

认知诊断模型属性层级关系和 Q 矩阵的 联合验证方法：面向实践的视角*

汪玲玲¹ 孙小坚²

(¹ 沈阳师范大学教育科学学院, 沈阳 110034) (² 西南大学教师教育学院, 重庆 400715)

摘 要 在认知诊断评估实践中, Q 矩阵和属性层级关系的构建正确与否都会影响认知诊断模型参数估计的准确性以及被试的分类准确率。属性层级关系和 Q 矩阵通常依赖领域专家判断实现, 目前已经有一些研究对 Q 矩阵或者属性层级关系分别进行检验修正。本文提出一种基于贝叶斯网条件独立性检验的方法联合验证 Q 矩阵和属性层级关系, 通过两个模拟研究考察了该方法的联合修正准确率, 以及修正准确率的具体影响因素。结果表明, 在 Q 矩阵错误率处于中等或以下水平时, 该方法能够有效修正 Q 矩阵和属性层级关系, 尤其在题目质量较高样本量充足测验长度较长的情况下, 联合修正效果更好。最后将该算法应用于具体认知诊断评估实践中, 对专家界定的属性层级关系和 Q 矩阵进行联合的基于数据的检验修正, 结果表明修正后的模型拟合更好。

关键词 认知诊断, 属性层级关系, Q 矩阵, 贝叶斯网

分类号 B841

1 引言

认知诊断理论(cognitive diagnostic theory, CDT, Leighton & Gierl, 2007)作为新一代测量理论的核心, 可以促进形成性评价和个性化学习。在认知诊断实践中, 通过认知心理学的分析, 从个体的认知过程、加工技能或知识结构中抽取出细粒度多维度的认知属性, 认知属性也是刻画测验项目的最重要的特征, 通过各种心理计量模型对被试在项目上的作答反应进行分析, 进而对被试的属性掌握情况进行诊断分类。测验项目与认知属性之间的映射关系通过 Q 矩阵表征(Tatsuoka, 1983, 1985)。同时, 一个认知诊断测验中包含的认知属性之间通常具有某种心理上或逻辑上的层级关系, 属性层级关系是认知模型构建的核心(丁树良 等, 2012), 而认知模型正是认知诊断测验所要回答的核心问题——受试者在测验领域的认知加工过程是怎样的(姜宇,

2020)。Q 矩阵和属性层级关系的构建正确与否都会影响认知诊断模型参数估计的准确性以及被试的分类准确率(Chiu, 2013; de la Torre, 2009; Liu et al., 2017; Rupp & Templin, 2008)。通常情况下, 属性层级关系和 Q 矩阵主要都是依赖领域专家的经验判断实现, 不同专家根据自身经验确定的 Q 矩阵和属性层级关系可能存在着差异, 与被试在作答过程中的真实认知过程也可能存在差异, 即原始的 Q 矩阵和属性层级关系都有较大可能包含错误的设定。鉴于二者对于模型参数估计非常重要, 所以有必要对专家主观判断的 Q 矩阵和属性层级关系进行检验修正。

在属性层级关系检验方面, 有研究者在规则空间模型(RSM)及属性层级模型(AHM)的框架下提出层级一致性指标(Hierarchy Consistency Index, HCI, Cui, 2007; Cui & Leighton, 2009; Wang & Gierl, 2011)对属性层级关系进行检验。喻晓锋等(2011)提

收稿日期: 2024-08-06

* 教育部人文社会科学基金青年项目(22YJCZH16); 辽宁省属本科高校基本科研业务费专项资金资助; 国家自然科学基金青年项目(32400930); 教育部人文社会科学基金青年项目(22YJC880065)。

通信作者: 汪玲玲, E-mail: wlling-007@163.com

出使用贝叶斯网方法进行结构学习,从被试的作答反应所获得的属性掌握模式中挖掘属性层级关系,研究显示,通过该方法获得的属性层级关系可作为确定实际属性层级关系的参考。但该方法所直接分析的数据是被试的属性掌握模式,需要通过分析被试的作答反应和 Q 矩阵去估计,而且其估计方法较为初级,“将被试正确作答的项目包含的属性向量进行或运算,得到所有被试的属性掌握模式”在复杂情况下这种不依赖诊断模型估计属性模式的方法不能完全保证估计的准确性。还有研究者在参数化的对数线性认知诊断模型(LCDM, Henson, et al., 2009)基础上开发了层级诊断分类模型(HDCM),创建了一个能够将属性层次结构参数化的心理测量模型和框架,用以证明或证伪某种属性层次结构(Templin & Bradshaw, 2014b)。HDCM 中, Templin 和 Bradshaw (2014b)提出采用似然比检验(likelihood ratio test, LR, 简称 LR 统计量)的方法对属性层级关系进行检验。这种检验通常是针对某一事先设定的属性层级结构的整体进行检验,接受或者拒绝该结构,并不能直接提供对该属性层级结构进行修正的参考信息。

另一方面,对专家初步标定的 Q 矩阵进行检验和修正的方法已经有很多研究。例如,非参数方法包括欧氏距离法(Chiu, 2013)、海明距离(汪大勋等, 2018)等,这些方法具有样本量要求小、易实现等优点,但严格的前提条件限制了这些方法的拓展性及实用性(刘彦楼, 吴琼琼, 2023)。研究者开发的参数化修正方法主要有: δ 法(de la Torre, 2008)、 γ 法(涂冬波等, 2012)、S 统计量方法(Liu et al., 2012)、迭代修正序列搜索(iterative modified sequential search; Terzi & de la Torre, 2018)、RMSEA 统计量(Kang et al., 2019)、加权残差 R 法(Yu & Cheng, 2020)、最优反应分布纯度方法(李佳等, 2022)等。以上方法主要是在特殊的 CDM 框架下(如 DINA、DINO、R-RUM 等)开发的方法。在饱和 CDM 框架下(如 G-DINA, generalized deterministic input noisy output “and” gate; de la Torre, 2011)的参数化 Q 矩阵修正方法主要包括: GDI (G-DINA discrimination index)方法(de la Torre & Chiu, 2016)、TLP (truncated L1 penalty function)方法(Xu & Shang, 2018)、相对拟合统计量方法(Wang et al., 2020; 汪大勋等, 2020)等。其中 Ma 和 de la Torre (2020)提出的 GDI 和 Wald 检验相结合的 Stepwise 方法对多级计分模型中的 Q 矩阵进行修正,同时这种方法也适用于 0-1 计分的

GDINA 模型的 Q 矩阵修正。

在以上介绍的 Q 矩阵检验方法研究中,通常假定属性是相互独立的,没有考虑属性层级关系对方法的影响。同样,已有的检验属性层级关系的方法通常假设 Q 矩阵界定是准确的,而没有考虑包含失误信息的 Q 矩阵界定情况。而在实践中,这二者之间是密不可分相互影响的,并且由于专家经验界定的主观性,二者都不可避免同时存在一些界定偏差信息。Wang 和 Lu (2021)提出两种探索性方法直接从数据中学习属性层级关系,并且这种对属性层级的估计方法是在未知 Q 矩阵信息的情况下进行的。Ma 等(2022)提出一种惩罚似然法,从数据中首先确定属性个数,然后联合估计属性层级关系和 Q 矩阵,这种方法有很强的统计学理论意义,同时也非常复杂。综上,以上两种方法虽然都同时考虑了属性层级关系和 Q 矩阵,但是其运行逻辑都是从带有噪音的数据中直接估计 Q 矩阵和属性层级关系,计算较为复杂,所以以上两个研究其模拟实验所采用的属性个数都较少: 3~4 个。虽然两种方法都是从无到有的更一般化的方法,但是当属性个数较多时,这两种直接从噪音数据中联合估计 Q 矩阵和属性层级关系的方法其运行复杂度可想而知。而在具体的认知诊断评估实践中通常是:在开发的认知诊断测验中,专家根据经验能够对属性层级关系和 Q 矩阵进行初步界定,提供相当程度的有价值的先验信息,只不过这种根据主观经验确定的质性方法难免包含错误,所以对已经界定的属性层级关系和 Q 矩阵进行联合检验和修正的方法同样为实践所急需。基于此,本研究提出了一种基于贝叶斯网视角的属性层级关系和 Q 矩阵的联合检验修正方法。本文接下来介绍基于贝叶斯网的联合验证方法的原理和具体算法实现流程。第三、四部分通过模拟和实证研究对该方法的检验修正效能和影响因素进行评估,最后讨论总结。

