

方差—协方差矩阵在认知诊断中的作用

吴琼琼¹, 赵悦¹, 刘彦楼^{2*}

(1. 曲阜师范大学心理学院, 济宁 273165; 2. 曲阜师范大学教育大数据研究院, 济宁 273165)

摘要: 认知诊断模型(cognitive diagnostic model, CDM)的统计检验目前主要包括模型参数的标准误估计、项目功能差异检验、项目水平模型比较、Q 矩阵修正、属性层级关系探索等 5 个重要的研究领域。方差—协方差矩阵(信息矩阵的逆矩阵)在 CDM 的以上 5 种统计检验中具有基础和核心的作用。文章评述了方差—协方差矩阵在 CDM 的统计检验中的作用, 梳理了以往研究者提出的信息矩阵估计方法的发展思路和脉络。最后对已有研究存在的重要问题进行讨论和展望。

关键词: 认知诊断模型; 方差—协方差矩阵; 信息矩阵; 标准误; Q 矩阵修正

中图分类号: B841.2

文献标识码: A

文章编号: 1003–5184(2023)03–0262–07

1 引言

经典心理测量理论和项目反应理论中通常采用单一的测验分数来描述被试在某个阶段的学习效果, 作为新一代心理测量理论的认知诊断模型(cognitive diagnostic model, CDM)改变了这一传统形式, 这也是 CDM 近年来得到广泛关注和飞速发展的重要原因。CDM 采用现代统计方法和计算机技术, 诊断被试的潜在属性(例如认知结构、认知过程、知识、技能、策略、人格特质或心理障碍等), 为教师提供及时的反馈, 为个性化教学提供有效的干预和指导。目前, CDM 已广泛应用于心理、教育、精神病理学等领域(Sorrel et al., 2016)。

CDM 的统计检验包括模型参数的标准误(standard error, SE)估计、项目功能差异(differential item functioning, DIF)检验、项目水平模型比较、Q 矩阵修正、属性层级关系探索等 5 个重要的研究领域。信息矩阵的逆矩阵就是方差—协方差矩阵, 方差—协方差矩阵在 CDM 的统计检验中具有基础和核心的作用(von Davier & Haberman, 2014)。方差—协方差矩阵能够根据被试的作答反应数据估计模型参数, 对方差—协方差矩阵对角元素取平方根可以获得模型参数估计值的 SE, 用来描述模型参数估计值的不确定性(Philipp et al., 2018); 方差—协方差矩阵可以构建 Wald 统计量进行项目功能差异检验(Liu et al., 2019; Ma et al., 2021)、项目水平模型比较(Liu et al., 2019; Ma & de la Torre, 2019)、Q 矩阵修正(汪大勋等, 2020; Ma & de la Torre, 2020a)等。除此之外, 研究者(姜宇, 2020; Liu et al., 2021; Templin & Bradshaw, 2014)指出方差—协方差矩阵可以探索属性层级关系, 进而指导教学过程, 例如高一珠

等(2017)基于属性层级关系来探索学习进阶过程。

十几年来, 针对信息矩阵(或方差—协方差矩阵)在 CDM 的统计检验中的作用, 研究者开展了深入的研究, 并提出了很多信息矩阵估计方法, 主要包括: 考虑全部项目参数的不完整信息矩阵(de la Torre, 2009)、考虑单个项目参数的不完整信息矩阵(de la Torre, 2011)、完整的经验交叉相乘信息矩阵(empirical cross-product information matrix, XPD; 刘彦楼等, 2016; Philipp et al., 2018)、完整的观察信息矩阵(observed information matrix, Obs; 刘彦楼等, 2016; Liu et al., 2021)、完整的三明治信息矩阵(sandwich-type information matrix, Sw; Liu, Xin et al., 2019; Liu et al., 2021)等。厘清这些方法的发展思路和脉络、解析不同方法之间的联系、优劣, 评述不同方法在 CDM 统计检验中的表现, 不仅能够为研究者开发新的信息矩阵估计方法提供思路, 还能为实践者探索信息矩阵的新应用提供依据。

基于此, 文章从模型参数的 SE 估计、DIF 检验、模型比较、Q 矩阵修正以及属性层级关系探索这 5 个 CDM 的主要研究领域出发, 详细评述了方差—协方差矩阵在统计检验中的作用, 然后对现有的信息矩阵估计方法进行综合分析, 最后对已有研究存在的重要问题进行讨论和展望。

