REGULARIZED LATENT CLASS ANALYSIS WITH APPLICATION IN COGNITIVE DIAGNOSIS

Yunxiao Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying

Abstract

Diagnostic classification models are confirmatory in the sense that the relationship between the latent attributes and responses to items is specified or parameterized. Such models are readily interpretable with each component of the model usually having a practical meaning. However, parameterized diagnostic classification models are sometimes too simple to capture all the data patterns, resulting in significant model lack of fit. In this paper, we attempt to obtain a compromise between interpretability and goodness of fit by regularizing a latent class model. Our approach starts with minimal assumptions on the data structure, followed by suitable regularization to reduce complexity, so that readily interpretable, yet flexible model is obtained. An expectation-maximization type algorithm is developed for efficient computation. It is shown that the proposed approach enjoys good theoretical properties. Results from simulation studies and a real application are presented.

Key words: diagnostic classification models, latent class analysis, regularization, consistency, EM algorithm

1. Introduction

Diagnostic classification models provide multidimensional classifications of respondents for the purpose of a fine-grained diagnosis. Such models have recently become important in educational assessment, psychiatric evaluation, and many other disciplines (Rupp and Templin, 2008; Rupp, Templin and Henson, 2010). For instance, diagnostic classification models have been used to identify students' mastery of different skills based on their responses to testing items, and to diagnose patients' presence of mental health disorders based on their responses to diagnostic questions. Various diagnostic classification models have been developed. A short and incomplete list of works includes Junker and Sijtsma (2001); Tatsuoka (2002); de la Torre and Douglas (2004); DiBello *et al.* (1995); Templin and Henson (2006); Tatsuoka (1985, 2009); Leighton *et al.* (2004); de la Torre (2011); Henson *et al.* (2009); von Davier (2005, 2008); von Davier and Yamamoto (2004); Rupp *et al.* (2010).

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 infers the underlying attribute profiles. The key component in the model specification is the relationship between the observed item responses and the latent attribute profiles. This relationship typically involves the Q-matrix, which assigns each of the items a subset of attributes and specifies the way attributes interact with each other (e.g. compensatory or conjunctive), making diagnostic classification models interpretable and confirmatory.

In many situations, parametric diagnostic classification models (e.g. DINA, DINO, NIDA, etc.) are oversimplified and not flexible enough to capture all the important data patterns, resulting in significant lack of fit. In addition, one should be careful about interpreting these models, due to possible nonidentifiability issues, as voiced, for example, in von Davier (2014), which shows that the DINA model that assumes conjunctive relationship among the attributes is mathematically equivalent to a more general compensatory diagnostic model with a transformed attribute space and a transformed Q-matrix. The general diagnostic classification models, such as the log-linear cognitive diagnosis model (LCDM; Henson *et al.*, 2009), the generalized diagnostic model (von Davier, 2008), and the generalized DINA model (de la Torre, 2011) assume a more flexible relationship between the attribute profiles and responses. However, even for these models, the Q-matrix and the number of attributes associated with the items are specified subjectively and may not be accurate, contributing to the model's lack of fit. The misspecification of the Q-matrix may lead to inaccurate inferences on the latent attribute profiles. Liu *et al.* (2012, 2013), Chen *et al.* (2015b), and Chen *et al.* (2015a) address this issue by constructing the Q-matrix by an objective fashion.

In this paper, we propose a modeling and inference approach that aims to obtain a model that fits the data well and is also simple enough to interpret. To this end, we start with an exploratory latent class model as an exploratory tool (Lazarsfeld *et al.*, 1968; Goodman, 1974a,b), assuming each individual subject belongs to one of M latent classes. The identifiability of such latent class models has been studied, for example, in Allman *et al.* (2009) and Xu (2016). Responses to items are assumed to be independent of each other given the latent class membership, that is, all the dependence among item responses is induced by the latent class membership. This is known as the local independence assumption. This unrestricted latent class model is saturated in the sense that all multivariate discrete distributions can be expressed as a mixture of finitely many independent distributions. As such, this exploratory latent class model is capable of providing a good fit to essentially all data patterns.

However, unrestricted latent class models usually lack interpretability. This is due to the fact that it often includes too many parameters. As a result, it is difficult to identify any pattern or to extract any practical interpretation out of the fitted model. Our approach to this is to reduce the model complexity by regularizing the parameter space.

The regularization we propose to use is based on the following observation. In areas where diagnostic classification models are applied, including educational assessment and mental health diagnosis, latent classes are usually parameterized by multiple attributes (interpreted as skills or mental health disorders). Very often, each item is associated with only a subset of the attributes. Technically speaking, the item response distribution does not depend on the values of certain attributes. Under the latent class model, this means the item response distribution is the same for subjects belonging to several latent classes. This feature is in fact common to all existing diagnostic classification models. We hope that the estimated latent class model also possesses such a pattern, the benefit of which will be explored in the sequel. Therefore, we impose regularization favoring models in which the latent classes are merged and the merging patterns are item-specific. More specifically, we impose a high penalty on the item response functions that take too many distinct values. With this regularization, the fitted model displays a partially merged pattern. The resulting item response function could be completely identical for a subset of latent classes, suggesting that this item does not differentiate among this set of latent classes. The proposed latent class model with the partially merged pattern generalizes the binary skills model in Haertel (1989), in which each item response function admits only two values.

The partially merged pattern describes the relationship between the latent classes and items, that is, groups of latent classes that each item can differentiate. Thus, the partially merged pattern, if correctly estimated, can greatly facilitate the interpretation of the latent classes. Based on the partially merged pattern, one could further establish a partial order relationship among the latent classes and reconstruct a multi-dimensional attribute parametrization of the latent classes. Additional uses and interpretations of the partially merged pattern will be discussed in Section 2.4.

We emphasize that the proposed method is complementary to the confirmatory cognitive diagnostic analysis that typically pre-specifies a Q-matrix and the way the attributes interact. If the signal is strong enough (for instance, simulated data), we are able to reconstruct the Q-matrix and the way the attributes interact through the estimated partially merged pattern.

Consistency results are established under mild regularity conditions; both the model parameters and the partially merged pattern can be consistently estimated by the proposed regularized estimator. Moreover, an efficient computational algorithm is developed. We apply the proposed method to a social anxiety disorder dataset and illustrate its uses with simulation studies.

We proceed as follows. The proposed regularized latent class analysis is described in Section 2 and its theoretical properties discussed in Section 3. Simulation studies are presented in Sections 4. In Section 5, the model is applied to a data set on social anxiety disorder. Section 6 contains a summary. An efficient algorithm is developed and related computational issues are discussed in the appendix.

2. Regularized Latent Class Analysis

First, we provide a brief review of diagnostic classification models and unrestricted latent class models. Then we propose a regularized latent class model that is intermediate between the unrestricted latent class model and parametric diagnostic classification models. The proposed model can be viewed as a generalization of the binary skills model proposed by Haertel (1989). Finally, as opposed to the unrestricted case, the regularized latent class analysis, by learning the partially merged pattern from data, can be used to aid in the construction of a confirmatory diagnostic classification model; for example, the learned partially merged pattern can be used to reparameterize the latent classes by attribute profiles, decide whether binary or polytomous attributes should be used, reconstruct the partial order of the latent classes, etc.

2.1. Diagnostic Classification Models and Latent Class Models

We consider a test consisting of J items taken by N subjects. Let $\mathbf{R} = (R^1, ..., R^J)$ denote the vector of responses to the J items. To simplify discussion, we assume that the responses are all binary, that is $R^j \in \{0, 1\}$. Our approach can be extended to other types of responses.

Diagnostic classification models. Diagnostic classification models assume that a subject's responses to items are governed by his/her latent (unobserved) attribute profile that forms a K-dimensional binary vector, 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/nonmastery of skill k. Both $\boldsymbol{\alpha}$ and $\mathbf{R}$ are subject-specific and we will later use the subscript i to index subjects, that is, $\boldsymbol{\alpha}_i$ and $\mathbf{R}_i$ are the latent attribute profile and response vector for subject $i = 1, ..., N$.

The dependence of item responses on attributes is typically described by the so-called Q-matrix. 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 with the presence of attribute k, $q_{jk} = 0$ otherwise. The precise relationship (e.g. conjunctive, compensatory, etc.) depends on the model parametrization. We use $\boldsymbol{\theta}$ as a generic notation for item-specific 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 *item response function*, the probability for subjects with attribute profile $\boldsymbol{\alpha}$ to provide a positive response to item j, i.e.,

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

Different specifications of parametric forms of $c_{j,\boldsymbol{\alpha}}$ as a function of Q, $\boldsymbol{\alpha}$, and $\boldsymbol{\theta}$ give rise to different diagnostic classification models, an example of which is the log-linear cognitive diagnosis model (LCDM)

$$\begin{aligned}&\operatorname{logit}(c_{j,\boldsymbol{\alpha}})\\ =&\beta_{j,0}+\sum_{k=1}^{K}\beta_{j,k}q_{jk}\alpha_k+\sum_{1\le k_1<k_2\le K}\beta_{j,k_1k_2}q_{jk_1}q_{jk_2}\alpha_{k_1}\alpha_{k_2}+\cdots+\beta_{j,12\cdots K}\prod_{k=1}^{K}q_{jk}\alpha_k.\end{aligned}\tag{2}$$

The LCDM is a saturated model that includes many diagnostic classification models, such as the DINA, DINO, and NIDA, as special cases. This model will be revisited in the sequel. Furthermore, $\boldsymbol{\alpha}_i$'s are independent and identically distributed following

$$\pi_{\boldsymbol{\alpha}}\triangleq P(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}),$$

where $\sum_{\boldsymbol{\alpha}\in\{0,1\}^K}\pi_{\boldsymbol{\alpha}}=1$.

Latent class models. Diagnostic classification models belong to the family of latent class models. In particular, a latent class model assumes that each subject i belongs to one of M latent classes denoted by $m_i\in\{1,2,...,M\}$. The membership indicators are independent and identically distributed with

$$\pi_k=P(m_i=k)\quad\text{for } k=1,...,M,\tag{3}$$

where $\sum_{k=1}^{M}\pi_k=1$. Let $\boldsymbol{\pi}=(\pi_1,...,\pi_M)$. The responses to items are conditionally independent given the latent class m,

$$P(R^j=1|m)=c_{j,m}.\tag{4}$$

Diagnostic classification models impose constraints or structures on the latent variable space and on the parametrization of the response function $c_{j,m}$, through which model components become interpretable. For instance, the latent class membership m is parameterized by a binary vector $\boldsymbol{\alpha}$, each element of which is interpreted as the mastery of a skill or the presence of a mental health disorder. In addition, the item response functions of diagnostic classification models also admit certain constraints and structures. For example, $c_{j,\boldsymbol{\alpha}}$ is often assumed to be monotone non-decreasing in $\boldsymbol{\alpha}$. That is, for attribute profiles $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$ and $\boldsymbol{\alpha}' = (\alpha'_1, ..., \alpha'_K)$, $\alpha_k \geq \alpha'_k$ for all k, implies $c_{j,\boldsymbol{\alpha}} \geq c_{j,\boldsymbol{\alpha}'}$ for all j. This means that a student with attribute profile $\boldsymbol{\alpha}$ is more capable than a student with profile $\boldsymbol{\alpha}'$ and has a higher probability of solving any problem correctly. Lastly, diagnostic classification models also admit some particular interaction among attributes (e.g. conjunctive, compensatory, etc.).

2.2. Partially Merged Latent Classes

Another distinct feature of diagnostic classification models concerns the parametrization of the item response function, which is the focus of the present development. We elaborate with a toy example based on the LCDM, consisting of three arithmetic problems and two attributes. It admits the following self-explanatory Q-matrix.

$$Q = \begin{array}{ccc}
\hline\hline
 & \text{subtraction} & \text{multiplication} \\
\hline
7-2 & 1 & 0 \\
5\times 2 & 0 & 1 \\
(7-2)\times 2 & 1 & 1 \\
\hline
\end{array} \tag{5}$$

The attribute profile $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ contains two elements, i.e. subtraction and multiplication, that stratify the entire population into four latent classes. For instance, the item "$7-2$" only requires attribute subtraction. Then, according to the LCDM specification in (2), $c_{1,(0,0)} = c_{1,(0,1)}$ and $c_{1,(1,0)} = c_{1,(1,1)}$. Similarly, item 2 only requires attribute "multiplication", and thus $c_{2,(0,0)} = c_{1,(1,0)}$ and $c_{1,(0,1)} = c_{1,(1,1)}$. It is in fact a common feature among most diagnostic classification models that an item is not associated with all attributes. We call this the *partially merged* pattern; for each item, there are certain subgroups of latent classes for which the item response function takes identical values.

For unrestricted latent class models, the partially merged pattern does not exist or is at least not particularly enforced. The response function $c_{j,m}$ is usually different for all latent classes. From the modeling point of view, the partially merged pattern is an intermediate step between exploratory models and confirmatory diagnostic classification models.

2.3. Regularized Latent Class Model with Partially Merged Pattern

We now propose to regularize on a latent class model so that the estimated model displays the partially merged pattern. We start with the latent class model (3) and (4). Let R_i^j denote subject i's response to item j. The likelihood function is

$$L(\mathbf{c}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \sum_{m=1}^{M} \Big\{ \pi_m \prod_{j=1}^{J} c_{j,m}^{R_i^j} (1 - c_{j,m})^{1-R_i^j} \Big\}, \tag{6}$$

where $\mathbf{c} = (c_{j,m} : 1 \leq j \leq J,\ 1 \leq m \leq M)$ is the item response function and $\boldsymbol{\pi} = (\pi_1, ..., \pi_M)$ is the latent class distribution. For now, we assume that the number of latent classes M is known; estimation of M when it is unknown will be considered subsequently. The maximum likelihood estimator

$$(\hat{\mathbf{c}}, \hat{\boldsymbol{\pi}}) = \arg\max_{\mathbf{c}, \boldsymbol{\pi}} L(\mathbf{c}, \boldsymbol{\pi}).$$

does not give a partially merged pattern and is often unstable particularly when the sample size is not adequate. In order to obtain the partially merged pattern, we impose a regularization to obtain

$$(\hat{\mathbf{c}}^{\lambda}, \hat{\boldsymbol{\pi}}^{\lambda}) = \arg\max_{\mathbf{c},\boldsymbol{\pi}} \{ l(\mathbf{c},\boldsymbol{\pi}) - N\kappa_{\lambda}(\mathbf{c}) \}, \tag{7}$$

where $l(\mathbf{c},\boldsymbol{\pi}) = \log L(\mathbf{c},\boldsymbol{\pi})$ is the log-likelihood and $\kappa_{\lambda}(\mathbf{c})$ is the regularization on the item response function.