2 基于贝叶斯网的联合验证方法实现原理

贝叶斯网络(Bayesian Networks, BN)是结合概率论与图论的概率模型(张连文, 郭海鹏, 2006)。一个贝叶斯网络由一个结构模型和一组条件概率组成,其中结构模型是有向无环图,图中的节点表示随机变量,有向边表示变量之间的相关或因果依赖关系,有向边所联结的两个节点,发起的节点称为父节点,指向的节点称为子节点。这种依赖关系用

网络中每个节点在给定父节点前提下的条件概率来量化。认知诊断中的属性层级关系和 Q 矩阵信息可以表征为贝叶斯网络的结构。例如图 1 所示, 在一个包含 3 个属性和 5 个题目的认知诊断测验中, 属性层级关系以及 Q 矩阵信息可以综合的使用贝叶斯网的结构来表征, 其中椭圆表示认知属性, 而矩形则表示测验题目, 从该网络结构中可看到每个题目与所测量的属性之间的连接关系, 这是 Q 矩阵所包含的信息; 另外, 该网络还呈现了各属性之间的层级关系, 如属性 A1 是属性 A2 的先决属性, 而这两个属性也是 A3 的先决属性。贝叶斯网络中, 属性和题目之间的单向箭头(称为“边”)表示 Q 矩阵中元素“1”, 即各测验题目所考察的认知属性, 同理属性之间的边则表示属性之间具有相关关系。贝叶斯网络可以对网络中的各条边进行显著性检验, 以明确各条边是否具有存在的意义, 如此, 对网络结构中每一条边进行显著性检验, 等价于检验 Q 矩阵中每一个题目的属性标定是否正确, 与此同时也能够检验各属性之间的边的显著性, 即检验了属性层级关系的合理性。

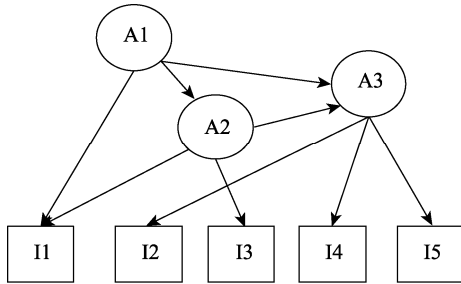


图 1 综合表征属性层级关系和 Q 矩阵的贝叶斯网路结构

在贝叶斯网中, 可以通过检验两个节点变量之间的条件独立性来检验联结它们之间的边的存在的合理性。这种检验可以通过条件卡方检验(Pearson Chi-square)或条件对数似然比检验(Log likelihood Ratio)实现(薛薇, 陈欢歌, 2012)。条件独立性检验的虚无假设为两个节点之间是条件独立的, 即不存在边, 如果检验得到的显著性水平非常低, 那么就可以拒绝虚无假设, 认为两个节点之间存在边的联结, 说明该题目测量了这个属性, 或者这条边连接的两个属性之间具有层级关系。在这两种检验中, 检验统计量都服从渐近卡方分布。这个条件卡方统计量定义为以下形式:

$$\chi^2(X_i, X_j | S) = \sum_k \chi^2(X_i, X_j | S = s_k) = \sum_{m,n,k} \frac{(N(x_i^m, x_j^n, s_k)N(s_k) - N(x_i^m, s_k)N(x_j^n, s_k))^2}{N(x_i^m, s_k)N(x_j^n, s_k)N(s_k)}$$

其中 $N(x_i^m, x_j^n, s_k)$ 表示变量 $X_i = x_i^m, X_j = x_j^n, S = s_k$ 时的实际观测次数, 其他的符号标记类同。在认知诊断评估情境下, 所有变量都是 0-1 二值计分, 所以这时卡方统计量的计算非常快捷。

综上可知, 使用贝叶斯网可以联合的同时验证 Q 矩阵和属性层级关系的正确性, 同时对于题目标定错误的属性, 或者错误设定的属性之间的关系, 也能够通过贝叶斯网中节点之间的独立性检验进行修正。与条件独立性检验针对网络中每一条边的合理性进行检验不同的是, 网络分数(network score)能够整体的评估一个贝叶斯网的网络结构与实际数据的拟合情况, 网络分数通常包括 AIC (Akaike, 1974), BIC (Schwarz, 1978), 和对数似然值(log-likelihood criterion, Pinheiro & Bates, 1995)。在 bnlearn R 软件包中, 除了常见的 BIC 指标, 还有一个 BDeu 指标(Bayesian Dirichlet equivalent uniform score, Heckerman et al., 1995), 这两个指标都能够直接实现。综上, 贝叶斯网的网络结构可以通过条件独立性检验或者描述模型数据拟合总体情况的网络分数进行验证。所以, 根据贝叶斯网的这种性质可以对综合表征属性层级关系和 Q 矩阵的贝叶斯网结构进行检验修正。使用贝叶斯网络进行属性层级关系和 Q 矩阵检验和修正的具体实现步骤如下:

步骤 1: 根据初始构建的 Q 矩阵和属性层级关系估计被试的属性掌握模式。

步骤 2: 初始 Q 矩阵可能存在属性冗余, 也可能存在属性缺失, 初始的属性层级关系也可能存在冗余和缺失。如果初始的贝叶斯网络结构仅仅根据初始 Q 矩阵和初始属性层级关系进行建构, 那么根据条件独立性检验只能检验出冗余的边, 而对于缺失的边无法检测到。所以我们初始构建一个贝叶斯网的饱和模型, 也就是假设每个题目测量了所有属性, 任意两个属性之间都有边的联结, 所有节点变量之间都存在边的连接, 这样能够对每一条边进行检验。

步骤 3: 将作答反应数据和已经估计的被试属性掌握模式输入这个贝叶斯网的饱和结构模型, 并检验其中每一条边存在的强度, 也就是属性节点与题目节点之间或者两个属性节点之间条件独立性关系的显著性水平, 保留检验显著的边, 进而得到一个校正的 Q 矩阵和校正的属性层级关系。根据预实验研究, 本研究所采用的显著性水平为 0.001。需要指出的是在 BN 中, 对所有边存在的显著性检验是一次同时进行的, 而不是每一条边分别进行, 这

也保证了该算法的计算效率。

步骤 4: 由于依赖初始的 Q 矩阵和属性层级关系估计的被试属性掌握模式存在误差, 所以根据这批带有噪声的数据进行条件独立性检验所修正的 BN 结构并不完全准确, 需要继续结合模型数据拟合指数(network score)进行校正。对于 Q 矩阵和属性层级关系中每一个被修正的元素, 计算修正之后的 BN 结构的拟合指数(本研究采用 BIC 指标, 经预实验, BIC 和 BDeu 两个指标的修正效果没有明显差异), 并与未修正之前的 BN 结构的拟合指数进行比较, 如果修正之后降低了拟合指标 BIC, 则保留此次修正结果, 反之则保留未修正之前的结果。根据总体模型数据拟合指标再次对 BN 的结构, 也就是 Q 矩阵和属性层级关系进行优化和校正。

步骤 5: 根据校正的 Q 矩阵和属性层级关系重新估计被试的属性掌握模式。

步骤 6: 将重新估计的被试属性掌握模式和作答反应数据再次输入 BN 模型中, 重复步骤 3、4、5, 可以预见第二次的迭代中, 被试的属性掌握模式估计更精确, 所以这批输入数据的噪声降低, 进而根据条件独立性和模型数据拟合指标综合之后修正的 BN 结构更加准确, 进而估计的被试的属性掌握模式也更加准确。根据预实验研究, 通常 3 次迭代之后, 便可以达到稳定的 BN 结构, 也就是稳定的 Q 矩阵和属性层级关系, 继续迭代并不能显著提高 Q 矩阵和属性层级关系的修正准确率。需要指出的是该算法的计算效率比较高, 即使是 3 次迭代, 仍然能够保证比较高的计算效率。在模拟研究中, 将最后求取的 Q 矩阵与真值比较, 计算这种方法的修正准确率, 同样的将最后确定的属性层级关系与生成数据的真值比较, 确认是否修正成功。检验贝叶斯网边的存在强度以及计算模型整体拟合指标的过程都可以通过 R 软件包(bnlearn; Scutari & Denis, 2021)实现。