2 方差—协方差矩阵在 CDM 的统计检验中的作用

2.1 模型参数的 SE 估计

在参数 CDM 中, 通过对方差—协方差矩阵对角线元素进行开方运算, 可获得项目参数 β 和结构参数 η (structural parameter) 的 SE 的估计值。但是, 仅获得模型参数的 SE 的估计值是不够的, 需要

* 通讯作者: 刘彦楼, E-mail: liuyanlou@163.com。

综合考虑置信区间(confidence interval, CI)来获得更准确的信息(汪文义等, 2020)。具体而言, 通过模型参数的 SE 与 z 检验结合获得模型参数估计值的 CI:

$$[\hat{\beta} \pm z_{\frac{\alpha}{2}} \cdot SE(\hat{\beta})] \quad (1)$$

以及

$$[\hat{\eta} \pm z_{\frac{\alpha}{2}} \cdot SE(\hat{\eta})] \quad (2)$$

其中, $\hat{\beta}$ 表示项目参数估计值, $SE(\hat{\beta})$ 表示项目参数的 SE ; $\hat{\eta}$ 表示结构参数估计值, $SE(\hat{\eta})$ 表示结构参数的 SE ; $z_{\frac{\alpha}{2}}$ 表示 $(1 - \frac{\alpha}{2})$ 的置信区间对应的标准正态分布。显著性水平一定时, SE 越小, CI 越小, 模型参数的估计值的离散程度越小, 更容易获得准确的估计值。

模型参数的 SE 估计的准确性影响着研究结果的准确性, 研究者(Liu, Xin et al., 2019; Philipp et al., 2018)在不同模型设定条件下进行了研究, 认为仅考虑项目参数的不完整信息矩阵会低估模型参数的 SE , 仅考虑单个项目参数的不完整信息矩阵会最大程度的低估模型参数的 SE , 完整信息矩阵计算的 SE 与不完整信息矩阵计算的 SE 相比, 偏差较小。

2.2 项目功能差异检验

DIF 指的是来自不同组但具有相同属性掌握模式的被试正确回答某个项目的概率不同(Hou et al., 2014)。例如, 测验中考察同一年级的男生和女生的阅读能力, 若呈现的是有关篮球的题目, 男生可能会因为接触篮球较多而表现出更高的水平, 那么该项目对这两个组来说是不公平的, 即该项目存在 DIF(魏丹等, 2020)。

采用方差-协方差矩阵构建 Wald 统计量用于 DIF 检验的原理是: 检验不同组被试对应的模型参数是否相同(王卓然等, 2014)。Hou 等(2014)在 DINA 模型(the deterministic inputs, noisy “and” gate model)下, 提出了使用 Wald 统计量来检验 DIF:

$$Wald = (R\hat{\beta}_j)'(R\hat{\Sigma}_jR')^{-1}(R\hat{\beta}_j) \quad (3)$$

其中, $\hat{\beta}_j = \hat{\beta}_{Rj} - \hat{\beta}_{Fj}$ 表示对照组与目标组中估计获得的模型参数估计值之差; R 是在不同模型框架下研究者自行构建的约束矩阵; $\hat{\Sigma}_j$ 是项目参数的方差-协方差矩阵。 $\hat{\Sigma}_j$ 是 Wald 统计量的核心。

采用 Wald 统计量进行 DIF 检验时, 有较高的统计检验力, 且能够检验出一致性 DIF 和非一致性的 DIF(王卓然等, 2014; Hou et al., 2020)。但是, 在部分研究(王卓然等, 2014; Hou et al., 2014; Hou et al., 2020)中, 用于 DIF 检验的 Wald 统计量采用的是不完整信息矩阵, 由于没有考虑结构参数, 存在

一类错误控制率的膨胀。基于此, 研究者(刘彦楼等, 2016; Liu, Yin et al., 2019; Ma et al., 2021)采用模拟研究比较了基于完整和不完整信息矩阵构建的 Wald 统计量在检验 DIF 时的表现, 发现完整信息矩阵构建的 Wald 统计量在检验 DIF 时, 在一类错误控制率和统计检验力方面有更好的表现。