The regularization term κ_{λ} is introduced so that the resulting estimate of the item response function displays the partially merged pattern. We let κ_{λ} be additive in the items. In particular, let $\mathbf{c}_j = (c_{j,1}, ..., c_{j,M})$ be the response function of item j. The regularization term takes the form

$$\kappa_{\lambda}(\mathbf{c}) = \sum_{j=1}^{J} p_{\lambda}(\mathbf{c}_j),$$

where $p_{\lambda}(\mathbf{c}_j)$ is the regularization on the response function for item j. In what follows, we specify p_{λ}. Let $c_{j,(1)} \leq c_{j,(2)} \leq ... \leq c_{j,(M)}$ be the order statistic of $\mathbf{c}_j = (c_{j,1}, ..., c_{j,M})$. The regularization function p_{λ} is chosen to favor $\mathbf{c}_j$ with more identical elements. Equivalently, we consider $d_m = c_{j,(m+1)} - c_{j,(m)}$. Each pair of identical elements in $\mathbf{c}_j$ corresponds to some $d_m = 0$. Thus, we choose p_{λ} to penalize nonzero values of d_m. We hope that if $c_{j,(m+1)} - c_{j,(m)}$ is very close to zero, then the penalty function will force it to be strictly zero and thus two latent classes are merged.

We propose to use the smoothly clipped absolute deviation penalty (SCAD; Fan and Li, 2001). In particular, we choose

$$p_{\lambda}(\mathbf{c}_j) = \sum_{m=1}^{M-1} p_{\lambda}^{SCAD}(c_{j,(m+1)} - c_{j,(m)}), \tag{8}$$

where p_λ^{SCAD} is defined as

$$p_\lambda^{SCAD}(x) = \begin{cases} \lambda|x| & \text{if } |x| \leq \lambda; \\ -\Big(\dfrac{|x|^2 - 2a\lambda|x| + \lambda^2}{2(a-1)}\Big) & \text{if } \lambda < |x| < a\lambda; \\ \dfrac{(a+1)^2\lambda^2}{2} & \text{if } |x| > a\lambda, \end{cases} \tag{9}$$

Remark 1. A natural alternative for p_λ seems to be the L_1 penalty (Tibshirani, 1996)

$$p_\lambda(\mathbf{c}_j) = \lambda|c_{j,(2)} - c_{j,(1)}| + \lambda|c_{j,(3)} - c_{j,(2)}| + ... + \lambda|c_{j,(M)} - c_{j,(M-1)}|.$$

However, this regularization function does not lead to a partially merged pattern. This is because $c_{j,(m)}$ is ordered in m, leading to the following simplification

$$p_\lambda(\mathbf{c}_j) = \lambda|c_{j,(M)} - c_{j,(1)}|.$$

Thus, such a regularization function results in either a fully merged pattern (i.e. all $c_{j,m}$'s are identical) or a completely non-merged pattern. Figure 1 compares the L_1 penalty function $p_\lambda^{L_1}(x) = \lambda|x|$ and the SCAD penalty function, which behaves the same as the L_1 penalty when x is close to zero (i.e. $|x| \leq \lambda$). If x is far away from zero (i.e. $|x| > a\lambda$), the penalty is constant; between λ and $a\lambda$ it is continuous. Thus, SCAD is a local penalty function, leading to the partially merged pattern. Moreover, SCAD has sound theoretical properties (Fan and Li, 2001).

==========================

Insert Figure 1 about here

==========================

On the regularization parameters. The penalty function defined by (9) contains two parameters a and λ, whose values need to be determined. We choose $a = 3.7$ as recommended by Fan and Li (2001). For λ, we use the generalized information criterion (GIC, Nishii *et al.*, 1984) to determine its values (see e.g. Wang *et al.*, 2007, 2009; Chen and Chen, 2008; Wang and Zhu, 2011; Fan and Tang, 2013). Specifically, for a given λ value, we compute the regularized estimator

$$(\hat{\mathbf{c}}^{\lambda}, \hat{\boldsymbol{\pi}}^{\lambda}) = \arg\max_{\mathbf{c},\boldsymbol{\pi}} \{l(\mathbf{c}, \boldsymbol{\pi}) - N\kappa_{\lambda}(\mathbf{c})\}. \tag{10}$$

The generalized information criterion corresponding to the regularized estimator given λ is computed as follows

$$\mathrm{GIC}(\lambda) = -2\ l(\hat{\mathbf{c}}^{\lambda}, \hat{\boldsymbol{\pi}}^{\lambda}) + a_N(\dim(\hat{\mathbf{c}}^{\lambda}) + M - 1) \tag{11}$$

where a_N is a positive number depending on the sample size N, $\dim(\hat{\mathbf{c}}^{\lambda})$ is the total number of distinct response probabilities in vector $\hat{\mathbf{c}}^{\lambda}$, and $M - 1$ counts the number of parameters in $\boldsymbol{\pi}$. Typically, a_N represents the level of penalty on model complexity. We consider two choices of a_N, including

1. GIC_1: $a_N = \log(N)$. This choice corresponds to the Bayesian information criterion.

2. GIC_2 : $a_N = \log\{\log(N)\}\log(N)$. This choice of a_N has been considered in Fan and Tang (2013) for high dimensional model selection. It deals with the case where the dimension of the parameter space d is a polynomial order of sample size N (i.e. $d = N^c$ for some $c > 0$). For the latent class model, it is more appropriate to use this choice as the dimension of the parameter space is $d = J \times M + M - 1$, which is relatively large compared to the sample size. See Fan and Tang (2013) for theoretical properties associated with this choice.

Finally, the regularization parameter is selected to minimize GIC(λ):

$$\lambda_* = \arg\min_{\lambda} \mathrm{GIC}(\lambda). \tag{12}$$

The final estimator is given by $(\hat{\mathbf{c}}^{\lambda_*}, \hat{\boldsymbol{\pi}}^{\lambda_*})$.

Our experience with simulations in Section 4 and real data analysis in Section 5 shows that the GIC as described above work well in practical settings. Moreover, GIC_2 tends to out perform GIC_1 in our simulation studies. The choice of λ controls the trade-off between the bias and variance of the resulting estimator. As will be shown in Section 3, when λ scales properly with sample size, the regularized estimator has desirable theoretical properties, including the convergence of the parameter estimate to its true value and the model selection consistency (i.e. recovering the underlying partially merged pattern).

On the number of latent classes M. The choice of M is a key issue. The general information criterion in (11) can be viewed as a function of both λ and M, denoted by GIC(λ, M). The tuning parameters λ and M may then be chosen jointly:

$$(\lambda_*, M_*) = \arg\min_{\lambda, M} \mathrm{GIC}(\lambda, M). \tag{13}$$

As before, both GIC_1 and GIC_2 are considered.

Summary of the regularized estimator. We propose to use the regularized estimator

$$(\hat{\mathbf{c}}^{\lambda,M}, \hat{\boldsymbol{\pi}}^{\lambda,M}) = \arg\max_{\mathbf{c},\boldsymbol{\pi}} \{l(\mathbf{c}, \boldsymbol{\pi}) - N\kappa_\lambda(\mathbf{c})\}$$

where $l(\mathbf{c}, \boldsymbol{\pi}) = \log L(\mathbf{c}, \boldsymbol{\pi})$ is the log-likelihood corresponding to (6) and κ_λ is the regularization function based on the SCAD penalty (9). With (λ_*, M_*) chosen as in (13), the final estimator is given by $(\hat{\mathbf{c}}^{\lambda_*,M_*}, \hat{\boldsymbol{\pi}}^{\lambda_*,M_*})$. For each (λ, M), the computation of $(\hat{\mathbf{c}}^{\lambda,M}, \hat{\boldsymbol{\pi}}^{\lambda,M})$ is carried out using an Expectation-Maximization algorithm described in the appendix.

2.4. Further Discussion on Partially Merged Pattern

We now discuss modeling questions that can be answered if a partially merged latent class pattern has been consistently identified based on the data.

Reconstructing the latent class parametrization. We illustrate the reconstruction of latent class parametrization via a simple yet illustrative example. Suppose that there are eight latent classes $m \in \{1, 2, ..., 8\}$. Consider an item that is able to differentiate among the following subsets of latent classes

$$\{1, 2, 3, 4\}, \quad \{5, 6, 7, 8\}.$$

Latent classes in curly braces admit the same values of the item response function. Suppose that the item is testing a single skill and thus latent classes 1, 2, 3, and 4 either all master or do not master the skill. Without loss of generality, let the response function take a higher value for the subset $\{1, 2, 3, 4\}$. Based on the estimated item response function and monotonicity, we create a dimension in the attribute profile, with $\{1, 2, 3, 4\}$ corresponding to one in this dimension and $\{5, 6, 7, 8\}$ corresponding to zero. This is the information obtained from one item. If the partially merged patterns of all items have been obtained, one may retrieve the entire attribute parametrization of the latent classes and the Q-matrix, which will be further illustrated in the simulation studies in Section 4. This reconstruction typically requires empirical knowledge of the items and the potential skills the items require, as in the social anxiety disorder example in Section 5.

Other structures of latent classes. Diagnostic classification models often assume the attributes comprising the K-dimensional attribute profile are binary, i.e. there are 2^K latent classes. Very often, not all 2^K latent classes exist in the population, for example, when there is a linear hierarchy among attributes, in which the presence of some attributes requires the presence of others (Leighton and Gierl, 2007). In this case, identifiability problems arise; for instance, the Q-matrix is usually not identifiable under such circumstances (Liu *et al.*, 2013). For a discussion on problems associated with linear hierarchies in diagnostic classification models, see von Davier and Haberman (2014). On the other hand, the regularized latent class model does not have this problem, as it has a built-in mechanism to estimate the number of latent classes, which automatically removes unnecessary latent classes.

Polytomous attributes. Another issue in diagnostic classification models is whether the attributes should be conceptualized as binary or polytomous (von Davier, 2005, 2008; Haberman *et al.*, 2008). The binary attribute structure imposes a clear distinction between the presence and absence of attributes/skills, whereas item response theory models assume the attributes are continuous; a polytomous attribute structure would be intermediate between these two. Whether the model should allow a polytomous attribute profile and how many levels each attribute admits are the key questions in this discussion. The partially merged latent class pattern systematically provides a solution to this problem. We first consider the unrestricted latent class model, under which each latent class admits its unique response distribution. If the number of classes is sufficiently large, then such models behave, to a degree, like a continuous latent factor model. As the latent classes gradually merge, the total number of possible distinct response distributions reduces. In the extreme case that all the latent classes merge together into a single class for an item, then this item does not have any differentiating power. The regularized latent class model has a built-in mechanism to control the merging process of latent classes for each item, which directly provides a choice of the number of levels for each attribute.

Reconstructing the partial order of latent classes. For most diagnostic classification models, the latent parameter space admits a partial order (Tatsuoka and

Ferguson, 2003). The partial order of the latent classes is also closely related to the concept of ordered latent classes (Croon, 1990, 1991), but is more general than the latter. More precisely, the partial order structure of the latent classes allows for two latent classes where members of a first class are better than members of a second class on a subset of items, while members of the second class are better than members of the first class on a different subset of items. On the other hand, the ordered latent classes assume a strict ordering of the latent classes. This structure is important for model interpretation. The partially merged pattern is necessary for reconstructing the partial order. Consider the grouping in (14) and suppose that $c_{1,1} = c_{1,4}$ and $c_{2,1} < c_{2,4}$. Thus, latent class 4 is more capable than latent class 1. If the item response function $c_{j,m}$ is estimated without any constraint (e.g. maximum likelihood estimate of general latent class models), then, due to the asymptotic normality and unbiasedness of maximum likelihood estimation, the estimated item response function admits $P(\hat{c}_{1,1} > \hat{c}_{1,4}) \approx 0.5$ and the partial order of the latent classes cannot be consistently estimated. The partially merged pattern, when estimated correctly, will force $\hat{c}_{1,1} = \hat{c}_{1,4}$ and thus the order can be estimated correctly. We illustrate the reconstruction of the partial order of latent classes in Section 4.

Another advantage of the partially merged pattern is to improve the estimation of the item response function. For example, in the arithmetic problem example in Section 2.2, when $c_{1,1}$ and $c_{1,4}$ are correctly merged, $c_{1,1}$ and $c_{1,4}$ are estimated by the pooled responses of both classes 1 and 4 in the regularized estimator.

3. Theoretical Properties

In this section, we present statistical properties of the regularized estimator $(\hat{\mathbf{c}}^{\lambda}, \hat{\boldsymbol{\pi}}^{\lambda})$ defined as in (10), assuming the number of latent classes M is known. As will be shown below, under suitable conditions, the regularized estimator is consistent for both parameter estimation and model selection (i.e. recovering the partially merged pattern). We first introduce the following notation and definition. Since the latent class distribution satisfies

$$\pi_M = 1 - \sum_{m=1}^{M-1} \pi_m,$$

the log-likelihood function defined as in (7) can be reparameterized by $\mathbf{c}$ and $\boldsymbol{\pi}_{-1} = (\pi_1, .., \pi_{M-1})^\top$. With a slight abuse of the notation, we denote the reparameterized log-likelihood function as $l(\mathbf{c}, \boldsymbol{\pi}_{-1})$. We formalize the concept of a partially merged pattern in the following definition.