3 Q 矩阵和属性层级关系的联合检验修正方法的影响因素分析

下面我们通过两个模拟研究具体考察各种影响因素对贝叶斯网检验修正 Q 矩阵和属性层级关系准确性的影响。模拟研究可以控制一些额外变量, 只考查抽象的测验情景下该方法的表现, 它可以提供基于贝叶斯网的检验方法基本的数值稳定性与精确性。另外, 目前已经开发很多 Q 矩阵修正方法, 这些方法通常都不考虑属性层级关系的影响。以往

对属性层级关系检验的研究中, 并没有严格的考察其修正准确率, 而且没有考虑 Q 矩阵的界定正确与否。已有的两项关于 Q 矩阵和属性层级关系的研究是联合估计, 而非联合检验, 综上, 本研究并没有基线方法作为对照。所以, 作为一项初步探索研究, 本文通过模拟研究探查该方法对属性层级关系和 Q 矩阵的修正准确率, 以及具体的影响因素, 以期对认知诊断评估的实践应用者提供有价值的参考信息。

3.1 模拟研究一

3.1.1 研究设计

本研究的 Q 矩阵测量 5 个属性, 测验长度分别 25, 40 两种水平, 25 题的 Q 矩阵具体设计参照姜宇(2020)和 Zhang 等(2024)的研究, 在 25 题基础上, 重复前 15 题并保证属性测量次数的平衡设计 40 题的 Q 矩阵。不失一般性, 认知诊断模型选取饱和 GDINA 模型。被试样本量设为两个水平 1000, 2000。题目参数的设置参照已有同类研究(李佳 等, 2022): 掌握项目 i 所测量的所有属性的被试和项目 i 的属性都没有掌握的被试对项目 i 作答正确的概率分别记为 $P_i(1)$ 和 $P_i(0)$, 项目质量分为高、低两个水平, 高质量项目的参数为: $P_i(0) \sim U(0.05, 0.25)$ 且 $P_i(1) \sim U(0.75, 0.95)$, 低质量的项目参数为: $P_i(0) \sim U(0.05, 0.4)$ 且 $P_i(1) \sim U(0.6, 0.95)$ 。然后模拟错误的 Q 矩阵, 错误率设置为 0(没有错误), 10%, 20%, 30% 四个水平, 在真实 Q 矩阵的基础上按错误率随机确定相应比例的项目, 错误设定参考已有研究(刘彦楼, 吴琼琼, 2023), 在两个约束条件下随机引入: (1) 所有项目必须至少测量一个属性; (2) 始终保留一个单位矩阵, 以满足模型可识别条件(Gu et al., 2018)。该测验的属性层级关系设定为直线型、收敛型、发散型和无结构型四种类型, 具体设置如图 2 所示。然后生成各种属性层级关系下的被试属性掌握模式真值, 被试从所有可能的属性掌握模式中随机抽取产生。根据所采用的认知诊断模型和题目参数模拟被试在题目上的作答反应, 生成被试作答反应数据。

产生模拟数据之后, 使用 HDCM 估计被试的属性掌握模式(通过 R 软件包 GDINA 实现), 在估计之前设定认知属性的层级关系, 在原有的属性层级关系真值的基础上, 分别产生有缺失边, 冗余边的属性层级关系, 以随机产生的这些模拟的属性层级关系为初值, 估计被试的属性掌握模式。然后根据前面所述步骤使用贝叶斯网联合修正属性层级

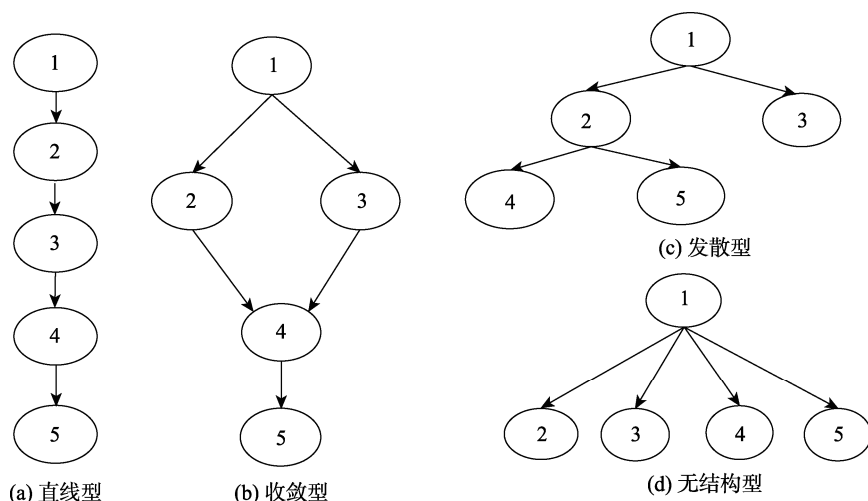


图 2 模拟研究设置的 5 个属性的 4 种真实的属性层级关系及 7 个属性的发散结构

关系和 Q 矩阵。需要特别说明的是, 在 4 种基本的属性层级模型中, 可以随机生成具有缺失值的边, 具体如下: 在每一种层级关系中, 对于任意两个属性之间的连接关系, 随机选定一对属性, 将其连接关系置为缺失。在生成冗余边时, 对于发散型和无结构型, 仍存在 5 个属性之间的冗余关系, 例如在图 2 所示的发散型结构中, 可能存在冗余边的属性对有(A2, A3), (A4, A5), (A3, A4), (A3, A5), 在无结构型中, 可能存在冗余的属性对有(A2, A3), (A3, A4), (A4, A5), (A2, A4), (A2, A5), (A3, A4), (A3, A5), (A4, A5)。而在直线型中, 5 个属性之间已经完全联结, 在任意两个属性对之间即使没有直接的边联结, 但是也会存在通过其他节点的联结, 所以并不能生成有效的属性冗余关系。同样道理, 在收敛型结构中, 只有一个有效冗余边, 即(A2, A3), 但由于模拟研究需要多次模拟随机生成冗余边, 故在生成冗余边时, 也不再考虑收敛型结构。综上, 在缺失边的情境下, 4 种属性层级结构都考虑, 在冗余边的情境下, 仅使用发散型和无结构型两种结构。综上, 本研究在属性边缺失情境下模拟数据的设置有: 2 (测验长度) $\times 2$ (样本量水平) $\times 2$ (题目参数水平) $\times 4$ (属性层级类型) $\times 4$ (Q 矩阵错误率) = 128 种情况, 在属性边冗余情境下模拟数据的设置为: 2 (测验长度) $\times 2$ (样本量水平) $\times 2$ (题目参数水平) $\times 2$ (属性层级类型) $\times 4$ (Q 矩阵错误率) = 64 种情况。

3.1.2 评价指标

在以上设计的每种实验条件下, 分别随机产生 100 批模拟数据, 然后使用模拟的带有错误信息的属性层级初始值估计属性掌握模式, 通过贝叶斯网

条件独立性检验和模型数据拟合指标的结合, 最后得到修正的属性层级关系和修正的 Q 矩阵, 将最终的修正结果与属性层级关系真值进行比较, 如果完全相同, 则代表修正正确一次, 最后计算每种情境下 100 次修正属性层级关系的正确次数作为评价指标。Q 矩阵的修正准确率所采用的指标为每次修正后的 Q 矩阵与真实 Q 矩阵题目测量模式的一致性作为模式判准率(Pattern classification rate, PCR)和每次修正后的 Q 矩阵与真实 Q 矩阵属性的一致性作为平均属性判准率(Average attribute classification rate, AACR), 以及 TPR (True positive rate)和 TNR (True negative rate)。其中 TPR 指错误标定的属性正确修改的比率, TNR 指正确界定的属性未被修改的比率(由于篇幅限制, 这部分评价指标结果可以联系作者提供)。所有实验均重复 100 次, 然后计算每个指标 100 次的平均值。

