection of the design matrix, thereby allowing for valid information in regards to the existence and degree of item interactions. The following section shows how this information can be used for drawing meaningful inference about test items through an empirical data analysis. Third, the algorithm provides an accurate approximation of the prior distribution despite the unknown latent class relationship. The proposed procedure can therefore be used when the prior distribution needs to be empirically determined during the parameter estimation.

6 Real Data Example

6.1 Data

This section reports an application of the GDCM to real data collected from a well-known personality questionnaire. The data set consists of responses of 824 females to Eysenck Personality

| Graph | $K = 3$ | $K = 4$ |
|-------|--------|--------|
|       | $N = 500$ | $N = 1000$ | $N = 3000$ | $N = 500$ | $N = 1000$ | $N = 3000$ |
| Null  | .027   | .017   | .009   | .040   | .027   | .016   |
| Pair  | .024   | .016   | .008   | .040   | .028   | .017   |
| Triplet | .043   | .038   | .034   | .064   | .056   | .051   |

Table 4: Recovery of Prior Distribution
Questionnaire (EPQ)-Revised (EPQ-R; S. B. G. Eysenck, Eysenck, & Barrett, 1985), a widely used survey for measuring individual differences in personality across different cultures. The questionnaire assesses personality traits of an individual in aspects of Psychoticism, Extraversion, Neuroticism, and Lie. Psychoticism (P) implies an inclination toward being impersonal, aggressive, lacking in empathy, and incompliant with rules. Extraversion (E) is characterized by sociability, activity, and impulsiveness. Neuroticism (N) is associated with high levels of negative perception such as being nervous, apprehensive, and highly-strung. Lie (L) is related to faking good of scores on the other scales.

The question inventory has 100 yes/no items in its full-scale version (H. J. Eysenck & Eysenck, 1975) and 48 items in its short version (S. B. G. Eysenck et al., 1985). The present study focuses on a test set of 36 items from the short scale version, excluding the L scale that measures consistency in item responses. Each of the personality dimensions was measured on a continuous scale by 12 items. In this particular analysis, we aimed to classify observed response patterns in terms of a small number of latent classes by defining each scale on a binary status. Each status indicates whether or not a subject possesses a propensity for a particular scale. A similar strategy has been practiced in the DCM literature, for example, for diagnosing pathological gamblers (Templin & Henson, 2006) or evaluating language proficiency (Jang, 2009; Roussos, DiBello, Henson, Jang, & Templin, 2010).

6.2 Goodness-of-fit

To evaluate the fit of the GDCM, we compared the observed log-likelihood to an empirical distribution obtained via parametric bootstrapping under the presumed model. The detailed procedure is described as follows. Let \((\hat{\beta}, \hat{\Phi}, \hat{\pi})\) denote the maximal pseudo-likelihood estimates of the fitted model. The logarithm of the joint likelihood given the estimated parameters is then calculated as

\[
l_0 = \log \prod_{i=1}^{N} f(x_i | \hat{\beta}, \hat{\Phi}, \hat{\pi}) = \sum_{i=1}^{N} \log \left( \sum_{\alpha} \hat{\pi}_{\alpha} \exp \left( x_i^\top \hat{\beta} \alpha + \frac{1}{2} x_i^\top \hat{\Phi} x_i \right) \right) - \log z(\hat{\beta}, \hat{\Phi}),
\]

where \((x_1, \ldots, x_N)\) is the set of observed responses. The plausibleness of the observed log-likelihood value, \(l_0\), can be evaluated by comparing it against an empirical distribution of log-likelihoods predicted by the fitted model. In this study 500 independent bootstrap data sets were
used to establish the reference distribution. Each data set contained responses generated via Gibbs sampler under the same dimension with the observed one. Let \((x_1^b, \ldots, x_N^b)\) denote the data set obtained from the \(b\)-th bootstrapping. The log of the joint likelihood for the bootstrap data is computed as

\[
l^b = \log \prod_{i=1}^{N} f(x_i^b | \hat{\mathbf{B}}, \hat{\Phi}, \hat{\pi}) = \sum_{i=1}^{N} \log \left( \sum_{\alpha} \hat{\pi}_\alpha \exp \left( x_i^b \hat{\mathbf{B}}\alpha + \frac{1}{2} x_i^b \hat{\Phi} x_i^b \right) \right) - \log z(\hat{\mathbf{B}}, \hat{\Phi}).
\]

The last term, \(\log z(\hat{\mathbf{B}}, \hat{\Phi})\), in both the \(l^o\) and \(l^b\) expressions is fixed at a constant and hence can be ignored when comparing the log-likelihoods across the resamples. As such, the present study reports the unnormalized log-likelihood values by discounting the last term. For comparison purpose, the standard DCM without the graphical component was also considered and fitted to the observed data. Applying the same procedure of parametric bootstrapping, the empirical distribution of \((l^1, \ldots, l^{500})\) was obtained under the DCM and used as a reference for evaluating the plausibility of the model.