Definition 1. Two item response functions $\mathbf{c}$ and $\tilde{\mathbf{c}}$ have the same partially merged pattern if for all $1 \leq j \leq J$ and $1 \leq m_1, m_2 \leq M$ the following statements hold:

(i) if $c_{j,m_1} < c_{j,m_2}$, then $\tilde{c}_{j,m_1} < \tilde{c}_{j,m_2}$;

(ii) if $c_{j,m_1} = c_{j,m_2}$, then $\tilde{c}_{j,m_1} = \tilde{c}_{j,m_2}$.

We write $\mathbf{c} \sim \tilde{\mathbf{c}}$, when they have the same partially merged pattern.

We then impose the following regularity conditions. Denote by $(\mathbf{c}^0, \boldsymbol{\pi}^0)$ the true model parameters.

A1 The analysis is constrained to the following parameter space:

$$\Theta = \{(\mathbf{c}, \boldsymbol{\pi}) : \delta \leq c_{j,m} \leq 1 - \delta, j = 1, ..., J, m = 1, ..., M,$$
$$\text{and } \delta \leq \pi_1 < \pi_2 < ... < \pi_M, \sum_{m=1}^{M} \pi_m = 1\},$$

where δ is a positive constant and $(\mathbf{c}^0, \boldsymbol{\pi}^0) \in \Theta$.

A2 $(\mathbf{c}^0, \boldsymbol{\pi}^0)$ is identifiable over Θ. That is,

$$E[l(\mathbf{c}^0, \boldsymbol{\pi}^0)] > E[l(\mathbf{c}, \boldsymbol{\pi})],$$

for all $(\mathbf{c}, \boldsymbol{\pi}) \in \Theta$ such that $(\mathbf{c}, \boldsymbol{\pi}) \neq (\mathbf{c}^0, \boldsymbol{\pi}^0)$, where the log-likelihood function $l(\mathbf{c}, \boldsymbol{\pi})$ is defined in (7) and the expectation is with respect to the responses from the true model.

A3 The Fisher information matrix of the reparameterized log-likelihood

$$I(\mathbf{c}^0, \boldsymbol{\pi}^0_{-1}) = E\Big[\frac{1}{N}\nabla l(\mathbf{c}^0, \boldsymbol{\pi}^0_{-1})\nabla l(\mathbf{c}^0, \boldsymbol{\pi}^0_{-1})^\top\Big]$$

is positive definite, where $\nabla l(\mathbf{c}^0, \boldsymbol{\pi}^0_{-1})$ is the gradient with respect to $(\mathbf{c}, \boldsymbol{\pi}_{-1})$ evaluated at the true parameters and the expectation is with respect to the responses from the true model.

Remark 2. A1 is assumed for technical convenience and the constant δ can be sufficiently small (e.g., 10^{-5}). It rules out singular cases where the log-likelihood function is negative infinity. In addition, it avoids nonidentifiability due to label switching, by pinning down the order of $\pi_1, ..., \pi_M$. In practice, to solve the optimization (10) with the constraint $\pi_1 < \pi_2 < \cdots < \pi_M$, we first solve the unconstrained one and then switch the labels of the latent classes according to the order of $\hat{\pi}_1, ..., \hat{\pi}_M$. Assumption A2 requires the identifiability of the true model. We refer to Allman *et al.* (2009) and Xu (2016) for discussions on model identifiability in general and the identifiability of latent class models in particular. Assumption A3 is a standard assumption for the maximum likelihood estimator of a general parametric model to achieve the same convergence rate (e.g. Lehmann and Casella, 2006). It ensures the regularized estimator to achieve a $O_P(1/\sqrt{N})$ convergence rate (see Theorem 1). Assumptions similar to A1-A3 are made in Fan and Li (2001).

In what follows, the theoretical properties of the regularized estimator are established, under a proper scaling of the tuning parameter as the sample size grows. The tuning parameter is denoted by λ_N, where the subscript N indicates the scaling.

The first theorem is on the consistency of parameter estimate.

Theorem 1. Under assumptions A1-A3, choose λ_N such that $\lambda_N \to 0$ as $N \to \infty$. Let $(\hat{\mathbf{c}}^{\lambda_N}, \hat{\boldsymbol{\pi}}^{\lambda_N})$ be the optimizer of (10) restricted to Θ. Then, $(\hat{\mathbf{c}}^{\lambda_N}, \hat{\boldsymbol{\pi}}^{\lambda_N})$ is a consistent estimator of $(\mathbf{c}^0, \boldsymbol{\pi}^0)$. Moreover, $(\hat{\mathbf{c}}^{\lambda_N}, \hat{\boldsymbol{\pi}}^{\lambda_N}) = (\mathbf{c}^0, \boldsymbol{\pi}^0) + O_P(1/\sqrt{N})$ as $N \to \infty$. That is, for each positive constant $\varepsilon > 0$, there exists a constant C such that

$$\limsup_{N\to\infty} P\Big(\|(\hat{\mathbf{c}}^{\lambda_N}, \hat{\boldsymbol{\pi}}^{\lambda_N}) - (\mathbf{c}^0, \boldsymbol{\pi}^0)\| > \frac{C}{\sqrt{N}}\Big) \leq \varepsilon,$$

where $\|\cdot\|$ is the Euclidean norm for vectors.

The proof is given in given in Appendix C. The second theorem shows that the partially merged pattern can be consistently recovered.

Theorem 2. Under assumptions A1-A3, choose λ_N such that $\lambda_N \to 0$ and $\lambda_N\sqrt{N} \to \infty$ as $N \to \infty$. Then,

$$\lim_{N\to\infty} P\big(\hat{\mathbf{c}}^{\lambda_N} \sim \mathbf{c}^0\big) = 1.$$

The proof is given in given in Appendix D.

4. Simulation Study

In this section, simulation studies are conducted to evaluate the performance of the proposed method. We also develop visualization methods displaying the estimates as a function of the tuning parameter, which are informative for selecting a model and understanding the data structure.

4.1. Simulation Study 1

We illustrate the use of the proposed method using two simulated datasets, both generated from LCDMs with $K = 3$ attributes, $J = 12$ items, and $N = 1000$

examinees. The difference between the two datasets is that all eight attribute profiles exist for Dataset 1, while for Dataset 2, there are only six attribute profiles due to the presence of a linear hierarchy. We examine the choice of the number of latent classes, reconstruction of the partially merged pattern, partial order of the latent classes, the latent class parametrization, and the visualization of the solution paths.

Simulation setting. Both datasets share the same Q-matrix and the true item response functions (from a LCDM), given in Table 1. There are 9 items that measure a single attribute and therefore have two item response function levels. In addition, there are 3 items that measure two attributes and have three item response function levels. For Dataset 1, all attribute profiles exist and are generated from the uniform distribution:

$$p_{\boldsymbol{\alpha}} = 1/8, \ \forall \ \boldsymbol{\alpha} \in \{0,1\}^3.$$

For Dataset 2, a linear hierarchy is assumed: the second attribute can only be mastered when the first attribute has been mastered. We assume

$$p_{(0,1,0)} = 0, \quad p_{(0,1,1)} = 0 \ \text{ and } \ p_{\boldsymbol{\alpha}} = 1/6, \quad \forall \ \boldsymbol{\alpha} \neq (0,1,0) \text{ or } (0,1,1).$$

==========================

Insert Table 1 about here

==========================

On the number of latent classes and partially merged pattern. We consider the possible number of latent classes $M = 4, 6, 8, 10$, and 12. For each value of M, a solution path is constructed for $\lambda \in [0, 0.1]$. For both datasets, both GIC_1 and GIC_2 correctly select the number of latent classes ($M_* = 8$ for Dataset 1 and $M_* = 6$ for Dataset 2) and select the regularization parameters that result in the true partially merged pattern and partial order of the latent classes. We present the GIC_1, GIC_2, and the number of parameters for the selected models and compare it with that of the unrestricted latent class models in Table 2. According to Table 2, the selected model is not only more parsimonious than the unrestricted latent class models, but also fits the data better in terms of GIC_1 and GIC_2. In addition, the number of latent classes is more likely to be recovered when the partially merged pattern is pursued. In particular, the number of latent classes may be underestimated when only considering the unrestricted latent class models; for instance, for Dataset 1, $M = 6$ is preferred to $M = 8$ based on GIC_2 for unrestricted latent class models, when the dataset is generated with eight latent classes.

===========================

Insert Table 2 about here

===========================

Reconstructing latent class parametrization and Q-matrix. The partial orders of the latent classes based on $\hat{\mathbf{c}}^{\lambda_*}$ are shown in Figure 2. For Dataset 1, under the monotonicity constraint (i.e. $\boldsymbol{\alpha} \geq \boldsymbol{\alpha}'$ implies $c_{j,\boldsymbol{\alpha}} \geq c_{j,\boldsymbol{\alpha}'}$ for all j), the most parsimonious representation is a three-dimensional binary space. In addition, up to label switching, the reparametrization has to be

$$\text{C1} = (0,0,0), \quad \text{C2} = (1,0,0), \quad \text{C3} = (0,1,0), \quad \text{C4} = (1,1,0),$$

$$\text{C5} = (0,0,1), \quad \text{C6} = (1,0,1), \quad \text{C7} = (0,1,1), \quad \text{C8} = (1,1,1).$$

This reparametrization recovers the attribute profiles in the true model. With this reparametrization of the latent classes, the Q-matrix in Table 1 can be perfectly recovered. For Dataset 2,

$$\text{C1} = (0,0,0), \quad \text{C2} = (1,0,0), \quad \text{C3} = (1,1,0),$$

$$\text{C4} = (0,0,1), \quad \text{C5} = (1,0,1), \quad \text{C6} = (1,1,1)$$

provides the most parsimonious reparametrization (up to label switching) when only binary attributes are considered. Based on this latent class reparametrization, the latent classes $(0, 1, 0)$ and $(0, 1, 1)$ do not exist, which suggests the presence of a linear hierarchy of Attributes 1 and 2. In addition, the Q-matrix in Table 1 can still be recovered, given this latent class reparametrization. If polytomous attributes are allowed, the latent classes can be alternatively reparameterized by two attributes:

$$\text{C1} = (0{,}0),\ \ \text{C2} = (1{,}0),\ \ \text{C3} = (2{,}0),\ \ \text{C4} = (0{,}1),\ \ \text{C5} = (0{,}2),\ \ \text{C6} = (2{,}2),$$

where the first dimension combines the original attributes 1 and 2 and the second dimension combines the original attributes 2 and 3.

==========================

Insert Figure 2 about here

==========================

Visualization of the solution paths. In the proposed approach, for each M, a solution path is created as a function of the tuning parameter λ. The solution paths provide information on the uncertainty of the model selection. We monitor the solution paths by plotting the item response functions for each item. We illustrate the visualization in Figure 3 using items 1, 4, 7, 10, 11, and 12 of Dataset 1, where each panel represents an item, the x-axis represents the value of λ, and the y-axis represents the value of the item response function. Each circle represents a particular value of $\hat{c}^{\lambda}_{j,m}$ and the size of the circle is proportional to $\hat{\pi}^{\lambda}_{m}$. The latent classes are identified by different colors. Based on Figure 3, we observe that for each item, there are multiple distinct response probabilities when λ is small. Then, as λ increases, response probabilities that are close tend to merge, resulting in partially merged patterns.

==========================

Insert Figure 3 about here

==========================

4.2. Simulation Study 2

In this study, we evaluate the performance of the proposed method using replicated datasets under various simulation settings.

Simulation setting. We generate data from the LCDM with $K = 3$ and $K = 4$. For $K = 3$, sample sizes $N = 500$, 1000, 2000, and 4000 are considered and the rest of the settings are the same as in Dataset 1 in Study 1. For $K = 4$, 18 items and sample sizes $N = 2000$, 4000, and 8000 are considered. Again, attribute profiles are generated from the uniform distribution

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

The Q-matrix and item response functions are shown in Table 3, where each of the first 12 items measures a single attribute and each of the last 6 items measures two attributes. In addition, items 1-12 have two item response function levels, items 13-15 have three levels, and items 16-18 have four levels.

==========================

Insert Table 3 about here

==========================

Evaluation criteria. For evaluation, we consider the following criteria to account for the correct selection of the partially merged pattern, the partial order of the latent classes, and the number of latent classes.

E1: If the number of latent classes M is known, we consider the frequency that the true partially merged pattern and partial order of the latent classes are captured by $\hat{\mathbf{c}}^{\lambda}$ for at least one value of λ on the solution path.

E2: If M is known, we consider the frequency with which the generalized information criteria selects λ_* such that $\hat{\mathbf{c}}^{\lambda_*}$ corresponds to the true partially merged pattern and partial order of the latent classes.

E3: If M is unknown, we consider the frequency with which the generalized information criteria selects the correct number of latent classes.

E4: If M is unknown, we consider the frequency with which the generalized information criteria selects both the correct number of latent classes and tuning parameter λ_* such that $\hat{\mathbf{c}}^{\lambda_*}$ corresponds to the true partially merged pattern and partial order of the latent classes.

Results. The results are displayed in Table 4. In particular, for the case $K = 3$ and $N = 1000$, results similar to those obtained in the analysis of Dataset 1 in Study 1 are often observed under the same simulation setting. In addition, we observe that as the sample size increases, the proposed approach performs better under all criteria. According to E1, given that M is known, the true partially merged pattern and partial order of the latent classes are captured by the solution path with probability close to 1 when $K = 3$ and $N \geq 1000$ and when $K = 4$ and $N \geq 4000$. According to E2 to E4, GIC_2 performs better than GIC_1 in terms of the selection of tuning parameter and the number of latent classes. In particular, whenever the solution path captures the true model, GIC_2 is able to correctly select the true number of latent classes and the partially merged pattern with probability close to 1 under all settings, while GIC_1 has some chance to miss the true model. Whenever GIC_1 fails, it tends to overfit the models. This suggests that the high dimensional scaling considered in GIC_2 is more appropriate than GIC_1 (Bayesian information criterion), considering the dimension of the parameter space that we choose a model from and the sample size. Finally, when $K = 3$, even with a relatively small sample size $N = 500$, the proposed method preforms reasonably well. Specifically, given M is known, the true partially merged pattern and partial order of the latent classes are captured by the solution path 80% of the time. Moreover, GIC_2 correctly selects the number of latent classes 96% of the time, and correctly selects both the number of classes and the partially merged pattern 72% of the time.