3.1.3 研究结果

使用贝叶斯网条件独立性检验方法修正属性层级关系的结果如表 1、表 2 所示, 在 Q 矩阵设定完全正确的情况下, 无论题目是高质量还是低质量, 无论属性关系是缺失还是冗余, 无论样本量是较大还是较小, 无论测验长短, 在各种属性层级结构中, 该算法都能够完全准确的修正属性层级关系, 在 100 次的模拟中, 完全修正正确次数达到 100%。而当 Q 矩阵存在界定错误时, 随着 Q 矩阵中错误元素的占比增大, 在各种条件下, 对属性层级的修正准确率逐渐降低。具体而言, 在 Q 矩阵失误概率为 10%的情况下, 无论题目质量高低, 样本量大小, 测验长短, 在 4 种属性层级结构下, 无论属性关系缺失还是冗余, 仍然能够达到较高的完全正确的修

正比例, 除个别情况, 大部分情况下, 仍能达 90% 或以上的比例。随着 Q 矩阵中错误率的提高, 对属性层级完全修正正确的比例逐渐降低, 这时这种比例也受到测验长度, 题目质量和样本量的同时影响, 例如在 Q 矩阵错误率为 20% 时, 当题目质量较低时, 测验长度和样本量能够显著提高这种完全修正正确的比例, 除极个别情况, 大多情况下的完全正确

的比例仍达到 80% 以上。4 种不同的属性层级类型下, 无结构型的修正正确次数在样本量较低时受题目质量和 Q 矩阵错误率影响较大, 其他几种类型的属性层级中, BN 在各种情况下的表现并没有明显的差异, 在其他条件设置相同的情况下, 属性缺失和属性冗余不同的错误类型中, BN 的表现没有明显的差异。

表 1 属性缺失情境下 4 种层级结构完全正确修正次数(100 次)

属性缺失			发散型		无结构型		线性型		聚合型	
Q-error	test length	g-s	2000	1000	2000	1000	2000	1000	2000	1000
0%	25	0.05-0.25	100	100	100	100	100	100	100	100
		0.05-0.4	100	100	100	100	100	100	100	100
	40	0.05-0.25	100	100	100	100	100	100	100	100
		0.05-0.4	100	100	100	100	100	100	100	100
10%	25	0.05-0.25	100	84	100	96	96	92	96	93
		0.05-0.4	90	82	98	96	89	87	96	91
	40	0.05-0.25	100	100	100	100	100	100	100	100
		0.05-0.4	98	92	95	95	96	91	97	93
20%	25	0.05-0.25	79	75	90	90	81	74	89	83
		0.05-0.4	74	64	82	50	76	66	82	75
	40	0.05-0.25	91	85	98	95	96	97	98	96
		0.05-0.4	82	76	85	57	85	84	88	90
30%	25	0.05-0.25	70	47	54	56	68	58	64	58
		0.05-0.4	60	30	46	18	64	60	58	54
	40	0.05-0.25	77	76	75	75	93	84	85	85
		0.05-0.4	70	62	53	44	75	72	74	61

表 2 属性冗余情境下 2 种层级结构完全正确修正次数(100 次)

属性冗余			发散型		无结构型	
Q-error	test length	g-s	2000	1000	2000	1000
0%	25	0.05-0.25	100	100	100	100
		0.05-0.4	100	100	100	100
	40	0.05-0.25	100	100	100	100
		0.05-0.4	100	100	100	100
10%	25	0.05-0.25	98	98	100	98
		0.05-0.4	96	92	99	86
	40	0.05-0.25	100	100	100	93
		0.05-0.4	100	100	100	90
20%	25	0.05-0.25	90	89	90	89
		0.05-0.4	82	64	83	49
	40	0.05-0.25	100	95	100	87
		0.05-0.4	95	93	90	80
30%	25	0.05-0.25	64	60	63	57
		0.05-0.4	60	54	49	46
	40	0.05-0.25	80	75	90	80
		0.05-0.4	80	65	58	55

在属性层级界定存在缺失或者冗余的情况下,使用贝叶斯网方法对界定失误的 Q 矩阵进行修正的准确率(包括 PCR 和 AACR)如表 3、表 4 所示。在 Q 矩阵错误率在 10%时,4 种属性层级结构存在缺失和冗余时,无论题目质量高低,无论样本量大小,无论测验长短,贝叶斯网方法对 Q 矩阵的修正准确率能够达到较高水平,即使在题目质量较低的情况下,除个别情况,模式判准率都能够达到 70%以上,平均属性判准率除个别情况都能达到 95%以上。在 Q 矩阵错误率为 20%的情况下,题目质量较高时,PCR 仍然能够达到 70%以上,AACR 能够达到 92%以上,而在题目质量较低的情况下,PCR 在较多情境下能够达到 50%以上,除极个别,AACR 达到 90%以上。

在 Q 矩阵错误率为 30%时,PCR 在短测验题目质量较高的情况下能够达到 53%以上,除极个别,AACR 达到 90%以上,长测验情境能够一定程度上弥补题目质量较低的影响,除无结构型的属性关系,其他情况下,PCR 仍然能够达到 50%以上,在题目质量较低并且短测验情境下,PCR 和 AACR 都比较低。从总体上看,样本量对该方法的修正准确率的影响也比较大,在样本量从 1000 提高到 2000 时,每种情境下的修正准确率都得到较大的提高。测验长度对修正准确率也有明显的影响,尤其是在 Q 矩阵错误率较高或者题目质量较低的情况下,相比短测验情境,长测验情境下 BN 的修正准确率提高更多。

表 3 属性缺失情境下 Q 矩阵的修正准确率(PCR 和 AACR)

属性缺失的 PCR			发散型		无结构型		线性型		聚合型	
Q-error	test length	g-s	2000	1000	2000	1000	2000	1000	2000	1000
10%	25	0.05-0.25	0.94	0.84	0.906	0.856	0.876	0.786	0.882	0.88
		0.05-0.4	0.813	0.677	0.827	0.73	0.786	0.652	0.786	0.79
	40	0.05-0.25	0.964	0.891	0.945	0.883	0.93	0.853	0.945	0.901
		0.05-0.4	0.859	0.794	0.76	0.624	0.825	0.7	0.863	0.77
20%	25	0.05-0.25	0.864	0.768	0.847	0.77	0.832	0.692	0.832	0.802
		0.05-0.4	0.66	0.497	0.722	0.486	0.692	0.522	0.646	0.634
	40	0.05-0.25	0.936	0.869	0.925	0.821	0.929	0.826	0.93	0.83
		0.05-0.4	0.785	0.709	0.648	0.438	0.743	0.595	0.781	0.664
30%	25	0.05-0.25	0.695	0.559	0.602	0.577	0.608	0.536	0.616	0.658
		0.05-0.4	0.541	0.307	0.419	0.279	0.441	0.371	0.451	0.437
	40	0.05-0.25	0.825	0.689	0.823	0.715	0.835	0.784	0.798	0.778
		0.05-0.4	0.663	0.569	0.483	0.334	0.618	0.527	0.634	0.525
属性缺失的 AACR			发散型		无结构型		线性型		聚合型	
Q-error	test length	g-s	2000	1000	2000	1000	2000	1000	2000	1000
10%	25	0.05-0.25	0.988	0.967	0.98	0.97	0.975	0.954	0.976	0.975
		0.05-0.4	0.958	0.93	0.964	0.942	0.956	0.915	0.952	0.952
	40	0.05-0.25	0.993	0.976	0.989	0.977	0.986	0.968	0.989	0.978
		0.05-0.4	0.968	0.953	0.938	0.911	0.958	0.928	0.971	0.948
20%	25	0.05-0.25	0.97	0.943	0.964	0.949	0.958	0.921	0.962	0.953
		0.05-0.4	0.913	0.867	0.935	0.872	0.899	0.873	0.91	0.91
	40	0.05-0.25	0.986	0.972	0.985	0.961	0.984	0.96	0.986	0.959
		0.05-0.4	0.948	0.927	0.908	0.849	0.932	0.897	0.945	0.918
30%	25	0.05-0.25	0.912	0.871	0.88	0.887	0.876	0.859	0.885	0.897
		0.05-0.4	0.864	0.782	0.818	0.779	0.793	0.805	0.829	0.829
	40	0.05-0.25	0.95	0.916	0.956	0.926	0.953	0.943	0.944	0.944
		0.05-0.4	0.908	0.877	0.844	0.798	0.886	0.863	0.896	0.863