![Parameter Bootstrap for Checking Model Fit](image)

Figure 2: Parametric Bootstrap for Checking Model Fit
Note. The items from 1 to 12 correspond to the P scale; those from 12 to 24 belong to the E scale; and the rest is associated with the N scale. The color of the heat map shows intensity of an item relation in absolute value.

6.3 Results

In Figure 2 relative locations of the observed log-likelihood values are indicated in histograms of the empirical distributions under each GDCM and DCM. The results suggest that the GDCM fit the given data fairly well whereas the DCM tended to suffer from lack of model fit due conceivably to local item dependence. Specifically, the (unnormalized) log-likelihood observed under the GDCM was 8605.41, corresponding to a p-value 0.792 in the empirical distribution. The DCM displayed the log-likelihood 4169.73 with a p-value 0.012. Recall that the main difference between the two models is the presence of a Markov network, which is designed to account for local item dependence. The substantial improvement in the model fit implies that some of the questionnaire items may interact with each other beyond the latent variable. After including the Markov network, however, such item dependency seemed to be adequately taken into account while the psychometric model remained the same.

To provide a graphical representation of the item interactions, a heat map is plotted and presented in Figure 3. The heat map portrays pairwise associations of the items by indicating the
intensity of relationship via color. In Figure 3 there are a total of 36 nonzero edges with the sparsity level amounting to 94.29%. The darkness of the color represents the intensity of the item association. The heat map suggests that the degree of item interdependence varies depending on the items and the personality scale being assessed. The figure indicates that the items within the same scale were more likely to interact; although a few in number, there were also items interrelated across the scales. Overall, the P scale appeared to have the most dependent item pairs. The strongest interactions were observed in the E and N scales.

To lend some meaning to the locally dependent items and their corresponding estimated $\phi_{jj'}$ values, we further examined the specific item pairs with conditional dependence. Given in Table 5 are item pairs with the greatest positive dependency. The table suggests that the item pairs tended to share common stimuli or similar wording. The first three pairs of the items, for instance, were coupled by the words, ‘party,’ ‘nervousness,’ and ‘people,’ respectively. The other item pairs seemed to have commonality in that they ask the same particular aspect of the intended scale. For example, the items 1 and 11 seemed to be concerned with recklessness, one of the characteristics in the P scale. The items 1, 3, and 8 may be seen related in light of social non-conformity. The following item pairs pertained to the same aspects of the E scale, sociableness and talkativeness, respectively.

Notice that in Table 5 some items appear multiple times in the pairs, implying that the questions can be related through a bundle that consists of more than two items (e.g., items 1, 11, and 8). Table 6 lists those items connected via a clique. A clique is defined as a subset of items in which every pair of items in the subset is linked by an edge. Our analysis indicates that the maximal clique underlying the EPQ-R data consists of three items. Table 6 summarizes those observed with positive dependence. Table 6 suggests that items in the same personality scale were more likely to be interrelated with each other and the items within the same clique were associated with the certain aspect of the personality scale. The items in the first clique appeared to assess characteristics related to recklessness and inconsideration of an individual. The following clique seemed to measure a subject’s inclination toward sociableness and activeness in the E scale. The item clique corresponding to the N scale seemed to measure emotional unstableness of an individual.

The item pairs and cliques found above carry important implications for practice. The results suggest that some of the items in the questionnaire are over-related and may even be redundant.
### Table 5: Item Pairs with the Highest Positive Local Dependency

| $\hat{\phi}_{jj'}$ | Item | Scale | Question                                                                 |
|-------------------|------|-------|---------------------------------------------------------------------------|
| 2.696             | 18   | E     | Can you easily get some life into a rather dull party?                     |
|                   | 20   | E     | Can you get a party going?                                                |
| 1.755             | 31   | N     | Would you call yourself a nervous person?                                 |
|                   | 35   | N     | Do you suffer from ‘nerves’?                                              |
| 1.311             | 16   | E     | Do you enjoy meeting new people?                                           |
|                   | 19   | E     | Do you like mixing with people?                                           |
| 1.029             | 1    | P     | Would you take drugs which may have strange or dangerous effects?         |
|                   | 11   | P     | (R) Do you try not to be rude to people?                                  |
| 0.858             | 3    | P     | Do you think marriage is old-fashioned and should be done away with?      |
|                   | 8    | P     | (R) Do good manners and cleanliness matter much to you?                   |
| 0.726             | 1    | P     | Would you take drugs which may have strange or dangerous effects?         |
|                   | 8    | P     | (R) Do good manners and cleanliness matter much to you?                   |
| 0.713             | 15   | E     | Can you usually let yourself go and enjoy yourself at a lively party?     |
|                   | 19   | E     | Do you like mixing with people?                                           |
| 0.672             | 13   | E     | Are you a talkative person?                                               |
|                   | 24   | E     | (R) Are you mostly quiet when you are with other people?                  |

*Note.* (R) denotes reversed questions.