==========================

Insert Table 4 about here

==========================

5. An Application to Social Anxiety Disorder Data

The social anxiety disorder dataset is from the National Epidemiological Survey on Alcohol and Related Conditions (NESARC) (Grant *et al.*, 2003). It contains the binary responses (Yes/No) to thirteen diagnostic questions on social anxiety disorder from 728 white males between 25 and 50 years old. Social anxiety disorder, also called social phobia, is an anxiety disorder in which a person has an excessive and unreasonable fear of social situations. It is the most common anxiety disorder and one of the most common psychiatric disorders, with 12% of American adults having experienced it (Stein and Stein, 2008; Kessler *et al.*, 2005). These thirteen questions are designed according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (American Psychiatric Association, 1994) and are displayed in Table 6. The regularized latent class analysis may provide a better understanding of subtypes of the social anxiety disorder, which may help in the prevention and treatment of this disorder.

We consider $M \in \{4, 5, ..., 15\}$ and $\lambda \in [0, 0.1]$. As GIC_2 tends to outperform GIC_1 according to the simulation study, we present the results based on GIC_2. Based on GIC_2, $M_* = 5$, $\lambda_* = 0.093$, and the selected model contains 34 parameters (while a latent class model with five classes has 69 parameters if no partially merged pattern is enforced). The estimated item response functions and latent class distribution under the selected model is presented in Table 5 and the partial order of the latent classes is shown in Figure 4. In particular, items 3, 4, 6, and 13 have three different item response function levels and the rest have two levels.

Based on Table 5 and Figure 4, Class 3 is the most healthy group, with the lowest item response functions to all items among all latent classes. On the other hand, Class 2 has the highest probabilities of having all symptoms and therefore suffers most. This class may correspond to the generalized social anxiety disorder subtype ("fears most social situations"; American Psychiatric Association, 1994). The other latent classes suffer from some but not all symptoms. For example, Class 1 and 3 share the same item response probabilities, except that Class 1 has higher probabilities of having symptoms 1 to 4 that are all about "public performance". It means that people in Class 1 suffer from the fear of performing/speaking in front of other people, but do not have the other symptoms. This class may correspond to the public speaking phobia subtype that is well known in the literature. Individuals suffering from this subtype have heightened physiological response specific to public speaking or performance situations but are more similar to healthy controls in other situations (e.g. Dalrymple and D'Avanzato, 2013). In addition, Class 4 has the lowest response probabilities on items 9 to 11 that are about "communication with strangers" and the highest probabilities on the rest of the items. Compared to Class 1, Class 4 has strictly higher probabilities of having symptoms 3-8, 12, and 13. Furthermore, Class 5 has the highest probabilities on items 9 to 13, all of which are related to "communication with others". Compared to Class 3, Class 5 has strictly higher item response functions on items 6, and 9-13.

===========================

Insert Table 5 about here

===========================

===========================

Insert Figure 4 about here

==========================

Following the discussion above, we further reconstruct the parametrization of the latent classes. In particular, the most parsimonious reparametrization by binary vectors is as follows:

$$C1 = (1,0,0), \quad C2 = (1,1,1), \quad C3 = (0,0,0), \quad C4 = (1,1,0), \quad C5 = (0,0,1),$$

where the first attribute is about "public performance" (measured by items 1-4), the second attribute is about "public performance" (items 3-4), "being examined" (items 5-8) and "small group communication" (items 12-13), and the third attribute is about "being watched" (items 6 and 9) and "communication with others" (items 9-13). It should be noted that some attribute profiles are missing when reparameterizing the latent classes in this manner. For example, the attribute profile $(0,1,0)$ does not exist, perhaps because patients do not display symptons on "being examined" (items 5-8) and "small group communication" (items 12-13), unless they already display symptoms on "public performance" (items 1-4). The Q-matrix can be reconstructed as in Table 6. Moreover, when polytomous attributes are allowed, the five latent classes can be reparameterized by two attributes:

$$C1 : (1,0), \quad C2 = (2,1), \quad C3 = (0,0), \quad C4 = (2,0), \quad C5 = (0,1).$$

The first attribute is polytomous and it is about "public performance" (items 1-4), "being examined" (items 5-8), and "small group communication" (items 12-13). The second attribute is about "being watched" (items 6 and 9) and "communication with others" (items 9-13).

==========================

Insert Table 6 about here

============================

As demonstrated above, by seeking partially merged patterns of the latent classes, we find latent class models that not only fit data well, but also readily interpretable. In particular, the latent classes of social anxiety disorder respondents are reparameterized by binary/polytomous attributes and the Q-matrix is reconstructed. In addition, our results indicate that a unidimensional latent trait is not enough to capture the latent structure of the social anxiety disorder symptoms, as there exist latent classes that do not follow an order. For example, Class 5 has higher probabilities of having symptoms related to "communication with others", while Class 1 has higher probabilities of having fear of "public performance". To confirm findings from this exploratory analysis, subsequent confirmatory studies are needed, by making use of the diagnostic classification models.

6. Discussions and Summary

In this paper, we propose a latent class model with partially merged pattern for analyzing item response data. Model selection and parameter estimation are carried out simultaneously by a regularized estimator whose theoretical properties, including the model selection and parameter estimation consistencies, are established. For a given number of latent classes, solution paths of the item response functions and the distribution of the latent classes as a function of the tuning parameter are created, providing information on the data structure. In addition, based on the generalized information criteria, the number of latent classes and partially merged pattern are selected, which will help to build a confirmatory diagnostic classification model. In particular, two generalized information criteria are considered. One is the Bayesian information criterion (GIC_1) and the other is a modified Bayesian information criterion (GIC_2) that takes into consideration a high dimensional scaling.

According to the simulation studies, the latter outperforms the former, indicating that the high dimensional scaling in the information criterion is more suitable for this problem.

We evaluate the performance of the proposed approach through simulation studies and an application to the social anxiety disorder data. Simulation studies show that the proposed regularized latent class analysis using GIC_2 accurately recovers the partially merged pattern and partial order of the latent classes if the sample size is adequate. The regularized latent class analysis finds meaningful subgroups of patients who may correspond to different social anxiety disorder subtypes, providing guidance toward a subsequent confirmatory analysis. Subtypes of social anxiety disorder, once identified, may be useful for helping researchers to suggesting different pathways to the disorder and in efforts to prevent and treat the disorder.

References

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

American Psychiatric Association (1994) *Diagnostic and statistical manual of mental disorders, fourth edition.* Washington, DC: American Psychiatric Association.

Chen, J. and Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–771.

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

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

Croon, M. (1990) Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology*, **43**, 171–192.

Croon, M. (1991) Investigating mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology*, **44**, 315–331.

Dalrymple, K. and D'Avanzato, C. (2013) Differentiating the subtypes of social anxiety disorder. *Expert review of neurotherapeutics*, **13**, 1271–1283.

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. V., Stout, W. F. and Roussos, L. A. (1995) Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In *Cognitively diagnostic assessment* (ed. R. L. B. Paul D. Nichols, Susan F. Chipman), 361–389. Hillsdale, NJ: Erlbaum.

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

Fan, Y. and Tang, C. Y. (2013) Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 531–552.

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

Goodman, L. A. (1974a) The analysis of systems of qualitative variables when some of the variables are unobservable. Part I — A modified latent structure approach. *American Journal of Sociology*, **79**, 1179–1259.

Goodman, L. A. (1974b) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

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.

Haberman, S. J., von Davier, M. and Lee, Y.-H. (2008) *Comparison of multidimensional item response models: multivariate normal ability distributions versus multivariate polytomous ability distributions.* (ETS Research Rep. No. RR-08-45). Princeton, NJ: ETS.

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

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.

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.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R. and Walters, E. E. (2005) Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. *Archives of general psychiatry*, **62**, 593–602.

Lazarsfeld, P. F., Henry, N. W. and Anderson, T. W. (1968) *Latent structure analysis.* Boston, MA: Houghton Mifflin.

Lehmann, E. L. and Casella, G. (2006) *Theory of point estimation.* Springer Science & Business Media.

Leighton, J. and Gierl, M. (2007) *Cognitive diagnostic assessment for education: Theory and applications.* Cambridge, UK: Cambridge University Press.

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.

Li, X., Liu, J. and Ying, Z. (2016) Chernoff index for cox test of separate parametric families. *arXiv preprint arXiv:1606.08248.*

Liu, J., Xu, G. and Ying, Z. (2012) Data-driven learning of Q-matrix. *Applied Psychological Measurement*, **36**, 548–564.

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

Nishii, R. *et al.* (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, **12**, 758–765.

Rupp, A. and Templin, J. (2008) 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.* New York, NY: Guilford Press.

Stein, M. B. and Stein, D. J. (2008) Social anxiety disorder. *The Lancet*, **371**, 1115–1125.

Tatsuoka, C. (2002) Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **51**, 337–350.

Tatsuoka, C. and Ferguson, T. (2003) Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 143–157.

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.* New York, NY: Routledge.

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 (Statistical Methodology)*, **58**, 267–288.

von Davier, M. (2005) *A general diagnosis model applied to language testing data.* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

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. (2014) The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, **67**, 49–71.

von Davier, M. and Haberman, S. J. (2014) Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic' classification models — a commentary. *Psychometrika*, **79**, 340–346.

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

Wang, H., Li, B. and Leng, C. (2009) Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 671–683.

Wang, H., Li, R. and Tsai, C.-L. (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568.

Wang, T. and Zhu, L. (2011) Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, **102**, 1141–1151.

Xu, G. (2016) Identifiability of restricted latent class models with binary responses. *arXiv preprint arXiv:1603.04140.*

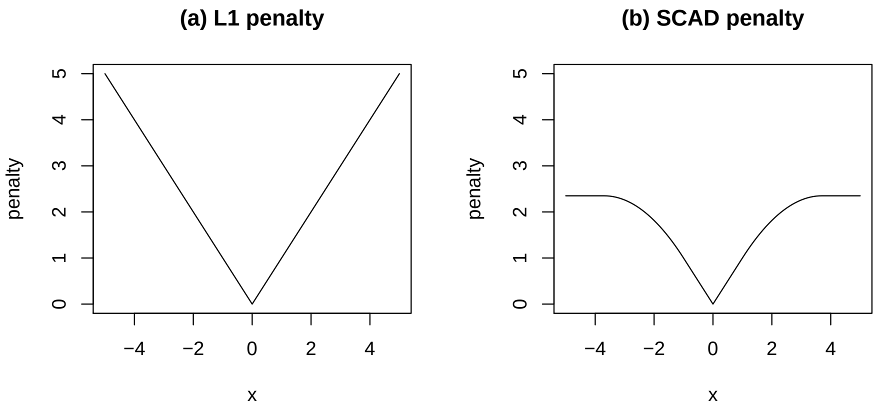

FIGURE 1.
The L_1 and SCAD penalty functions, where $\lambda = 1$ for both penalty functions and $a = 3.7$ for the SCAD penalty function.

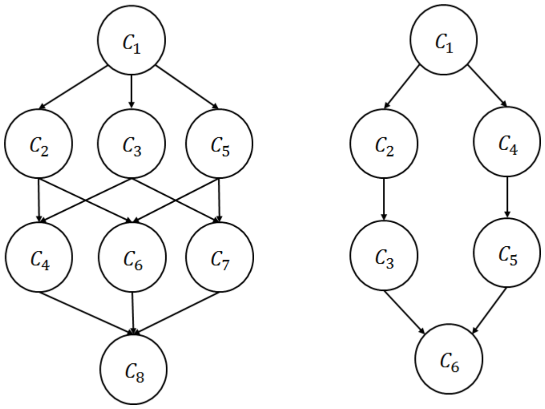

FIGURE 2.
Simulation study 1: the partial order of the latent classes based on regularized latent class analysis. Left: Dataset 1; Right: Dataset 2.

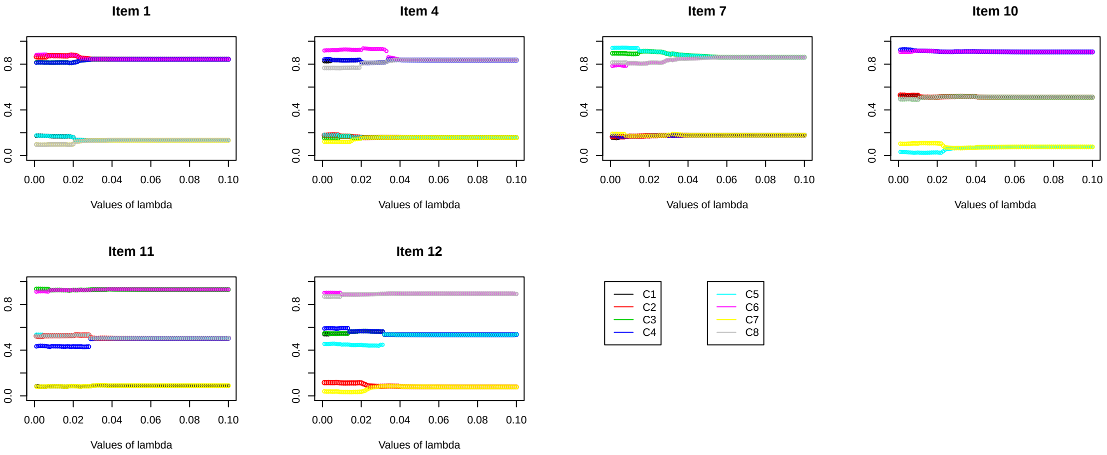

FIGURE 3.
Simulation study 1: the solution paths of the item response functions for items 1, 4, 7, 10, 11, and 12 of Dataset 1, given $M = 8$.