表 4 属性冗余情境下 Q 矩阵的修正准确率(PCR 和 AACR)

Q-error	test length	g-s	属性冗余的 PCR				属性冗余的 AACR			
			发散型		无结构型		发散型		无结构型	
			2000	1000	2000	1000	2000	1000	2000	1000
10%	25	0.05-0.25	0.929	0.861	0.909	0.854	0.985	0.971	0.982	0.969
		0.05-0.4	0.81	0.678	0.826	0.711	0.957	0.929	0.964	0.937
	40	0.05-0.25	0.968	0.909	0.953	0.88	0.993	0.98	0.991	0.976
		0.05-0.4	0.876	0.791	0.865	0.783	0.974	0.954	0.973	0.955
20%	25	0.05-0.25	0.867	0.782	0.828	0.755	0.969	0.95	0.958	0.944
		0.05-0.4	0.651	0.761	0.704	0.504	0.909	0.944	0.928	0.876
	40	0.05-0.25	0.936	0.868	0.93	0.81	0.986	0.97	0.986	0.958
		0.05-0.4	0.814	0.695	0.8	0.703	0.955	0.925	0.958	0.935
30%	25	0.05-0.25	0.681	0.595	0.612	0.534	0.908	0.883	0.883	0.871
		0.05-0.4	0.502	0.581	0.449	0.254	0.852	0.88	0.827	0.772
	40	0.05-0.25	0.855	0.719	0.858	0.718	0.96	0.926	0.967	0.926
		0.05-0.4	0.616	0.575	0.61	0.478	0.891	0.884	0.887	0.854

3.2 模拟研究二

3.2.1 研究设计

为了进一步探查贝叶斯网算法在属性个数较多的情境下的表现,本研究设定属性个数为 7 个,根据模拟研究一的结果,各种类型的属性层级结构下,BN 的修正准确率相差不大,不失一般性,本研究随机设定属性层级结构为发散结构,7 个属性的测验长度固定为 40 个题目,样本量同研究一的设置,为 2000,1000。在模拟研究一中由于属性个数有限,所设定的属性关系缺失或冗余的水平仅限定为一个边,在本研究中进一步设定更多的属性缺失和冗余情况,在原有的如图 3 所示的 7 个发散属性结构的基础上共设定 5 种错误属性层级结构:(1)随机缺失一个边;(2)随机缺失两个边;(3)随机冗余一个边;(4)随机冗余两个边;(5)既有随机缺失边又有随机冗余边。其他实验条件的设定:题目质量, Q 矩阵错误率等完全与模拟一相同,模拟数据产生模型依然采用 GDINA 模型。

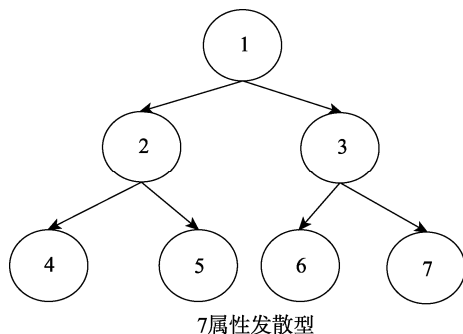


图 3 模拟研究设置的 7 个属性发散结构的属性层级关系

3.2.2 评价指标

在本研究中,主要考察在属性个数较多的情况下,不同属性缺失情况(5 个水平) × 不同题目质量(2 个水平) × 不同 Q 矩阵错误率水平(4 个水平) × 不同样本量水平(2 个水平) = 80 种情况下的检验修正准确率,评价指标同模拟研究一,依然采用 PCR 和 AACR,以及 100 次模拟中属性层级完全修正正确的次数。

3.2.3 研究结果

表 5 为 7 个属性的发散型结构的属性层级下,在样本量为 2000 时,5 种不同属性层级关系错误的情况下,各种实验情境下,贝叶斯网对 Q 矩阵的修正准确率(PCR, AACR)。由表可知,在各种类型的属性层级错误类型,其他条件相同的情况下,贝叶斯网对 Q 矩阵的修正准确率相差不大。在 Q 矩阵错误率为 10%时,无论题目质量高低,该方法对 Q 矩阵的修正准确率都达到很高的水平,在错误率为 20%时,题目质量较高的情况下, Q 矩阵修正准确率仍能够保持较高的水平,随着题目质量降低,修正准确率逐渐下降。当 Q 矩阵错误率为 30%时,修正准确率随着题目质量降低而逐渐降低。在 5 种不同属性错误类型下,贝叶斯网同时对属性层级完全修正正确的次数(100 次实验)如图 4 所示,当 Q 矩阵完全正确时,无论题目质量高低,贝叶斯网都能够完全正确的修正属性层级关系,之后随着 Q 矩阵错误率提高,题目质量降低,贝叶斯网对属性层级关系完全修正正确的次数逐渐下降。可以发现,即使在 7 个属性的诊断测验中,当 Q 矩阵错误率控制

在中等水平(20%以下)或者题目质量较好时, 贝叶斯网仍然能够有效的识别属性层级界定中的错误信息。表 6 为样本量 1000 时, 算法在各种实验条件下的表现, 相比于样本量 2000 的情况下, 各情境下的判准率稍有降低。图 5 为样本量 1000 时, 贝叶斯网在 100 次实验中, 对属性层级关系完全修正正确的次数, 可以看到相比于样本量为 2000 时的各种情境下, 算法的修正准确率稍有下降, 而且在样本量为 1000 时几种不同属性错误类型下的修正准确率之间的差异更大一些。

4 联合检验修正方法在实证数据中的应用

本部分进一步验证该方法在实证数据中的有效性, 该数据来自于美国大学生英语测验实际数据 (ECPE), Templin 和 Bradshaw (2014a)使用 ECPE 的数据研究并开发了 HDCM, 并试图采用 HDCM 发现并解释真实的 ECPE 数据的属性层级关系。ECPE 数据共 28 道题目, 测量三个属性: 词汇规则(a), 整合规则(b)语法规则(c)。根据 Templin 和 Bradshaw

表 5 7 个属性的发散结构下 Q 矩阵的修正准确率(样本量 = 2000)

PCR		缺失一个边	缺失两个边	冗余一个边	冗余两个边	既有缺失
Q-error	g-s					又有冗余
10%	0.05-0.25	0.849	0.859	0.836	0.861	0.846
	0.05-0.4	0.767	0.724	0.71	0.705	0.713
20%	0.05-0.25	0.739	0.693	0.739	0.728	0.716
	0.05-0.4	0.57	0.546	0.576	0.513	0.53
30%	0.05-0.25	0.51	0.511	0.578	0.51	0.555
	0.05-0.4	0.333	0.27	0.35	0.343	0.336

AACR		缺失一个边	缺失两个边	冗余一个边	冗余两个边	既有缺失
Q-error	g-s					又有冗余
10%	0.05-0.25	0.977	0.978	0.975	0.978	0.976
	0.05-0.4	0.964	0.953	0.953	0.952	0.955
20%	0.05-0.25	0.952	0.941	0.951	0.95	0.946
	0.05-0.4	0.914	0.912	0.919	0.904	0.909
30%	0.05-0.25	0.885	0.883	0.91	0.881	0.897
	0.05-0.4	0.815	0.834	0.839	0.842	0.839