2.3 项目水平模型比较

目前已经有 100 多种 CDM, 包括一般性的 CDM 和特殊的 CDM, 对一般性的 CDM 适当加以约束可以得到特殊 CDM。虽然一般性的 CDM 能够更好的拟合数据, 但主要存在以下问题: 首先, 参数过多使模型难以解释; 其次, 精确估计参数需要的样本量大; 最后, 属性分类准确性较低。而特殊模型在样本量较小的时候能获得较高的属性分类准确性, 所以在实践中选择模型—数据拟合良好的特殊模型是至关重要的。

De la Torre 和 Lee(2013)提出在 G-DINA 模型(the generalized DINA model)框架下, 使用 Wald 统计量进行项目水平的模型选择, 原理是: 可以在项目水平上检验所测量属性数目大于 1 的项目, 在对模型的整体拟合不造成显著影响的前提下, 将饱和 CDM 缩减为特殊 CDM。其形式为:

$$Wald = (R\hat{\beta}_j)'(R\hat{\Sigma}_jR')^{-1}(R\hat{\beta}_j) \quad (4)$$

其中, $\hat{\beta}_j$ 是项目 j 的项目参数极大似然估计向量。

Wald 统计量不仅能够识别出特殊模型, 而且能提高分类准确性(Ma & de la Torre, 2016)。但是, 部分研究(de la Torre & Lee, 2013; Ma & de la Torre, 2016)构建的 Wald 统计量存在一类错误控制率的膨胀。采用完整信息矩阵构建 Wald 统计量用于项目水平模型比较时, Wald 统计量表现出较为稳健的一类错误控制率和统计检验力(刘彦楼等, 2019; Liu, Andersson et al., 2019)。

2.4 Q 矩阵修正方法

Q 矩阵反映了测验的内部结构和内容设计, 通常由领域专家根据经验进行主观界定, 故原始 Q 矩阵中可能存在一些错误设定。Q 矩阵错误设定会降低模型参数估计的准确性, 产生较差的模型—数据拟合, 并导致错误的属性估计(Chiu, 2013)。

Ma 和 de la Torre(2020a)在 seq-GDINA 模型(the sequential GDINA model)中提出了 Stepwise 方法进行 Q 矩阵修正。Stepwise 方法首先采用 GDI 方法(de la Torre & Chiu, 2016)从单一属性的 q 向量中确定第一个所需属性, 再逐步多次采用 Wald 统计量决定是否增加或删除属性来选择正确的 q 向量。Wald 统计量用于 Q 矩阵修正的原理是: 假设项目 j 所对应的 q 向量定义了 2 个及以上的属性, 将

某一属性从 \mathbf{q} 向量中移除而没有导致模型一数据拟合变差,则这个属性就不是必需的。Wald 统计量的形式为:

$$\text{Wald} = [\mathbf{R} \mathbf{p}_j(\alpha_l)]' (\mathbf{R} \mathbf{V}_j \mathbf{R}')^{-1} [\mathbf{R} \mathbf{p}_j(\alpha_l)] \quad (5)$$

其中, $\mathbf{p}_j(\alpha_l)$ 表示拥有属性掌握模式为 α_l 的被试在项目 j 中正确作答概率的向量; \mathbf{V}_j 是项目正确作答概率的方差-协方差矩阵。

Ma 和 de la Torre (2019) 认为,虽然基于完整信息矩阵的 Wald 统计量更加准确,但计算时难度较大,故 Stepwise 方法中构建的 Wald 统计量采用不完整信息矩阵求逆获得的方差-协方差矩阵。但文章认为 Stepwise 方法在正确修正错误标定属性方面表现较差可能是因为 Wald 统计量的计算中采用了不完整的信息矩阵,故可以尝试采用不同类型的完整信息矩阵构建 Wald 统计量用于 Q 矩阵修正,与 Stepwise 方法进行比较。

Stepwise 方法用于 Q 矩阵修正一经提出,就引起了广泛的关注。例如,汪大勋等 (2020) 将相对拟合指标引入 seq - GDINA 模型中,发现相对拟合指标和 Stepwise 方法在类别水平的多值计分模型中均有较好的 Q 矩阵修正效果。此外,研究者可以尝试在更多情境下考察 Wald 统计量用于 Q 矩阵修正的表现,如属性个数有误时如何自动识别等,在多值计分认知诊断模型中开发更多新的 Q 矩阵修正方法与 Stepwise 方法进行比较也很有价值。