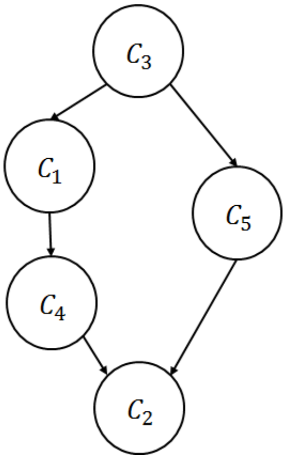

FIGURE 4.
Social anxiety disorder data: the partial order of the latent classes based on regularized latent class analysis

	Q			(0,0,0)	(1,0,0)	(0,1,0)	(1,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,1)
1	1	0	0	0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.85
2	1	0	0	0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.85
3	1	0	0	0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.85
4	0	1	0	0.15	0.15	0.85	0.85	0.15	0.15	0.85	0.85
5	0	1	0	0.15	0.15	0.85	0.85	0.15	0.15	0.85	0.85
6	0	1	0	0.15	0.15	0.85	0.85	0.15	0.15	0.85	0.85
7	0	0	1	0.15	0.15	0.15	0.15	0.85	0.85	0.85	0.85
8	0	0	1	0.15	0.15	0.15	0.15	0.85	0.85	0.85	0.85
9	0	0	1	0.15	0.15	0.15	0.15	0.85	0.85	0.85	0.85
10	1	1	0	0.10	0.50	0.50	0.90	0.10	0.50	0.50	0.90
11	1	0	1	0.10	0.50	0.10	0.50	0.50	0.90	0.50	0.90
12	0	1	1	0.10	0.10	0.50	0.50	0.50	0.50	0.90	0.90

Table 1.
Simulation study 1: the Q-matrix and the item response functions from a LCDM.

Dataset 1	GIC$_1$	GIC$_2$	Num-Par
Selected ($M_* = 8$)	*14194.4*	*14413.5*	34
Unrestricted ($M = 4$)	15064.8	15393.4	51
Unrestricted ($M = 6$)	14689.8	15185.9	77
Unrestricted ($M = 8$)	14595.8	15259.4	103
Unrestricted ($M = 10$)	14729.1	15560.1	129
Unrestricted ($M = 12$)	15859.5	14860.9	155
Dataset 2	GIC$_1$	GIC$_2$	Num-Par
Selected ($M_* = 6$)	*13676.0*	*13882.1*	32
Unrestricted ($M = 4$)	14179.3	14507.9	51
Unrestricted ($M = 6$)	13925.8	14421.9	77
Unrestricted ($M = 8$)	14059.8	14723.3	103
Unrestricted ($M = 10$)	14198.7	15029.8	129
Unrestricted ($M = 12$)	14336.3	15334.8	155

Table 2.
Simulation study 1: GIC$_1$, GIC$_2$, and the number of model parameters (Num-Par) for selected models and the unrestricted latent class models.

		Q			C1	C2	C3	C4	C5	C6	C7	C8
1	1	0	0	0	0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.85
2	1	0	0	0	0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.85
3	1	0	0	0	0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.85
4	0	1	0	0	0.15	0.15	0.85	0.85	0.15	0.15	0.85	0.85
5	0	1	0	0	0.15	0.15	0.85	0.85	0.15	0.15	0.85	0.85
6	0	1	0	0	0.15	0.15	0.85	0.85	0.15	0.15	0.85	0.85
7	0	0	1	0	0.15	0.15	0.15	0.15	0.85	0.85	0.85	0.85
8	0	0	1	0	0.15	0.15	0.15	0.15	0.85	0.85	0.85	0.85
9	0	0	1	0	0.15	0.15	0.15	0.15	0.85	0.85	0.85	0.85
10	0	0	0	1	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
11	0	0	0	1	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
12	0	0	0	1	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
13	1	1	0	0	0.10	0.50	0.50	0.90	0.10	0.50	0.50	0.90
14	1	0	1	0	0.10	0.50	0.10	0.50	0.50	0.90	0.50	0.90
15	1	0	0	1	0.10	0.50	0.10	0.50	0.10	0.50	0.10	0.50
16	0	1	1	0	0.10	0.10	0.35	0.35	0.65	0.65	0.90	0.90
17	0	1	0	1	0.10	0.10	0.35	0.35	0.10	0.10	0.35	0.35
18	0	0	1	1	0.10	0.10	0.10	0.10	0.35	0.35	0.35	0.35
					C9	C10	C11	C12	C13	C14	C15	C16
1					0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.85
2					0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.85
3					0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.85
4					0.15	0.15	0.85	0.85	0.15	0.15	0.85	0.85
5					0.15	0.15	0.85	0.85	0.15	0.15	0.85	0.85
6					0.15	0.15	0.85	0.85	0.15	0.15	0.85	0.85
7					0.15	0.15	0.15	0.15	0.85	0.85	0.85	0.85
8					0.15	0.15	0.15	0.15	0.85	0.85	0.85	0.85
9					0.15	0.15	0.15	0.15	0.85	0.85	0.85	0.85
10					0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
11					0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
12					0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
13					0.10	0.50	0.50	0.90	0.10	0.50	0.50	0.90
14					0.10	0.50	0.10	0.50	0.50	0.90	0.50	0.90
15					0.50	0.90	0.50	0.90	0.50	0.90	0.50	0.90
16					0.10	0.10	0.35	0.35	0.65	0.65	0.90	0.90
17					0.65	0.65	0.90	0.90	0.65	0.65	0.90	0.90
18					0.65	0.65	0.65	0.65	0.90	0.90	0.90	0.90

Table 3.

Simulation study 2: the Q-matrix and item response functions from a LCDM. The latent classes are parametrized as follows: C1 =(0,0,0,0), C2 =(1,0,0,0), C3 =(0,1,0,0), C4 =(1,1,0,0), C5 =(0,0,1,0), C6 =(1,0,1,0), C7 =(0,1,1,0), C8 =(1,1,1,0), C9 =(0,0,0,1), C10 =(1,0,0,1), C11 =(0,1,0,1), C12 =(1,1,0,1), C13 =(0,0,1,1), C14 =(1,0,1,1), C15 =(0,1,1,1), C16 =(1,1,1,1).

	$K=3$				$K=4$		
Sample Size	500	1000	2000	4000	2000	4000	8000
E1	80	100	100	100	78	100	100
E2(GIC$_1$)	64	92	96	99	44	96	98
E3(GIC$_1$)	58	64	68	82	63	66	85
E4(GIC$_1$)	39	58	64	81	25	63	83
E2(GIC$_2$)	72	98	99	100	53	97	99
E3(GIC$_2$)	96	100	100	100	99	100	100
E4(GIC$_2$)	72	98	99	100	53	97	99

Table 4.
Simulation study 2: the number of times among 100 replications that the evaluation criteria are satisfied under different simulation settings

Item	C1	C2	C3	C4	C5
1	0.95	0.95	0.34	0.95	0.34
2	0.87	0.87	0.10	0.87	0.10
3	0.66	0.79	0.03	0.79	0.03
4	0.57	0.89	0.20	0.89	0.20
5	0.10	0.63	0.10	0.63	0.10
6	0.07	0.22	0.07	0.22	0.15
7	0.17	0.65	0.17	0.65	0.17
8	0.14	0.64	0.14	0.64	0.14
9	0.00	0.21	0.00	0.00	0.21
10	0.14	0.91	0.14	0.14	0.91
11	0.09	0.83	0.09	0.09	0.83
12	0.05	0.47	0.05	0.47	0.47
13	0.00	0.25	0.00	0.05	0.25
$\hat{\boldsymbol{\pi}}^{\lambda_*}$	0.32	0.29	0.23	0.09	0.06

Table 5.
Social anxiety disorder data: the estimated item response functions and latent class distribution of the selected model.

Item	Have you ever had a strong fear or avoidance of	A1	A2	A3
1	speaking in front of other people?	1	0	0
2	taking part/ speaking in class?	1	0	0
3	taking part/ speaking at a meeting?	1	1	0
4	performing in front of other people?	1	1	0
5	being interviewed?	0	1	0
6	writing when someone watches?	0	1	1
7	taking an important exam?	0	1	0
8	speaking to an authority figure?	0	1	0
9	eating/drinking in front of other people?	0	0	1
10	having conversations with people you don't know well?	0	0	1
11	going to parties/social gatherings?	0	0	1
12	dating?	0	1	1
13	being in a small group situation?	0	1	1

Table 6.
Social anxiety disorder data: the item contents and the reconstructed loading structure under the latent class reparametrization $C1 = (1,0,0)$, $C2 = (1,1,1)$, $C3 = (0,0,0)$, $C4 = (1,1,0)$, and $C5 = (0,0,1)$.

Appendix

A. Estimation via the Expectation-Maximization Algorithm

We propose to use the expectation-maximization (EM) algorithm combined with the coordinate descent algorithm for the computation of the regularized estimator in (10) for given λ and M. The algorithm guarantees a monotone increasing objective function. Given initial values $\mathbf{c}$ and $\boldsymbol{\pi}$, the algorithm iterates between the E-step and M-step until convergence.

A.1. E-step

In the E-Step, one computes the Q-function,

$$Q(\mathbf{c}^*, \boldsymbol{\pi}^* | \mathbf{c}, \boldsymbol{\pi}) = E_{\mathbf{c},\boldsymbol{\pi}}\{\log L(\mathbf{c}^*, \boldsymbol{\pi}^*; \mathbf{R}_i, m_i, i = 1, ..., N) \mid \mathbf{R}_i, i = 1, ..., N\}. \tag{14}$$

The expectation is taken with respect to $m_i, i = 1, ..., N$. The notation $E_{\mathbf{c},\boldsymbol{\pi}}$ denotes the conditional distribution corresponding to parameters $\mathbf{c}$ and $\boldsymbol{\pi}$. The complete data log-likelihood function is

$$\log L(\mathbf{c}^*, \boldsymbol{\pi}^*; \mathbf{R}_i, m_i) = \sum_{i=1}^{N}\sum_{j=1}^{J} R_i^j \log c^*_{j,m_i} + (1 - R_i^j)\log(1 - c^*_{j,m_i}) + \sum_{i=1}^{N} \log(\pi^*_{m_i}).$$

Under the posterior distribution, m_i, $i = 1, ..., N$ are independent and the posterior distribution associated with the parameters $\mathbf{c}$ and $\boldsymbol{\pi}$ is

$$\begin{aligned} q_{im} :=& P_{\mathbf{c},\boldsymbol{\pi}}(m_i = m | \mathbf{R}_l, l = 1..., M) \\ =& P_{\mathbf{c},\boldsymbol{\pi}}(m_i = m | \mathbf{R}_i) \\ \propto& \prod_{j=1}^{J} c_{j,m}^{R_{ij}}(1 - c_{j,m})^{1-R_{ij}}\pi_m. \end{aligned}$$

The Q-function takes the following additive form,

$$\begin{aligned}
&Q(\mathbf{c}^*, \boldsymbol{\pi}^* | \mathbf{c}, \boldsymbol{\pi}) \\
=&\sum_{j=1}^{J} \sum_{i=1}^{N} \sum_{m=1}^{M} q_{im} \big[R_i^j \log c_{j,m}^* + (1 - R_i^j) \log(1 - c_{j,m}^*) \big] + \sum_{m=1}^{M} \sum_{i=1}^{N} q_{im} \log \pi_m^*.
\end{aligned} \tag{15}$$

A.2. M-step

The M-step consists of maximizing the regularized Q-function with respect to $(\mathbf{c}^*, \boldsymbol{\pi}^*)$

$$\max_{\mathbf{c}^*, \boldsymbol{\pi}^*} Q(\mathbf{c}^*, \boldsymbol{\pi}^* \mid \mathbf{c}, \boldsymbol{\pi}) - N \kappa_\lambda(\mathbf{c}^*).$$

Note that in the objective function, the term

$$\sum_{m=1}^{M} \sum_{n=1}^{N} q_{im} \log \pi_m^*$$

consists only of $\boldsymbol{\pi}^*$, and for each j the term

$$\sum_{i=1}^{N} \sum_{m=1}^{M} q_{im}^* \Big(R_i^j \log c_{j,m}^* + (1 - R_i^j) \log(1 - c_{j,m}^*) \Big) - N \sum_{m=1}^{M-1} p_\lambda^{SCAD} \big(c_{j,(m+1)}^* - c_{j(m)}^* \big)$$

consists only of $\mathbf{c}_j^*$. Therefore, we can maximize the Q-function w.r.t. $\boldsymbol{\pi}^*$ and each $\mathbf{c}_j^*$ independently. In particular,

$$\boldsymbol{\pi}^\dagger = \arg\max_{\boldsymbol{\pi}^*} \sum_{m=1}^{M} \sum_{n=1}^{N} q_{im} \log \pi_m^*$$

can be computed as follows

$$\pi_m^\dagger = \frac{\sum_{i=1}^{N} q_{im}}{\sum_{l=1}^{M} \sum_{i=1}^{N} q_{il}}, \quad m = 1, ..., M.$$

We maximize

$$Q_j(\mathbf{c}_j^*) = \sum_{m=1}^{M} a_m^j \log c_{j,m}^* + b_m^j \log(1 - c_{j,m}^*) - N \sum_{m=1}^{M-1} p_\lambda^{SCAD}\big(c_{j,(m+1)}^* - c_{j(m)}^*\big), \quad (16)$$

where $a_m^j = \sum_{i=1}^N q_{im} R_i^j$ and $b_m^j = \sum_{i=1}^N q_{im}(1-R_i^j)$. Here, a_m^j represents the expected number of respondents who are from latent class m and have responded correctly to item j, and b_m^j represents the expected number of respondents who are from latent class m and have responded incorrectly to item j, given the responses and the current parameter estimates.