表 6 7 个属性的发散结构下 Q 矩阵的修正准确率(样本量 = 1000)

PCR		缺失一个边	缺失两个边	冗余一个边	冗余两个边	既有缺失
Q-error	g-s					又有冗余
10%	0.05-0.25	0.728	0.745	0.728	0.735	0.646
	0.05-0.4	0.738	0.731	0.653	0.695	0.638
20%	0.05-0.25	0.632	0.626	0.475	0.635	0.484
	0.05-0.4	0.627	0.598	0.415	0.603	0.479
30%	0.05-0.25	0.434	0.431	0.408	0.428	0.19
	0.05-0.4	0.373	0.425	0.33	0.39	0.11

AACR		缺失一个边	缺失两个边	冗余一个边	冗余两个边	既有缺失
Q-error	g-s					又有冗余
10%	0.05-0.25	0.957	0.963	0.943	0.952	0.941
	0.05-0.4	0.96	0.959	0.958	0.958	0.94
20%	0.05-0.25	0.932	0.931	0.799	0.928	0.897
	0.05-0.4	0.931	0.928	0.887	0.929	0.897
30%	0.05-0.25	0.871	0.868	0.863	0.868	0.61
	0.05-0.4	0.851	0.863	0.835	0.861	0.527

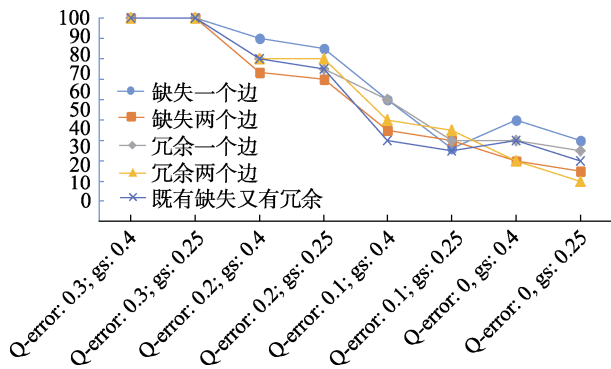


图 4 BN 对 7 个属性的发散结构属性层级关系在各种实验条件下的完全修正正确次数(样本量=2000)
(100 次实验完全修正正确的次数)

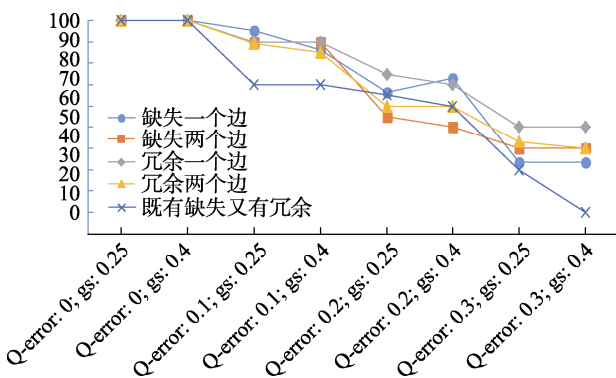


图 5 BN 对 7 个属性的发散结构属性层级关系在各种实验条件下的完全修正正确次数(样本量 = 1000)
(100 次实验完全修正正确的次数)

(2014a)的研究,这 3 个属性呈直线型关系,即 $a \rightarrow b \rightarrow c$ 。测验的 Q 矩阵和被试数据可以公开从 CDM R 软件包获得。

使用贝叶斯网进行检验修正的具体分析过程如下:先构建一个饱和的贝叶斯网,即 3 个属性之间都包含有向边(边的方向性依照初始建构的理论关系,即共有三条边, $a \rightarrow b$, $b \rightarrow c$, $a \rightarrow c$),每个题目与每个属性之间也包含有向边(不失一般性,方向性统一为属性指向题目),注意这个饱和模型的构建仅仅是为了完整的检验该测验中每一个可能存在的题目与属性以及属性与属性之间的联系,同模拟研究设置。根据已有的实证数据估计被试对 3 个认知属性的知识掌握状态,然后将被试作答反应数据和估计的知识掌握状态数据输入网络,对这一网络中的每一条边进行条件独立性检验。这个过程本质上也是对通过理论建构的模型(同时包含 Q 矩阵和属性层级关系)和实证数据之间的拟合程度进行检验的一个过程。通过假设检验得到的信息可以为进一步修正原有的理论建构的模型提供启示。但是在实证数据中,模型修正的过程并非完全依靠统

计检验的结果,也要结合领域专家对实际问题的判断。

使用贝叶斯网对初始建构的属性层级关系和 Q 矩阵进行联合检验后,发现原来设定的属性层级关系以及属性和题目之间的联结关系一部分得到了统计上的验证,但是也有一部分检验显示这些边的存在强度并不高,就是说两个节点之间可能是条件独立的,也有少量原来 Q 矩阵中没有发现的题目与属性之间的联系在经过贝叶斯网条件独立检验之后发现两个节点之间并非独立,而是具有依赖关系的,初次经过贝叶斯网条件独立检验发现属性之间的关系为: $a \rightarrow c$, $b \rightarrow c$ 。然后邀请专家再次对这些有争议存在的边进行讨论判定,包括属性层级的边。这样结合贝叶斯网条件独立性检验的结果,该认知诊断测验中所初始构建的 3 个认知属性之间的层级关系在经过统计检验和专家再次判定之后仍然保持原来的关系,但是 Q 矩阵有一些修正(具体修正过程信息可以联系作者提供)。接下来,根据贝叶斯网方法所验证的属性层级关系以及所修正的新的 Q 矩阵,重新估计被试的属性掌握模式。最后对初始建构的理论模型和经过检验修正之后的模型进行比较,以模型数据拟合指数进行评估。通过 BIC 拟合指标的数据显示经过统计检验和专家判定结合修正的新模型(属性层级保持原有关系, Q 矩阵有修正)与数据拟合是更好的(见表 7)。所以基于贝叶斯网的条件独立性检验的方法能够对属性层级关系和 Q 矩阵进行联合验证,并为最终优化模型提供有价值的信息。

表 7 BN 对实证数据的模型修正后的拟合指标

属性层级关系	Q 矩阵	BIC
$a \rightarrow b \rightarrow c$	原始	43815.53
$a \rightarrow c, b \rightarrow c$	BN 修正	44071.94
$a \rightarrow b \rightarrow c$	BN 修正	43724.51

5 结论与讨论

本研究提出了一种基于贝叶斯网条件独立性检验的属性层级关系和 Q 矩阵的联合检验方法。通过两个模拟研究系统的考察了样本量、Q 矩阵错误率、题目质量高低、测验长度、属性层级关系类型、属性标定错误类型,认知属性个数等因素对 BN 方法在检验 Q 矩阵和属性层级关系方面的准确性。结果表明,当 Q 矩阵界定完全正确时,在各种实验条件下, BN 方法都能够完全有效的修正属性层级关

系。当Q矩阵中的错误率控制在中等水平或以下(20%以下)时,在各种属性层级结构类型,各种属性层级错误类型下,BN方法都能够较好的同时修正属性层级关系和Q矩阵,随着Q矩阵错误率的增加,二者的修正准确率都逐渐下降,尤其当题目质量较低时,Q矩阵和属性层级关系的修正准确率都降到最低。除此之外,样本量对属性层级关系和Q矩阵的修正准确率也有明显的影响,在样本量为2000时各种情况下的修正准确率明显高于样本量为1000时的结果;测验长度对修正准确率的影响也非常明显,这主要是由于,当测验长度增加后,能够有更充分的题目证据估计初始的知识状态。并且该方法在属性个数较多的情况下,对Q矩阵和属性层级的修正结果依然较好,尤其是Q矩阵错误率较低和题目质量较高时。另一方面,不同的属性层级错误类型或者错误严重程度在一定程度上对修正准确率的影响较低。综合以上分析结果,本研究建议,在对属性层级关系和Q矩阵进行联合修正时要注意题目质量,并尽可能的保证样本量的充足,在保证题目质量的情况下尽量扩大测验长度。此外,虽然本研究也考察了在Q矩阵错误率较高时该算法的表现,但是通常情况下,在实际应用情境中,如果Q矩阵错误率达到30%~40%,一般建议重新标定Q矩阵。而在模拟实验中,设置高错误率的Q矩阵是为了更能直接考察方法的效能。