### Table 6: Cliques Consisting of Three Items with Positive Local Dependency

| $\sum \hat{\phi}_{jj'}$ | Item | Scale | Question                                                                 |
|--------------------------|------|-------|---------------------------------------------------------------------------|
| 2.131                    | 1    | P     | Would you take drugs which may have strange or dangerous effects?         |
|                          | 8    | P     | (R) Do good manners and cleanliness matter much to you?                   |
|                          | 11   | P     | (R) Do you try not to be rude to people?                                  |
| 1.504                    | 15   | E     | Can you usually let yourself go and enjoy yourself at a lively party?     |
|                          | 19   | E     | Do you like mixing with people?                                           |
|                          | 21   | E     | Do you like plenty of bustle and excitement around you?                  |
| 0.823                    | 25   | N     | Does your mood often go up and down?                                     |
|                          | 26   | N     | Do you ever feel ‘just miserable’ for no reason?                          |
|                          | 27   | N     | Are you an irritable person?                                              |

*Note.* The first column gives the sum of $\hat{\phi}_{jj'}$s ($j \neq j'$) in the clique. The item cliques had positive $\hat{\phi}_{jj'}$s for all edges.
From the psychometric perspective, it calls for corrective actions, such as revising questions, removing repeated items, or explicitly accounting for the degree of conditional item dependence, for the benefit of making valid inference about the item and person parameters.

7 Discussion

A major goal of this article has been to provide an analytical tool for identifying local item dependence and estimating the degree of such dependency within the DCM framework. The DCM is a statistical approach that models individual response patterns via a small number of latent classes. It determines whether an examinee possesses a particular set of attributes and thereby allows for individualized feedback on the examinee’s cognitive status. Because of its capacity to provide fine-grained information about examinee’s mastery of attributes, the DCM has been widely applied in educational, psychological, and clinical settings for supporting formative assessments (e.g., H. Liu, You, Wang, Ding, & Chang, 2013; Templin & Henson, 2006).

In DCMs statistical inference about the parameters hinges on a very restrictive assumption that item responses are locally independent within a latent class. The focus of the present article is to relax this assumption and provide a flexible modeling framework for analyzing data encountered in real applications. Specifically, the suggested model, namely the graphical DCM (GDCM), attempts to address surface local dependence in which items are interrelated due to shared features, such as common stimuli, similar wording, carry-over effects, etc. The GDCM is constructed by introducing a Markov network into the DCM such that it can describe pairwise associations of items that cannot be explicated by the latent class being modeled. The design matrix emerged from the Markov network also allows for a graphical representation of a network that underlies the test items. The numerical experiments using simulation suggest that the GDCM reduces bias in parameter estimation and gives an accurate summary of the conditional dependence relationships among the items. The analysis of empirical data indicates that when a test includes items interacting beyond the latent variable, the GDCM can not only improve the model fit but also offer a meaningful description of such dependency.

The GDCM and the estimation method proposed in this article can have several implications in practice. First, the GDCM is by far the most flexible framework for modeling the cognitive
process and item interaction simultaneously. The cognitive component is defined on the generalized form and thus can accommodate a wide range of restricted DCMs. The graphical model provides an explicit description of item relations without requiring prior knowledge about the underlying structure, thereby supporting exploratory investigation into a set of test items. In practice the information provided by the design matrix can serve as evidence of test validity or can be utilized at test development or assembly stages to improve the quality of assessments. Second, the present framework ensures interpretability of the model parameters by means of a sparse graph. The sparsity constraint imposed on the Markov network can further expedite the estimation of a graph in that the coordinate descent algorithm cycles through only a small subset of nonzero edges (most coordinates that are zero never become nonzero). Third, the GDCM, though aimed at surface local dependence, can also be used for examining underlying local item dependence that arises from testlets such as reading passage, graph or chart problems, or laboratory scenarios. This is because a testlet itself can be considered a common stimulus that invokes inter-item correlation.

While this is only an initial research effort to expand the DCMs to realistic settings, the implied results are promising enough to warrant further study. An important extension would be to explore the potential use of the GDCM in complex settings, such as missing data, stochastic test design, polytomous response data, and data with covariates. In particular, given the increasing popularity of performance task items and constructed response items, an extension of the GDCM to the polytomous response data seems worthy of follow-up research. A common strategy for calibrating polytomous data within the DCM framework has been to dichotomize item responses so that they can be analyzed using dichotomous models (e.g., Johnson et al., 2013; Su, 2013). The dichotomization of polytomous items, however, can result in loss of information. In preference future research could consider directly incorporating a Markov network into polytomous DCMs (e.g., de la Torre, 2010; Hansen, 2013; Templin, Henson, Rupp, Jang, & Ahmed, 2008; von Davier, 2008) in an analogous manner with the dichotomous case.