Let

$$\mathbf{c}_j^\dagger = \arg\max_{\mathbf{c}_j} Q_j(\mathbf{c}_j). \quad (17)$$

We first show the result for the order of $c_{j,m}^\dagger, m = 1, ..., M$.

Proposition 1. Let $x_{j,m}^* = \frac{a_m^j}{a_m^j + b_m^j}$ and $c_{j,m}^\dagger$ be defined in (17), $j = 1, ..., J$, $m = 1, ..., M$. Then for each j, the order of $c_{j,1}^\dagger, ..., c_{j,M}^\dagger$ is the same as that of $x_{j,1}^*, ..., x_{j,M}^*$. That is, for $l \neq s, 1 \leq l, s \leq M$, if $x_{j,l}^* \geq x_{j,s}^*$ then $c_{j,l}^\dagger \geq c_{j,s}^\dagger$.

Because of this proposition, the computation in (17) is greatly simplified. That is, instead of looking for the solution on the whole domain $[0,1]^M$, we only need to focus on a much smaller subspace (whose volume is $1/(M!)$) that is decided by the order of $x_{j,1}^*, ..., x_{j,M}^*$. On knowing the order of $c_{j,1}^\dagger, ..., c_{j,M}^\dagger$, we parameterize the maximization problem by the order statistics. For instance, if $x_{j,1}^* < ... < x_{j,M}^*$, then $c_{j,(m)}^\dagger = c_{j,m}^\dagger$. In this case, we write

$$\begin{aligned} Q_j^r(c_{j,(1)}^*, d_1, ..., d_{M-1}) = &\sum_{m=1}^{M} \left[a_m^j \log\left(c_{j,(1)}^* + \sum_{l=1}^{m-1} d_l\right) + b_m^j \log\left(1 - c_{j,(1)}^* - \sum_{l=1}^{m-1} d_l\right)\right] \\ &- N \sum_{m=1}^{M-1} p_\lambda^{SCAD}(d_m), \end{aligned}$$

where $d_l = c^*_{j,(l+1)} - c^*_{j,(l)}$. Then we apply the coordinate descent algorithm to the reparametrized function $Q^r_j(c_{j,(1)}, d_1, ..., d_{M-1})$ subject to the constraint that $c_{j,(1)}, d_1, ..., d_{M-1} \geq 0$ and $c_{j,(1)} + \sum_{m=1}^{M-1} d_m \leq 1$. For more details about the coordinate descent algorithm, see Friedman *et al.* (2010).

B. Proof of Proposition 1

Proof. For simplicity of notation, we assume $M = 2$ and $x^*_{j,1} \leq x^*_{j,2}$. For $M > 2$, the proof is similar. Assume to the contrary that $c^\dagger_{j,1} > c^\dagger_{j,2}$. Then according to (17)

$$Q_j(c^\dagger_{j,1}, c^\dagger_{j,2}) \geq Q_j(c^\dagger_{j,1}, c^\dagger_{j,1}) \text{ and } Q_j(c^\dagger_{j,1}, c^\dagger_{j,2}) \geq Q_j(c^\dagger_{j,2}, c^\dagger_{j,2}).$$

According to (16), this can be simplified to

$$a^j_2 \log c^\dagger_{j,2} + b^j_2 \log(1 - c^\dagger_{j,2}) - 2p^{SCAD}_\lambda(c^\dagger_{j,1} - c^\dagger_{j,2}) \geq a^j_2 \log c^\dagger_{j,1} + b^j_2 \log(1 - c^\dagger_{j,1}) \tag{18}$$

$$a^j_1 \log c^\dagger_{j,1} + b^j_1 \log(1 - c^\dagger_{j,1}) - 2p^{SCAD}_\lambda(c^\dagger_{j,1} - c^\dagger_{j,2}) \geq a^j_1 \log c^\dagger_{j,2} + b^j_1 \log(1 - c^\dagger_{j,2}) \tag{19}$$

Because $p^{SCAD}_\lambda(c^\dagger_{j,1} - c^\dagger_{j,2}) \geq 0$, (18) and (19) are still true by removing the term $-2p^{SCAD}_\lambda(c^\dagger_{j,1} - c^\dagger_{j,2})$. According to the definition of $x^*_{j,1}$ and $x^*_{j,2}$, we have

$$x^*_{j,2} \log c^\dagger_{j,2} + (1 - x^*_{j,2}) \log(1 - c^\dagger_{j,2}) \geq x^*_{j,2} \log c^\dagger_{j,1} + (1 - x^*_{j,2}) \log(1 - c^\dagger_{j,1})$$
$$x^*_{j,1} \log c^\dagger_{j,1} + (1 - x^*_{j,1}) \log(1 - c^\dagger_{j,1}) \geq x^*_{j,1} \log c^\dagger_{j,2} + (1 - x^*_{j,1}) \log(1 - c^\dagger_{j,2})$$

Adding these two inequalities up gives

$$(x^*_{j,2} - x^*_{j,1})\Big(\log c^\dagger_{j,2} - \log(1 - c^\dagger_{j,2}) - \log c^\dagger_{j,1} + \log(1 - c^\dagger_{j,1})\Big) > 0.$$

Therefore

$$\log c^{\dagger}_{j,2} - \log(1 - c^{\dagger}_{j,2}) > \log c^{\dagger}_{j,1} - \log(1 - c^{\dagger}_{j,1}). \tag{20}$$

However, the function $\log x - \log(1 - x)$ is strictly increasing for $x \in (0, 1)$, so (20) is impossible. This finishes the proof. □

C. Proof of Theorem 1

Proof. Throughout the proof, we write $a_N = o(b_N)$ for two sequence of vectors a_N and b_N if $\|a_N\|/\|b_N\|$ tend to zero and $a_N = O(b_N)$ if $\|a_N\|/\|b_N\|$ is bounded when N varies. Moreover, for two sequences of random vectors a_N and b_N, we write $a_N = O_P(b_N)$ if $\|a_N\|/\|b_N\|$ converges to zero in probability and $a_N = O_P(b_N)$ if $\|a_N\|/\|b_N\|$ is tight in probability. To simplify the notation, we denote the true model parameters as $(\mathbf{c}, \boldsymbol{\pi})$ and write $\boldsymbol{\theta} = (\mathbf{c}, \boldsymbol{\pi}_{-1})$, $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{c}}^{\lambda_N}, \hat{\boldsymbol{\pi}}^{\lambda_N}_{-1})$ and $\boldsymbol{\theta}' = (\mathbf{c}', \boldsymbol{\pi}'_{-1})$. Note that the event $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \geq \frac{C}{\sqrt{N}}$ implies

$$\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\| \geq \frac{C}{\sqrt{N}}, \boldsymbol{\theta}' \in \Theta} \{l(\boldsymbol{\theta}') - N\kappa_\lambda(\mathbf{c}')\} \geq l(\boldsymbol{\theta}) - N\kappa_\lambda(\mathbf{c}).$$

Therefore, it is sufficient to show that for each $\varepsilon > 0$, there exists a sufficiently large constant C, such that

$$\limsup_{N\to\infty} P\Big(\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\| \geq \frac{C}{\sqrt{N}}, \boldsymbol{\theta}' \in \Theta} \{l(\boldsymbol{\theta}') - N\kappa_\lambda(\mathbf{c}')\} \geq l(\boldsymbol{\theta}) - N\kappa_\lambda(\mathbf{c})\Big) \leq \varepsilon.$$

We split the probability above into two parts,

$$P\Big(\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\| \geq \frac{C}{\sqrt{N}}, \boldsymbol{\theta}' \in \Theta} \{l(\boldsymbol{\theta}') - N\kappa_\lambda(\mathbf{c}')\} \geq l(\boldsymbol{\theta}) - N\kappa_\lambda(\mathbf{c})\Big) \leq I_1 + I_2,$$

where

$$I_1 = P\Big(\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\| \geq \varepsilon_1, \boldsymbol{\theta}' \in \Theta} \{l(\boldsymbol{\theta}') - N\kappa_{\lambda_N}(\mathbf{c}')\} \geq l(\boldsymbol{\theta}) - N\kappa_{\lambda_N}(\mathbf{c})\Big)$$

and

$$I_2 = P\Big(\sup_{\frac{C}{\sqrt{N}} \leq \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \varepsilon_1, \boldsymbol{\theta}' \in \Theta} \{l(\boldsymbol{\theta}') - N\kappa_{\lambda_N}(\mathbf{c}')\} \geq l(\boldsymbol{\theta}) - N\kappa_{\lambda_N}(\mathbf{c})\Big).$$

Here, ε_1 is a positive constant independent of N, whose value will be chosen later. We present upper bounds for I_1 and I_2 separately. The next lemma, whose proof is given in Appendix E, provides an upper bound for I_1.

Lemma 1. For any fixed $\varepsilon_1 > 0$, there exists a positive constant ε_2 (depending on ε_1) such that for sufficiently large N, we have $I_1 \leq e^{-\varepsilon_2 N}$.

We proceed to the I_2 term. We first analyze

$$\sup_{\frac{C}{\sqrt{N}} \leq \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \varepsilon_1, \boldsymbol{\theta}' \in \Theta} \{l(\boldsymbol{\theta}') - l(\boldsymbol{\theta}) - N\kappa_{\lambda_N}(\mathbf{c}') + N\kappa_{\lambda_N}(\mathbf{c})\}.$$

It is straightforward to check that for $\boldsymbol{\theta}' \in \Theta$, there exists a sufficiently large positive constant η, such that

$$\|\nabla l(\boldsymbol{\theta}')\| \leq \eta N, \ \|\nabla^2 l(\boldsymbol{\theta}')\| \leq \eta N \text{ and } \|\nabla^3 l(\boldsymbol{\theta}')\| \leq \eta N, \tag{21}$$

where $\nabla^2 l$ and $\nabla^3 l$ denote vectors consisting of all second and third partial derivatives of l, respectively. According to (21), we compute the Taylor expansion of $l(\boldsymbol{\theta}')$ around $\boldsymbol{\theta}$ for $\boldsymbol{\theta}' \in \Theta$

$$\begin{aligned} & l(\boldsymbol{\theta}') - l(\boldsymbol{\theta}) \\ =& (\boldsymbol{\theta}' - \boldsymbol{\theta})^\top \nabla l(\boldsymbol{\theta}) - \frac{1}{2} N (\boldsymbol{\theta}' - \boldsymbol{\theta})^\top I(\boldsymbol{\theta})(\boldsymbol{\theta}' - \boldsymbol{\theta}) \\ & + O_P(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \sqrt{N}) + O(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^3 N). \end{aligned} \tag{22}$$

In (22), the term $O_P(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \sqrt{N})$ corresponds to the remainder term for the second derivatives at $\boldsymbol{\theta}$ and the term $O(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^3 N)$ corresponds to the terms involving third derivatives. Note that for $\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \varepsilon_1$, there exists a positive constant C_2, independent of ε_1, such that $O(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^3 N) \leq C_2 \varepsilon_1 \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 N$. Thus, the $O(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^3 N)$ term is dominated by the second term, that is,

$$-\frac{1}{2} N(\boldsymbol{\theta}' - \boldsymbol{\theta})^\top I(\boldsymbol{\theta})(\boldsymbol{\theta}' - \boldsymbol{\theta}) + O(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^3 N) \leq -\frac{1}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 N(\inf_{\|v\|=1} v^\top I(\boldsymbol{\theta})v - \varepsilon_1 C_2). \tag{23}$$

Also note that $O_P(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \sqrt{N}) = O_P(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \sqrt{N})$, and $\nabla l(\boldsymbol{\theta}) = O_P(\sqrt{N})$. Thus,

$$(\boldsymbol{\theta}' - \boldsymbol{\theta})^\top \nabla l(\boldsymbol{\theta}) + O_P(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \sqrt{N}) = O_P(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \sqrt{N}). \tag{24}$$

Combining (22), (23) and (24) gives

$$\begin{aligned}
&\sup_{\frac{C}{\sqrt{N}} \leq \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \varepsilon_1} l(\boldsymbol{\theta}') - l(\boldsymbol{\theta}) \\
\leq &\sup_{\frac{C}{\sqrt{N}} \leq \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \varepsilon_1} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| O_P(\sqrt{N}) - \frac{\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 N}{2}(\inf_{\|v\|=1} v^\top I(\boldsymbol{\theta})v - \varepsilon_1 C_2),
\end{aligned}$$

which is further bounded above by

$$\sup_{\frac{C}{\sqrt{N}} \leq \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \varepsilon_1} l(\boldsymbol{\theta}') - l(\boldsymbol{\theta}) \leq \frac{C}{\sqrt{N}} O_P(\sqrt{N}) - \frac{C^2}{2}(\inf_{\|v\|=1} v^\top I(\boldsymbol{\theta})v - \varepsilon_1 C_2).$$

Therefore, by choosing ε_1 sufficiently small, we have

$$\sup_{\frac{C}{\sqrt{N}} \leq \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \varepsilon_1} l(\boldsymbol{\theta}') - l(\boldsymbol{\theta}) \leq -\frac{C^2}{4} \inf_{\|v\|=1} v^\top I(\boldsymbol{\theta})v + O_P(1)C. \tag{25}$$

We proceed to the penalty term. For simplicity of discussion, we only state the proof for the case where there is no $j \in \{1, ..., J\}$ such that $c_{j,1} = c_{j,2} = ... = c_{j,M}$. That is, all items have discrimination power. When there are items that have the same item response function among all the latent classes, the proof is similar, and is thus omitted.