进一步的,为了充分探查该算法在实践应用中的可扩展性,本研究将该算法用于实证数据,结果表明BN算法能够为专家初始建构的属性层级关系和Q矩阵的修正提供有价值的参考信息,结合专家的最终判断,修正之后的新模型与实证数据的拟合指标更优。值得指出的是,与其他所有Q矩阵修正方法类似,即使在模拟研究中BN方法能够保证对属性层级关系和Q矩阵的修正准确性,但是在实践应用中,仍然不能完全依靠数据驱动的算法确定最后的属性层级关系和Q矩阵,这些修正算法能提供可靠的Q矩阵和属性层级关系的修正信息,但是对于一些有争议的题目或属性仍然需要领域内容专家的判断才能最后确定。

需要特别指出的是,在本研究中使用BN方法进行联合验证之前需要对被试的属性掌握模式做初始的估计,在本文的模拟研究部分采纳传统的认知诊断模型实现。在实践中,还需要根据实证数据恰当的选择诊断模型,如果不考虑数据的模型选择也可以使用贝叶斯网直接实现估计(Wang & Lu,

2021)。并且由于该算法的实现依赖传统认知诊断模型估计被试的属性掌握模式,所以本研究的模拟实验部分的样本量采用1000,2000。而在Wang和Lu(2021)的联合估计方法中所使用的最小样本量也是1000。在最新的只有Q矩阵估计或修正的相关研究中,设置了样本量较小情境下,例如N为500时算法的表现,在更小样本条件下检验方法的性能,能够更好地贴合实际应用场景。但是因为目前本研究的算法运行依赖传统认知诊断模型估计被试属性掌握模式,并且由于篇幅的限制,本研究作为一项初始的探索性研究,所考察的变量个数也有限,所采用的生成数据的模型只有GDINA模型,模型需要估计的参数较多,所以要达到较高精度的参数估计对样本量的需求较大。未来研究应继续考察该方法在简单模型(如DINA、DINO、ACDM等)中的效果,与此同时,可以考虑这些简单模型在样本量较小(如 $N=500$)时算法的性能。

此外本研究采用迭代的方式进行联合验证,但是根据前期预实验,这种迭代次数不是越多越好,迭代次数对修正准确率优化的效果边际效益递减,基本3次迭代后即可达到对属性层级和Q矩阵的稳定的修正结果,而在实证研究中,可以根据具体情况再设定具体的迭代次数,如果一次联合验证即可以提供非常丰富的修正信息,那么也可以使用一次修正的结果,然后结合专家的经验判断最后确定属性层级和Q矩阵。虽然BN方法实现的是对属性层级和Q矩阵的联合验证,但是该算法的计算效率仍然能够得到一定程度的保证,根据本研究的结果,测验长度为25,样本量为2000迭代3次对Q矩阵和属性层级进行一次联合验证的算法运行时间为123 s。根据李佳等(2022)对Q矩阵修正方法的研究中的报告,在测验长度为20,样本量为300时,基尼系数方法的循环修正运行时间为147 s,不考虑其他硬件条件等细微的差异,可以看到虽然本研究在Q矩阵修正基础上联合修正了属性层级关系,但是在算法运行时间上与同类的单独修正Q矩阵的算法仍然保持在一个同等的时间复杂度水平上。

最后,虽然已经有研究(喻晓锋等,2011)提出使用贝叶斯网结构学习K2算法进行数据驱动的属性层级关系探索,但是在其算法中,根据模拟得到的观察反应模式和测验项目,将被试正确作答的项目包含的属性向量进行“或”运算,得到所有被试的属性掌握模式,将它作为结构学习的数据集。这实际就是在根据Q矩阵和作答反应估计被试的属性

模式,但是这种估计方法比较初级,遇到复杂模型时不一定有效。而且这种方法估计的被试属性掌握模式之前,已经根据专家界定的属性层级关系编制题目确定 Q 矩阵,所以初始对被试的估计过程依靠的就是专家预先确定的属性层级关系。尽管已有少量研究提出了使用 LR 检验对属性层级进行验证的方法(Templin & Bradshaw, 2014a),但是该方法更多的是对属性层级进行整体的检验,通过模型比较的方式检验每一种假设属性层级的合理性,并没有对属性层级的修正准确率进行探查。而本研究可以对属性层级中任意两个认知属性之间的关系进行数据驱动的检验修正,而且这种修正是同时进行的,也就是说只要执行一次 BN 的条件独立性检验,能够得到所有联结边的合理性检验结果。未来研究中可以考虑在 Q 矩阵设定完全正确的情况下,比较本研究的 BN 方法和 LR 检验(Templin & Bradshaw, 2014a)对属性层级关系的修正效率。

目前,还没有研究提出能够对属性层级关系和 Q 矩阵进行联合验证的方法。有研究者提出不带任何先验信息的情况下从噪音数据中联合估计属性层级和 Q 矩阵(Wang & Lu, 2021; Ma et al., 2022),这两个研究都有着非常强的统计学理论意义,同时算法的复杂性也显而易见,所以其模拟研究都采用较少的属性个数(3~4 个)。而在认知诊断评估实践中,领域专家或者测验开发者通常对属性层级关系以及 Q 矩阵会提供相当的先验知识,所以面对实践需求,更需要一种能够同时检验和修正已有的属性层级关系和 Q 矩阵的方法,对已经提出的认知属性层级的理论模型和测验 Q 矩阵进行数据驱动的验证并根据实证的作答反应数据提供修正信息。所以,本研究并没有对 BN 方法的联合验证的准确性进行基线比较,因为目前还没有研究能够同时对属性层级关系和 Q 矩阵进行联合验证。但是本研究从另一个角度,在控制属性独立的情况下,对 BN 方法修正 Q 矩阵的准确性进行对比评估,结果表明,在属性独立的情况下,与已有的 Stepwise 方法(Ma & de la Torre, 2020)相比,BN 方法在 Q 矩阵错误率较高时更有优势(由于篇幅限制这部分的研究结果可以联系作者提供)。

参 考 文 献

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <http://dx.doi.org/10.1109/TAC.1974.1100705>
- Cui, Y. (2007). *The hierarchy consistency index: Development and analysis* [Unpublished doctoral dissertation]. University of Alberta.
- Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, 46(4), 429–449.
- Chiu, C.-Y. (2013). Statistical refinement of the Q-Matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37(8), 598–618. <https://doi.org/10.1177/0146621613488436>
- de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45(4), 346–362.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115–130.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179–199.
- de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, 81(2), 253–273.
- Ding, S. L., Mao, M. M., Wang, W. Y., Luo, F., & Cui, Y. (2012). Evaluating the consistency of test items relative to the cognitive model for educational cognitive diagnosis. *Acta Psychologica Sinica*, 44(11), 1535–1546.
- [丁树良, 毛萌萌, 汪文义, 罗芬, Cui, Y. (2012). 教育认知诊断测验与认知模型一致性的评估. *心理学报*, 44(11), 1535–1546.]
- Gu, Y., Liu, J., Xu, G., & Ying, Z. (2018). Hypothesis testing of the Q-matrix. *Psychometrika*, 83(3), 515–537.
- Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3), 197–243.
- Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(3), 191–210.
- Jiang, Y. (2020). *Research on the test method of attribute hierarchy based on information matrix* [Unpublished doctoral dissertation]. Beijing Normal University.
- [姜宇. (2020). 基于信息矩阵的属性层级关系检验方法研究 (博士学位论文). 北京师范大学.]
- Kang, C. H., Yang, Y. K., & Zeng, P. H. (2019). Q-matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement*, 43(7), 527–542.
- Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and application*. Cambridge: Cambridge University Press.
- Li, J., Mao, X., & Wei, J. (2022). A simple and effective new method of Q-matrix validation. *Acta Psychologica Sinica*, 54(8), 996–1008.
- [李佳, 毛秀珍, 韦嘉. (2022). 一种简单有效的 Q 矩阵修正新方法. *心理学报*, 54(8), 996–1008.]
- Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-Matrix. *Applied Psychological Measurement*, 36(7), 548–564.
- Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement*, 77(2), 220–240.
- Liu, Y., & Wu, Q. (2023). An empirical Q-matrix validation method using complete information matrix in cognitive diagnostic models. *Acta Psychologica Sinica*, 55(1), 142–158.
- [刘彦楼, 吴琼琼. (2023). 认知诊断模型 Q 矩阵修正: 完整信息矩阵的作用. *心理学报*, 55(1), 142–158.]