2.5 属性层级关系探索

研究者 (Leighton et al., 2004) 认为,属性可以以特定的形式形成层级结构,因为在心理与教育实践中一些属性需要建立在其他属性的基础上才能掌握。故准确识别 CDM 中属性层级关系能够使研究者深入地理解被试的心理过程,是成功进行认知诊断的关键。

在预先没有确定属性之间的层级关系的前提下,基于方差-协方差矩阵构建的 z 统计量可以用来探索属性层级关系 (姜宇, 2020; Liu et al., 2021; Templin & Bradshaw, 2014)。它的原理是:对每一个结构参数是否为 0 进行假设检验,确定哪些结构参数是不允许存在的,获得可能的属性掌握模式,达到对属性掌握模式的内部结构进行判断的目的。 z 统计量的形式为:

$$z = \frac{\hat{\eta}}{SE(\hat{\eta})} \quad (6)$$

Liu 等 (2021) 的研究表明 z 统计量与似然比检验 (Templin & Bradshaw, 2014) 获得的结果相似。即基于方差-协方差矩阵构建的 z 统计量在探索属性层级关系中有较好的表现。但遗憾的是,目前关于

z 统计量探索属性层级关系的研究较少,未来研究者可以尝试在不同属性层级下 (例如,线型、金字塔型等),采用不同类型的信息矩阵考察 z 统计量在探索属性层级关系中的表现。

3 信息矩阵估计方法

3.1 结构参数及其表达形式

CDM 中,研究者 (Liu et al., 2016; Philipp et al., 2018) 指出 CDM 中同时存在两种类型的参数:项目参数和结构参数。结构参数指的是被试总体的潜在属性掌握模式分布状况。例如,测验中共有 2 个属性 $K = 2$,那么所有可能的属性掌握模式有 4 种,即:

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \quad (7)$$

在这个测验中被试所有可能的属性掌握模式 (α_l) 有 $L = 4$ 种; $p(\alpha_l)$ 为被试总体中拥有第 l 种属性掌握模式 α_l 的分布比例,例如 $p(\alpha_1)$ 是被试总体中具有第 1 种属性掌握模式 $\alpha_1 = (0, 0)'$ 的分布比例。

结构参数的表达形式有两种:第一种是 Liu 等人 (2016) 将 CDM 中的结构模型 (即用于描述被试所有可能的属性掌握模式分布的模型) 表达为:

$$p(\alpha_l) = p(\alpha_l | \eta) = \frac{\exp(\eta_l)}{\sum_{l=1}^L \exp(\eta_l)} \quad (8)$$

上式中, $p(\alpha_l | \eta)$ 表示从总体中随机抽取的一个被试拥有第 l 种属性掌握模式的概率, $\eta = (\eta_1, \dots, \eta_L)'$ 。因为全部的属性掌握模式概率之和为 1,将最后一个属性掌握模式对应的参数约束为 $\eta_L = 0$ 。

第二种是 Philipp 等 (2018) 直接使用结构参数 π 表示被试在各个属性掌握模式上的分布比例,并用向量 $\pi = (\pi_1, \dots, \pi_L)'$ 表示:

$$\pi_l = p(\alpha_l) \quad (9)$$

根据信息矩阵是否考虑结构参数、以及结构参数表达形式的不同,可以将其分为不同的类型。

3.2 不完整的信息矩阵

De la Torre (2009) 最先提出了 CDM 中关于全部项目参数估计值的信息矩阵估计方法 \mathcal{J}_{D09} :

$$\mathcal{J}_{D09} = \frac{\partial \mathcal{A}(\mathbf{x})}{\partial \boldsymbol{\beta}} \cdot \frac{\partial \mathcal{A}(\mathbf{x})}{\partial \boldsymbol{\beta}'} \quad (10)$$

其中, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_J)'$ 为模型中的全部项目参数, J 表示所有项目的数量。 \mathbf{x} 表示所有被试在所有项目的作答反应矩阵, $\mathcal{A}(\mathbf{x}) = \sum_{i=1}^I \log \sum_{l=1}^L L(\mathbf{x}_i |$

$\alpha_i)p(\alpha_i)$ 表示的是作答反应矩阵的对数似然函数, $L(x_i | \alpha_i)$ 表示属性掌握模式为 α_i 的被试 $i \in (1, \dots, I)'$ 的作答反应向量 x_i 的似然。