Define a function $\mathbf{gap}(\boldsymbol{\beta}) = \min\{|\beta_i - \beta_j| : \beta_i \neq \beta_j, i = 1, ..., M, j = 1, ..., M\}$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_M) \in R^M$ and there exist i and j such that $\beta_i \neq \beta_j$. Note that the difference of order statistics $c_{j,(m+1)} - c_{j,(m)}$ is either zero or greater than $\frac{\mathbf{gap}(\mathbf{c}_j)}{4}$. Recall in the definition (9), $p^{SCAD}_{\lambda_n}(x) = \frac{(a+1)^2\lambda^2}{2}$ for all $|x| \geq a\lambda$. Thus, the penalty term $p^{SCAD}_{\lambda_N}(c_{j,(m+1)} - c_{j,(m)})$ is either 0 (when $c_{j,(m+1)} - c_{j,(m)} = 0$) or $\frac{(a+1)^2\lambda^2}{2}$ (when $c_{j,(m+1)} - c_{j,(m)} > 0$) for N sufficiently large such that $\lambda_N < \frac{\min_{1\leq j\leq J} \mathbf{gap}(\mathbf{c}_j)}{4a}$. Therefore,

$$\begin{aligned}&\kappa_{\lambda_N}(\mathbf{c})\\=&\sum_{j=1}^{J}\sum_{m=1}^{M-1} p_{\lambda_N}(c_{j,(m+1)} - c_{j,(m)}) = \frac{(a+1)^2\lambda^2}{2}\sum_{j=1}^{J}\text{Card}(\{m : c_{j,(m+1)} - c_{j,(m)} > 0\}),\end{aligned} \tag{26}$$

where Card$(\cdot)$ denotes the number of elements in a set. On the other hand, we have the following lemma on $\kappa_{\lambda_N}(\mathbf{c}')$, whose proof is given in Appendix E,.

Lemma 2. If $\|\mathbf{c}' - \mathbf{c}\| < \frac{1}{4}\min_{1\leq j\leq J} \mathbf{gap}(\mathbf{c}_j)$ and $\lambda_N \leq \frac{1}{4a}\min_{1\leq j\leq J} \mathbf{gap}(\mathbf{c}_j)$, then

$$\kappa_{\lambda_N}(\mathbf{c}') \geq \frac{(a+1)^2\lambda^2}{2}\sum_{j=1}^{J}\text{Card}(\{m : c_{j,(m+1)} - c_{j,(m)} > 0\}).$$

The above lemma and (26) show that $\kappa_{\lambda_N}(\mathbf{c}') - \kappa_{\lambda_N}(\mathbf{c}) \geq 0$ for $\lambda_N \leq \frac{\min_{1\leq j\leq J} \mathbf{gap}(\mathbf{c}_j)}{4a}$. Combine this with (25), we have that for sufficiently large N,

$$\sup_{\frac{C}{\sqrt{N}} \leq \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \varepsilon_1} \{l(\boldsymbol{\theta}') - l(\boldsymbol{\theta}) - N(\kappa_{\lambda_N}(\mathbf{c}') - \kappa_{\lambda_N}(\mathbf{c}))\} \leq -\frac{C^2}{4}\inf_{\|v\|=1} v^\top I(\boldsymbol{\theta})v + O_P(1)C.$$

Note that $\inf_{\|v\|=1} v^\top I(\boldsymbol{\theta})v$ is equal to the smallest eigenvalue of $I(\boldsymbol{\theta})$, which is positive by Assumption A3. Therefore, we have

$$I_2 = P\Big(\sup_{\frac{C}{\sqrt{N}} \leq \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \varepsilon_1, \boldsymbol{\theta}' \in \Theta} \{ l(\boldsymbol{\theta}') - N\kappa_{\lambda_N}(\mathbf{c}') \} \geq l(\boldsymbol{\theta}) - N\kappa_{\lambda_N}(\mathbf{c}) \Big) \leq \frac{\varepsilon}{2}$$

for C sufficiently large. Combining our results for I_1 and I_2, we conclude the proof.

□

D. Proof of Theorem 2

Proof. We first present a useful lemma, whose proof is given in Appendix E.

Lemma 3. There exist constants C and C_1 such that

$$P\Big(\sup_{\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \frac{C}{\sqrt{N}}} \|\nabla l(\boldsymbol{\theta}')\| \leq C_1\sqrt{N}, \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \leq \frac{C}{\sqrt{N}} \Big) > 1 - \varepsilon$$

for sufficiently large N.

Let the event $\Omega_1 = \Big\{ \sup_{\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \frac{C}{\sqrt{N}}} \|\nabla l(\boldsymbol{\theta}')\| \leq C_1\sqrt{N}, \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \leq \frac{C}{\sqrt{N}} \Big\}$. It is sufficient to show that on the event Ω_1, $\hat{\mathbf{c}}^{\lambda_N}$ and $\mathbf{c}$ have the same partially merged pattern for N large enough. We prove this by contradiction. Assume that on the contrary, the partially merged pattern of $\hat{\mathbf{c}}^{\lambda_N}$ and $\mathbf{c}$ are different, we will construct a $\tilde{\boldsymbol{\theta}} \in \Theta$ such that

$$l(\tilde{\boldsymbol{\theta}}) - N\kappa_{\lambda_N}(\tilde{\mathbf{c}}) > l(\hat{\boldsymbol{\theta}}) - N\kappa_{\lambda_N}(\hat{\mathbf{c}}^{\lambda_N}), \tag{27}$$

which contradicts the definition of $\hat{\boldsymbol{\theta}}$. Without loss of generality, assume that $\hat{\mathbf{c}}_1^{\lambda_N}$ and $\mathbf{c}_1$ have different partially merged patterns. That is, there exist $m_1, m_2 \in \{1, ..., M\}$ such that $c_{1,m_1} \leq c_{1,m_2}$ but $\hat{c}_{1,m_1}^{\lambda_N} > \hat{c}_{1,m_2}^{\lambda_N}$. There are two cases: (1) $c_{1,m_1} < c_{1,m_2}$ and (2) $c_{1,m_1} = c_{1,m_2}$. Because on the event Ω_1, $|\hat{c}_{1,m_i}^{\lambda_N} - c_{1,m_i}| < \frac{C}{\sqrt{N}}$ $(i = 1, 2)$, the first case is not possible when N is sufficiently large. Thus, we only need to consider the second case where $c_{1,m_1} = c_{1,m_2}$ and $\hat{c}_{1,m_1}^{\lambda_N} > \hat{c}_{1,m_2}^{\lambda_N}$. Define two sets of indices as follows.

$$A = \{m_2 \in \{1, ..., M\} : \exists m_1 \in \{1, ..., M\} \text{ such that } c_{1,m_1} = c_{1,m_2} \text{ and } \hat{c}^{\lambda_N}_{1,m_1} > \hat{c}^{\lambda_N}_{1,m_2}\},$$

and

$$B = \{l \in \{1, ..., M\} : \hat{c}^{\lambda_N}_{1,l} = \min_{m \in A} \hat{c}^{\lambda_N}_{1,m}\}.$$

The set B is a subset of A, collecting the indices that $\hat{c}^{\lambda_N}_{1,m}$ is minimized. Due to the assumption above, both A and B are non-empty sets. Now we construct $\tilde{\mathbf{c}}$ as follows.

$$\tilde{c}_{1,m} = \begin{cases} \hat{\mathbf{c}}^{\lambda_N}_{1,m} + \Delta \text{ if } m \in B \\ \hat{\mathbf{c}}^{\lambda_N}_{1,m} \text{ if } m \notin B \end{cases},$$

where Δ is a sufficiently small positive number that will be chosen later. For $j = 2, ..., J$ and $m = 1, ..., M$, we keep $\tilde{c}_{j,m} = \hat{c}^{\lambda_N}_{j,m}$. We also set $\tilde{\boldsymbol{\pi}}_{-1} = \hat{\boldsymbol{\pi}}^{\lambda_N}_{-1}$. That is, $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ are the same except for $\tilde{c}_{1,m}$ where $m \in B$. We proceed to compare $l(\tilde{\boldsymbol{\theta}}) - N\kappa_{\lambda_N}(\tilde{\mathbf{c}})$ and $l(\hat{\boldsymbol{\theta}}) - N\kappa_{\lambda_N}(\hat{\mathbf{c}}^{\lambda_N})$. Because $\tilde{\boldsymbol{\theta}}$ and $\tilde{\mathbf{c}}$ depend on Δ, we write $\tilde{\boldsymbol{\theta}}(\Delta)$ and $\tilde{\mathbf{c}}(\Delta)$ to indicate this dependence.

Lemma 4. On the event Ω_1, for N sufficiently large, $\kappa_{\lambda_N}(\tilde{\mathbf{c}}(\Delta))$ is differentiable at 0. Furthermore, $\frac{d\kappa_{\lambda_N}(\tilde{\mathbf{c}}(\Delta))}{d\Delta} = -\lambda_N$.

The lemma above allows us to take the derivative of $q(\Delta) = l(\tilde{\boldsymbol{\theta}}(\Delta)) - N\kappa_{\lambda_N}(\tilde{\mathbf{c}}(\Delta))$ with respect to Δ on the event Ω_1,

$$\dot{q}(0) = \sum_{m \in B} \frac{\partial l(\hat{\boldsymbol{\theta}})}{\partial c_{1,m}} + N\lambda_N.$$

Recall that on event Ω_1, $|\sum_{m\in B}\frac{\partial l(\hat{\boldsymbol{\theta}})}{\partial c_{1,m}}| \leq C_1\sqrt{N}\text{Card}(B)$. This, together with Lemma 4, gives

$$\dot{q}(0) \geq \sqrt{N}(-C_1\text{Card}(B) + \sqrt{N}\lambda_N).$$

Note that $\sqrt{N}\lambda_N \to \infty$ as $N \to \infty$. Thus, $\dot{q}(0) > 0$ for sufficiently large N. This implies that $q(\Delta) > q(0) = l(\hat{\boldsymbol{\theta}}) - N\kappa_{\lambda_N}(\hat{\mathbf{c}}^{\lambda_N})$ for sufficiently small positive Δ. It further implies that (27) holds for such $\tilde{\boldsymbol{\theta}}(\Delta)$, contradicting the definition of $\hat{\boldsymbol{\theta}}$. □

E. Proof of supporting Lemmas

Proof of Lemma 1. Note that the event

$$\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|\geq\varepsilon_1,\boldsymbol{\theta}'\in\Theta}\{l(\boldsymbol{\theta}') - N\kappa_{\lambda_N}(\mathbf{c}')\} \geq l(\boldsymbol{\theta}) - N\kappa_{\lambda_N}(\mathbf{c})$$

implies

$$\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|\geq\varepsilon_1,\boldsymbol{\theta}'\in\Theta}\{l(\boldsymbol{\theta}') - l(\boldsymbol{\theta})\} \geq N\inf_{\boldsymbol{\theta}'\in\Theta}\{\kappa_{\lambda_N}(\mathbf{c}') - \kappa_{\lambda_N}(\mathbf{c})\}.$$

Thus, we have an upper bound for the probability

$$\begin{aligned}&P\Big(\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|\geq\varepsilon_1,\boldsymbol{\theta}'\in\Theta}\{l(\boldsymbol{\theta}') - N\kappa_{\lambda_N}(\mathbf{c}')\} \geq l(\boldsymbol{\theta}) - N\kappa_{\lambda_N}(\mathbf{c})\Big)\\ \leq&P\Big(\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|\geq\varepsilon_1,\boldsymbol{\theta}'\in\Theta}\{l(\boldsymbol{\theta}') - l(\boldsymbol{\theta})\} \geq N\inf_{\boldsymbol{\theta}'\in\Theta}\{\kappa_{\lambda_N}(\mathbf{c}') - \kappa_{\lambda_N}(\mathbf{c})\}\Big).\end{aligned} \tag{28}$$

According to the definition of κ_{λ_N}, we have $0 \leq \kappa_{\lambda_N}(\mathbf{c}') \leq J(M-1)\times\frac{(a+1)^2\lambda_N^2}{2}$. Therefore, (28) is further bounded above by

$$\begin{aligned}&P\Big(\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|\geq\varepsilon_1,\boldsymbol{\theta}'\in\Theta}\{l(\boldsymbol{\theta}') - N\kappa_{\lambda_N}(\mathbf{c}')\} \geq l(\boldsymbol{\theta}) - N\kappa_{\lambda_N}(\mathbf{c})\Big)\\ \leq&P\Big(\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|\geq\varepsilon_1,\boldsymbol{\theta}'\in\Theta}\{l(\boldsymbol{\theta}') - l(\boldsymbol{\theta})\} \geq C_3N\lambda_N^2\Big)\end{aligned}$$

where $C_3 = J(M-1) \times \frac{(a+1)^2}{2}$ is a constant. Note that $\lambda_N \to 0$ as $N \to \infty$, so the right-hand side of the above display is the type I error probability of the generalized likelihood ratio test with a $e^{o(N)}$ cut-off value for testing

$$H_0 : \boldsymbol{\theta}' = \boldsymbol{\theta} \text{ against } H_1 : \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \geq \varepsilon_1, \boldsymbol{\theta}' \in \Theta, \tag{29}$$

whose exponential decay rate has been established in Lemma 3 in Li *et al.* (2016), that there exists a rate $\rho > 0$ such that $P\Big(\sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|\geq\varepsilon_1,\boldsymbol{\theta}'\in\Theta}\{l(\boldsymbol{\theta}') - l(\boldsymbol{\theta})\} \geq C_2 N \lambda_N^2\Big) = e^{-(\rho+o(1))N}$. Choosing ε_2 to be positive and smaller than ρ, we conclude our proof.