- Ma, C., Ouyang, J., & Xu, G. (2022). Learning latent and hierarchical structures in cognitive diagnosis models. *Psychometrika*, 88(1), 175–207.
- Ma, W., & de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology*, 73(1), 142–163.
- Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1), 12–35.
- Rupp, A. A., Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, 68(1), 78–96.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
- Scutari, M., & Denis, J. B. (2021). *Bayesian networks with examples in R*. New York: Chapman and Hall/CRC.
- Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, 10(1), 55–73.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345–354.
- Templin, J., & Bradshaw, L. (2014a). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 79(2), 317–339.
- Templin, J., & Bradshaw, L. (2014b). The use and misuse of psychometric models. *Psychometrika*, 79(2), 347–354.
- Terzi, R., & de la Torre, J. (2018). An iterative method for empirically-based Q-matrix validation. *International Journal of Assessment Tools in Education*, 5(2), 248–262.
- Tu, D. B., Cai, Y., & Dai, H. Q. (2012). A new method of Q-matrix validation based on DINA model. *Acta Psychologica Sinica*, 44(4), 558–568.
- [涂冬波, 蔡艳, 戴海琦. (2012). 基于 DINA 模型的 Q 矩阵修正方法. *心理学报*, 44(4), 558–568.]
- Wang, C., & Gierl, M. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, 48(2), 165–187.
- Wang, D., Gao, X., Cai, Y., & Tu, D. (2020). A method of Q-matrix validation for polytomous response cognitive diagnosis model based on relative fit statistics. *Acta Psychologica Sinica*, 52(1), 93–106.
- [汪大勋, 高旭亮, 蔡艳, 涂冬波. (2020). 基于类别水平的多级计分认知诊断 Q 矩阵修正: 相对拟合统计量视角. *心理学报*, 52(1), 93–106.]
- Wang, D. X., Cai, Y., & Tu, D. B. (2020). Q-matrix estimation methods for cognitive diagnosis models: Based on partial known Q-matrix. *Multivariate Behavioral Research*. <https://doi.org/10.1080/00273171.2020.1746901>
- Wang, D.-X., Gao, X.-L., Han, Y.-T., & Tu, D.-B. (2018). A simple and effective Q-matrix estimation method: From non-parametric perspective. *Journal of Psychological Science*, 41(1), 180–188.
- [汪大勋, 高旭亮, 韩雨婷, 涂冬波. (2018). 一种简单有效的 Q 矩阵估计方法开发: 基于非参数化方法视角. *心理科学*, 41(1), 180–188.]
- Wang, C., & Lu, J. (2021). Learning attribute hierarchies from data: Two exploratory approaches. *Journal of Educational and Behavioral Statistics*, 46(1), 1–27. <https://doi.org/10.3102/1076998620931094>
- Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, 113(523), 1284–1295.
- Xue, W., & Chen, H. G. (2012). *Data mining based on Clementine*. China Renmin University Press.
- [薛薇, 陈欢歌. (2012). *基于 Clementine 的数据挖掘*. 中国人民大学出版社.]
- Yu, X. F., & Cheng, Y. (2020). Data-driven Q-matrix validation using a residual-based statistic in cognitive diagnostic assessment. *British Journal of Mathematical and Statistical Psychology*, 73(Suppl 1), 145–179.
- Yu, X. F., Ding, S. L., Qin, C. Y., & Lu, Y. N. (2011). Application of Bayesian networks to identify hierarchical relation among attributes in cognitive diagnosis. *Acta Psychologica Sinica*, 43(3), 338–346.
- [喻晓锋, 丁树良, 秦春影, 陆云娜. (2011). 贝叶斯网在认知诊断属性层级结构确定中的应用. *心理学报*, 43(3), 338–346.]
- Zhang, L. W., & Guo, H. P. (2006). *Introduction to Bayesian networks*. Beijing: Science Press.
- [张连文, 郭海鹏. (2006). *贝叶斯网引论*. 北京: 科学出版社.]
- Zhang, X. Q., Jiang, Y., Xin, T., & Liu, Y. L. (2024). Iterative attribute hierarchy exploration methods for cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, <https://doi.org/10.3102/10769986241268906>

An approach that can validate both Q-matrices and attribute hierarchies in cognitive diagnosis models: From the empirical application perspective

WANG Ling-Ling¹, SUN Xiao-Jian²

⁽¹⁾ School of Educational Science, Shenyang Normal University, Shenyang 110034, China)

⁽²⁾ College of Teacher Education, Southwest University, Chongqing 400715, China)

Abstract

Cognitive diagnostic models (CDMs) are developed to diagnostically evaluate subjects' cognitive strengths and weaknesses based on the Q-matrix mapping of the items and attributes. The traditional calibration of cognitive attributes in the Q-matrix mainly relies on the subjective judgment of experts. Due to the subjective

process of Q-matrix construction, there inevitably are more or less misspecifications in the Q-matrix, which, if left unchecked, may result in a serious negative impact on cognitive diagnostic assessment. From another important perspective, in the empirical applications of CDMs, cognitive attributes generally do not operate independently but rather belong to an interrelated network, and a certain psychological order, logical order, or hierarchical relationship may be present among the cognitive attributes. The correctness of both the Q-matrix and the attribute hierarchy significantly impacts the parameter estimation ability of a CDM and the accuracy of the examinee's classification result. Recently, considerable studies have developed approaches for validating Q-matrices or testing attribute hierarchies respectively. However, there is no method that can validate both the Q-matrix and the attribute hierarchy simultaneously. From the empirical application perspective, an approach that can simultaneously validate both a prespecified Q-matrix and an attribute hierarchy is more desirable.

An approach based on Bayesian networks (BN) for validating both Q-matrices and attribute hierarchies simultaneously is proposed in this research. To explore the performance of the BN method, this article conducted two simulation studies and one empirical data analysis to theoretically and practically evaluate the accuracy of the Q-matrix validation and attribute hierarchy correction processes. The correctness of each element in the Q matrix and the attributes hierarchy can be checked by testing the strength of edge existence in the network structure.

When validating the attribute hierarchy relationships and the Q-matrix jointly in the first simulation, we explore the effects of Q-matrix error rate, item quality, test length, sample size, and the attribute hierarchy type on the correction accuracy of both the Q-matrix and the attribute hierarchy. The results show that the BN method can effectively correct the Q-matrix and the attribute hierarchy simultaneously when the error rate of the Q-matrix is at a medium or low level, especially when the item quality is high or the sample size is sufficient or the test length is long, the accuracy of the correction is generally high. As the Q-matrix error rate increases and the quality of the items decreases, the correction accuracy gradually decreases. The BN method can correct the attribute hierarchies exactly right when the Q matrix is correct. The results in the second simulation show that when the attribute number in the Q-matrix increases, the BN method is still performing well. Different types of attribute hierarchy errors have a small impact on the correction accuracy across different conditions. The effectiveness of the BN method in the empirical dataset was demonstrated by the better model data fit index of BIC.

In conclusion, the initial specified Q-matrix and attribute hierarchy can be simultaneously validated via the BN method. Then the corrected Q-matrix and the refined attribute hierarchy obtained from the data-driven BN method can again be combined with the theoretical judgments of experts to obtain a more optimized model, finally achieving more accurate diagnostic outcomes in CDA practice.

Keywords cognitive diagnosis, attribute hierarchy relationships, Q-matrix, Bayesian network