Second, the assumption of a known item-attribute loading matrix may be relaxed in the future. While the present study has assumed that the item loading matrix is correctly identified for the sake of simplicity of discussion, in practice a true \( \mathbf{Q} \)-matrix is hardly known, and it may be subject to fallible judgment. Provided that a misidentified \( \mathbf{Q} \)-matrix can have a deleterious effect on the recovery of model parameters (e.g., Rupp & Templin, 2008), additional work needs to be done in
regards to the behavior of the GDCM under modest misspecification of $Q$. Associated with this problem, it would be fruitful to consider statistical methods for a fully exploratory situation in which neither $\Phi$ nor $Q$ are known. The useful literature in this connection is Y. Chen, Liu, Xu, and Ying (2015), which explores regularized maximum likelihood estimation of $B$ in the absence of $Q$. A similar technique can be employed to simultaneously identify $Q$ and $\Phi$ during the model estimation.

Finally, the present GDCM assumes no particular distribution on the underlying attributes. If, however, latent attributes are structured with a specific form (e.g., Leighton, Gierl, & Hunka, 2004; Su, 2013; Templin, Henson, Templin, & Roussos, 2008; von Davier, 2010), it seems more reasonable to estimate the model parameters under such restriction. A natural question arises as to the robustness of the GDCM in this event. It would also be worthwhile to investigate how the assumed attribute structure would shape the resulting dependence relationships among the items.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician, 24*, 179–195.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581–612.

Chen, C.-T., & Wang, W.-C. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement, 31*, 388–411.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

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

Chen, Y., Liu, J., & Ying, Z. (2015). Online item calibration for Q-matrix in CD-CAT. *Applied Psychological Measurement, 39*, 5–15.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*, 598–618.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Cox, D. R., & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika, 91*, 729–737.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numer. Math.*, 31, 377–403.

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. (2010, July). The partial-credit DINA model. Athens, GA: Paper presented at the international meeting of the Psychometric Society.

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

Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika, 81*, 425–455.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the eysenck personality questionnaire*. London: Hodder & Stoughton.

Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences, 6*, 21–29.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.*, 2, 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Sofwr.*, 33, 1–22.

Gibbons, R., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423–436.

Glas, C. A., & Suárez Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87–106.
Haertel, E. H. (1989). Using restricted latent class models to map the attribute structure of achievement items. *Journal of Educational Measurement, 26*, 333–352.

Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis* (Unpublished doctoral dissertation). University of California at Los Angeles.

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

Höfling, H., & Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research, 10*, 883–906.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion model to LanguEdge assessment. *Language Testing, 26*, 31–73.

Johnson, M., Lee, Y.-S., Sachdeva, R. J., Zhang, J., Waldman, M., & Park, J. Y. (2013, April). Examination of gender differences using the multiple groups DINA model. San Francisco: Paper presented at the Annual Meeting of the National Council on Measurement in Education.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359–1378.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147–154.

Leighton, J. P., Gierl, M. J., & 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.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics, 80*, 220–239.

Liu, H., You, X., Wang, W., Ding, S., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification, 30*, 152–172.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*, 548–564.
Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli, 19*, 1790–1817.

Liu, Y., & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement, 36*, 670–688.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2n contingency tables. *Journal of the American Statistical Association, 100*, 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*, 713–732.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of a general theory. *Multivariate Behavioral Research, 14*, 21–38.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379–396.

Roussos, L., DiBello, L., Henson, R., Jang, E., & Templin, J. (2010). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 35–69). Washington, DC, US: American Psychological Association.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 19*, 461–464.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247.

Stout, W. F. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485–518.

Su, Y.-L. (2013). *Cognitive diagnostic analysis using hierarchically structured skills* (Unpublished doctoral dissertation). University of Iowa.

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

Templin, J. L., & Henson, R. A. (2006). Measurement of psychology disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.
Templin, J. L., Henson, R. A., Rupp, A. A., Jang, E., & Ahmed, M. (2008, March). Cognitive diagnosis models for nominal response data. New York: Paper presented at the annual meeting of the National Council on Measurement in Education.

Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement, 32*, 559–574.

Tuerlinckx, F., & De Boeck, P. (2001). The effects of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*, 181–195.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report 05-16). Princeton: Educational Testing Service.

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. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28.

Wainer, H., Bradlow, E., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? what is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22–29.

Wang, W., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126–149.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*, 109–128.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika, 92*, 937–950.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Zhang, Y., Li, R., & Tsai, C. L. (2010). Regularization parameter selections via generalized
information criterion. *J. Am. Statist. Ass.*, 105, 312–323.