Proof of Lemma 2. Because $\kappa_{\lambda_N}(\mathbf{c}') = \sum_{j=1}^{J} p_{\lambda_N}(\mathbf{c}'_j)$, it is sufficient to show that for each $j \in \{1, ..., J\}$

$$p_{\lambda_N}(\mathbf{c}'_j) \geq \frac{(a+1)^2\lambda^2}{2}\mathrm{Card}(\{m : c_{j,(m+1)} - c_{j,(m)} > 0\}).$$

Similar to the discussion proceeding (26), we only need to prove that for each $j \in \{1, .., J\}$,

$$\mathrm{Card}(\{m : c'_{j,(m+1)} - c'_{j,(m)} \geq a\lambda_N\}) \geq \mathrm{Card}(\{m : c_{j,(m+1)} - c_{j,(m)} > 0\}). \tag{30}$$

We first prove that for each $m \in \{1, ..., M-1\}$, if $c_{j,(m+1)} - c_{j,(m)} > 0$, then there exists $m' \in \{1, ..., M-1\}$ such that

$$|c'_{j,m'} - c_{j,(m)}| \leq \frac{1}{4}\mathbf{gap}(\mathbf{c}_j) \text{ and } \min\{c'_{j,l} - c'_{j,m'} : c'_{j,l} > c'_{j,m'}\} \geq a\lambda_N. \tag{31}$$

To proceed, we define a set $D = \{l : c_{j,l} = c_{j,(m)}\}$. We choose $m' \in D$ such that $c'_{j,m'} = \max_{k\in D} c'_{j,k}$. Recall that we assume $\|\mathbf{c}' - \mathbf{c}\| \leq \frac{1}{4}\min_{1\leq j\leq J}\mathbf{gap}(\mathbf{c}_j)$. Thus, we have $|c'_{j,m'} - c_{j,(m)}| = |c'_{j,m'} - c_{j,m'}| \leq \frac{1}{4}\mathbf{gap}(\mathbf{c}_j)$. Moreover, for each l such that

$c'_{j,l} > c'_{j,m'}$, $l \notin D$, due to the choice of m'. We then show

$$c'_{j,l} > c_{j,(m)} + \frac{1}{2}\mathbf{gap}(\mathbf{c}_j)$$

by contradiction. If $c'_{j,l} \leq c_{j,(m)} + \frac{1}{2}\mathbf{gap}(\mathbf{c}_j)$, then

$$c_{j,l} < c_{j,(m)} + \mathbf{gap}(\mathbf{c}_j). \tag{32}$$

Since $c_{j,(m+1)} \geq c_{j,(m)} + \mathbf{gap}(\mathbf{c}_j)$, combining with (32) implies that

$$c_{j,l} = c_{j,(m)} \quad \text{or} \quad c_{j,l} < c_{j,(m)}.$$

On one hand, $c_{j,l} = c_{j,(m)}$ contradicts $l \notin D$. On the other, if $c_{j,l} < c_{j,(m)}$, then $c_{j,l} \leq c_{j,(m-1)}$ and

$$c'_{j,l} \leq c_{j,l} + \frac{1}{4}\mathbf{gap}(\mathbf{c}_j) \leq c_{j,(m-1)} + \frac{1}{4}\mathbf{gap}(\mathbf{c}_j) \leq c'_{j,m'},$$

contradicting $c'_{j,l} > c'_{j,m'}$. Therefore,

$$c'_{j,l} > c_{j,(m)} + \frac{1}{2}\mathbf{gap}(\mathbf{c}_j) \geq c'_{j,m'} + \frac{1}{4}\mathbf{gap}(\mathbf{c}_j) \geq c'_{j,m'} + a\lambda_N,$$

when $\lambda_N \leq \frac{1}{4a}\min_{j\in\{1,\ldots,J\}} \mathbf{gap}(\mathbf{c}_j)$. Therefore, (31) holds for $\lambda_N \leq \frac{1}{4a}\min_{j\in\{1,\ldots,J\}} \mathbf{gap}(\mathbf{c}_j)$. Notice that for different m such that $c_{j,(m+1)} - c_{j,(m)} > 0$, the corresponding m' such that (31) holds are distinct. Thus, (30) is proved.

Proof of Lemma 3. According to Theorem 1, for each ε, there exists a constant C such that for sufficiently large N,

$$P(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| > \frac{C}{\sqrt{N}}) < \frac{\varepsilon}{2}. \tag{33}$$

Now, for $\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \frac{C}{\sqrt{N}}$, we expand $\nabla l(\boldsymbol{\theta}')$ around $\boldsymbol{\theta}$,

$$\|\nabla l(\boldsymbol{\theta}') - \nabla l(\boldsymbol{\theta})\| \leq \sup_{\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \leq \frac{C}{\sqrt{N}}} \|\nabla^2 l(\tilde{\boldsymbol{\theta}})\| \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|.$$

By (21) the right-hand side of the above display is further bounded above by $\eta N \times \frac{C}{\sqrt{N}} = C\eta\sqrt{N}$. Thus, for $\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \frac{C}{\sqrt{N}}$,

$$\|\nabla l(\boldsymbol{\theta}')\| \leq \|\nabla l(\boldsymbol{\theta})\| + C\eta\sqrt{N}.$$

Taking the supremum with respect to $\boldsymbol{\theta}'$ in the above display, we have

$$\sup_{\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \frac{C}{\sqrt{N}}} \|\nabla l(\boldsymbol{\theta}')\| \leq \|\nabla l(\boldsymbol{\theta})\| + C\eta\sqrt{N}.$$

Note that $\|\nabla l(\boldsymbol{\theta})\| = O_P(\sqrt{N})$. This and the above display yield

$$\sup_{\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \frac{C}{\sqrt{N}}} \|\nabla l(\boldsymbol{\theta}')\| = O_P(\sqrt{N}).$$

Consequently, we can choose C_1 sufficiently large such that

$$P(\sup_{\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \leq \frac{C}{\sqrt{N}},} \|\nabla l(\boldsymbol{\theta}')\| > C_1\sqrt{N}) < \frac{\varepsilon}{2}.$$

We combine this with (33), concluding the proof.

Proof of Lemma 4. Let $K = \text{Card}(\{\hat{c}_{1,1}^{\lambda_N}, ..., \hat{c}_{1,M}^{\lambda_N}\})$ be the number of distinct values in $\hat{\mathbf{c}}_1^{\lambda_N}$. Define the vector of ordered distinct values in $\hat{\mathbf{c}}_1^{\lambda_N}$ as $\hat{\gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_K)^T$ such that $\hat{\gamma}_1 < \hat{\gamma}_2 < ... < \hat{\gamma}_K$ and $\{\hat{\gamma}_1, ..., \hat{\gamma}_K\} = \{\hat{c}_{1,1}^{\lambda_N}, ..., \hat{c}_{1,M}^{\lambda_N}\}$. We define $\tilde{\gamma}$ in the same manner. Let k^* satisfy $\hat{\gamma}_{k^*} = \min_{l \in A} \hat{c}_{1,l}^{\lambda_N}$. We choose $|\Delta| < \min\{\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*}, \hat{\gamma}_{k^*} - \hat{\gamma}_{k^*-1}\}$. Then $\tilde{\mathbf{c}}_1$ and $\hat{\mathbf{c}}_1^{\lambda_N}$ have the same partially merged pattern and for

$k = 1, ..., K$

$$\tilde{\gamma}_k = \begin{cases} \hat{\gamma}_k & \text{if } k \neq k^*, \\ \hat{\gamma}_k + \Delta & \text{if } k = k^*. \end{cases} \tag{34}$$

The penalty term for $\tilde{\mathbf{c}}_1$ is

$$p_{\lambda_N}(\tilde{\mathbf{c}}_1) = \sum_{k=1}^{K-1} p_{\lambda_N}^{SCAD}(\tilde{\gamma}_{k+1} - \tilde{\gamma}_k).$$

By (34), the above display becomes

$$p_{\lambda_N}(\tilde{\mathbf{c}}_1) = p_{\lambda_N}^{SCAD}(\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*} - \Delta) + p_{\lambda_N}^{SCAD}(\hat{\gamma}_{k^*} + \Delta - \hat{\gamma}_{k^*-1}) + \sum_{k \notin \{k^*, k^*-1\}} p^{SCAD}(\hat{\gamma}_{k+1} - \hat{\gamma}_k),$$

where we set $p_{\lambda_N}^{SCAD}(\hat{\gamma}_{k^*} + \Delta - \hat{\gamma}_{k^*-1})$ to be 0 if $k^* = 1$. We compare this with the penalty term of $\hat{\mathbf{c}}_1^{\lambda_N}$

$$p_{\lambda_N}(\tilde{\mathbf{c}}_1) - p_{\lambda_N}(\hat{\mathbf{c}}_1^{\lambda_N}) = q_1(\Delta) + q_2(\Delta), \tag{35}$$

where we define $q_1(\Delta) = p_{\lambda_N}^{SCAD}(\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*} - \Delta) - p_{\lambda_N}^{SCAD}(\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*})$ and $q_2(\Delta) = p_{\lambda_N}^{SCAD}(\hat{\gamma}_{k^*} + \Delta - \hat{\gamma}_{k^*-1}) - p_{\lambda_N}^{SCAD}(\hat{\gamma}_{k^*} - \hat{\gamma}_{k^*-1})$. We will show that $\dot{q}_1(0) = -\lambda_N$ and $\dot{q}_2(0) = 0$. To proceed, we first analyze $\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*}$. Let m_2^* satisfy $\hat{c}_{1,m_2^*}^{\lambda_N} = \min_{l \in A} \hat{c}_{1,l}^{\lambda_N} = \hat{\gamma}_{k^*}$. According to the definition of the set A, there exists m_1 such that $\hat{c}_{1,m_1}^{\lambda_N} > \hat{c}_{1,m_2^*}^{\lambda_N}$ and $c_{1,m_1} = c_{1,m_2^*}$. Note that $\hat{c}_{1,m_1}^{\lambda_N} \leq c_{1,m_1} + \frac{C}{\sqrt{N}}$ and $\hat{c}_{1,m_2^*}^{\lambda_N} \geq c_{1,m_2^*} - \frac{C}{\sqrt{N}}$ on the event Ω_1, so $\hat{c}_{1,m_1}^{\lambda_N} \leq \hat{c}_{1,m_2^*}^{\lambda_N} + \frac{2C}{\sqrt{N}}$. Recall that $\hat{\gamma}_{k^*} = \min_{l \in A} \hat{c}_{1,l}^{\lambda_N} = \hat{c}_{1,m_2^*}^{\lambda_N}$ and $\hat{\gamma}_{k^*+1} \leq \hat{c}_{1,m_1}^{\lambda_N}$. Thus, $\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*} \leq \frac{2C}{\sqrt{N}}$. Because $\lambda_N \sqrt{N} \to \infty$ as N grows large, $\frac{2C}{\sqrt{N}} < \frac{\lambda_N}{2}$ for sufficiently large N. Consequently, for $|\Delta| < \frac{C}{\sqrt{N}}$,

$$|\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*} - \Delta| < \lambda_N. \tag{36}$$

According to the definition of $p_{\lambda_N}^{SCAD}$ in (9), and (36), we have

$$q_1(\Delta) = \lambda_N |\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*} - \Delta| \text{ for } |\Delta| < \frac{C}{\sqrt{N}}.$$

Because $\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*} > 0$, for $|\Delta| < \min\{\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*}, \frac{C}{\sqrt{N}}\}$,

$$q_1(\Delta) = \lambda_N(\hat{\gamma}_{k^*+1} - \hat{\gamma}_{k^*} - \Delta).$$

Therefore,

$$\dot{q}_1(0) = -\lambda_N.$$

Now we proceed to the analysis of q_2. If $k^* = 1$, then $q_2(\Delta)$ is set to 0, and so is $\dot{q}_2(\Delta)$. We proceed to the case where $k^* \geq 2$. Choose m_1^* such that $\hat{c}_{1,m_1^*}^{\lambda_N} = \hat{\gamma}_{k^*-1}$. As $\hat{c}_{1,m_1^*}^{\lambda_N} = \hat{\gamma}_{k^*-1} < \hat{\gamma}_{k^*} = \hat{c}_{1,m_2^*}^{\lambda_N}$, we know $m_1^* \notin A$ and $c_{1,m_1^*} \neq c_{1,m_2^*}$ because of the definition of A and B. Furthermore, according to the analysis below (27), it is not possible to have $c_{1,m_1^*} > c_{1,m_2^*}$ on event Ω_1. Thus, we have $c_{1,m_1^*} < c_{1,m_2^*}$. Now let N be sufficiently large such that $\frac{2C}{\sqrt{N}} < \frac{c_{1,m_2^*} - c_{1,m_1^*}}{2}$, then $\hat{c}_{1,m_2^*}^{\lambda_N} - \hat{c}_{1,m_1^*}^{\lambda_N} > \frac{c_{1,m_2^*} - c_{1,m_1^*}}{2}$ on the event Ω_1. Thus, for $|\Delta| < \frac{c_{1,m_2^*} - c_{1,m_1^*}}{4}$ we have

$$\hat{\gamma}_{k^*} + \Delta - \hat{\gamma}_{k^*-1} > \frac{c_{1,m_2^*} - c_{1,m_1^*}}{4}.$$

By the definition in (9), for N sufficiently large such that $a\lambda_N < \frac{c_{1,m_2^*} - c_{1,m_1^*}}{4}$,

$$q_2(\Delta) = p_{\lambda_N}^{SCAD}(\hat{\gamma}_{k^*} + \Delta - \hat{\gamma}_{k^*-1}) = 0 \quad \text{for } |\Delta| < \frac{c_{1,m_2^*} - c_{1,m_1^*}}{4}.$$

Thus, $\dot{q}_2(0) = 0$. Combining this with $\dot{q}_1(0) = -\lambda_N$ and (35), $\frac{d}{d\Delta}\{p_{\lambda_N}(\tilde{\mathbf{c}}_1) - p_{\lambda_N}(\hat{\mathbf{c}}_1^{\lambda_N})\}|_{\Delta=0} = -\lambda_N$. We conclude the proof by noting that $\kappa_{\lambda_N}(\tilde{\mathbf{c}}) = \sum_{j=1}^J p_{\lambda_N}(\tilde{\mathbf{c}}_j)$ and that $\tilde{\mathbf{c}}_j = \hat{\mathbf{c}}_j^{\lambda_N}$ for $j \in \{2, ..., J\}$.