De la Torre(2011)提出仅考虑了单个项目参数的信息矩阵估计方法 \mathcal{J}_{D11} :

$$\mathcal{J}_{D11} = \frac{\partial \mathcal{A}(x)}{\partial \beta_j} \cdot \frac{\partial \mathcal{A}(x)}{\partial \beta'_j} \quad (11)$$

其中, β_j 是第 j 个项目的项目参数向量。由于 \mathcal{J}_{D11} 仅考虑了单个项目参数,所以是 \mathcal{J}_{D09} 的子矩阵。

3.3 完整的信息矩阵

3.3.1 期望信息矩阵

以公式中结构参数 η 的表达式为基础, Liu 等人(2016)提出了 CDM 中最为基础的信息矩阵——期望(费舍)信息矩阵 \mathcal{J}_E :

$$\mathcal{J}_E = E \left[\frac{\partial \mathcal{A}(x_i)}{\partial \gamma} \cdot \frac{\partial \mathcal{A}(x_i)}{\partial \gamma'} \right] \quad (12)$$

需要指出的是,尽管期望信息矩阵是度量模型参数估计值准确性的“黄金法则”(Tian et al., 2013),但 \mathcal{J}_E 仅具有理论上的价值。首先,期望信息矩阵 \mathcal{J}_E 需要求出关于被试 i 作答反应的期望,即计算所有可能的作答反应模式,但会随着项目数量的增多而呈现指数增长,项目量较大条件下无法计算。其次, \mathcal{J}_E 需要通过模型参数真值计算;然而,在实践中模型参数的真值是未知的。故实践中使用观察到的全部被试的作答反应矩阵取代对单个被试的作答反应求期望;使用模型参数的极大似然估计值取代模型参数的真值。

3.3.2 经验交叉相乘信息矩阵

刘彦楼等(2016)最早提出了同时考虑项目参数 β 和结构参数 η 的正确信息矩阵估计方法。他们在期望信息矩阵 \mathcal{J}_E 的基础上提出基于全部模型参数估计值及被试作答反应矩阵的信息矩阵估计方法 $\mathcal{J}_{XPD-\eta}$:

$$\mathcal{J}_{XPD-\eta} = \frac{\partial \mathcal{A}(x)}{\partial \gamma} \cdot \frac{\partial \mathcal{A}(x)}{\partial \gamma'} \quad (13)$$

$\mathcal{J}_{XPD-\eta}$ 是通过观察数据对数似然函数关于模型参数一阶导向量交叉相乘而获得,故称作经验交叉相乘信息矩阵。

Philipp 等(2018)同样认为应该将项目参数 β 和结构参数 π 均包含在信息矩阵中。为从理论上证明仅包含单个或全部项目参数的信息矩阵会低估模型参数的 SE ,他们将 XPD 信息矩阵 $\mathcal{J}_{XPD-\pi}$ 表达为以下 4 个部分的组合:

$$\mathcal{J}_{XPD-\pi} = \begin{pmatrix} \mathcal{J}_{\beta,\beta} & \mathcal{J}_{\beta,\pi} \\ \mathcal{J}_{\pi,\beta} & \mathcal{J}_{\pi,\pi} \end{pmatrix} \quad (14)$$

其中, $\mathcal{J}_{\beta,\beta}$ 是仅包含全部项目参数的信息矩阵,等价于 de la Torre (2009) 提出的信息矩阵估计方法

\mathcal{J}_{D09} 。结合公式可以发现,结构参数 π 构建的 XPD 矩阵能够以属性掌握模式分布比例的形式直接给出结构参数 SE 的估计值,便于后续进行统计检验(Liu et al., 2021)。

3.3.3 观察信息矩阵

刘彦楼等人(2016)也同时提出了观察信息矩阵 $\mathcal{J}_{Obs-\eta}$:

$$\mathcal{J}_{Obs-\eta} = - \frac{\partial^2 \mathcal{A}(x)}{\partial \gamma \partial \gamma'} \quad (15)$$

$\mathcal{J}_{Obs-\eta}$ 是通过负的观察数据对数似然函数关于模型参数二阶偏导数而计算的。

3.3.4 三明治信息矩阵

Liu, Xin 等(2019)在 Liu 等(2016)提出期望信息矩阵 \mathcal{J}_E 以及刘彦楼等(2016)提出的经验交叉相乘信息矩阵 $\mathcal{J}_{XPD-\eta}$ 以及观察信息矩阵 $\mathcal{J}_{Obs-\eta}$ 的基础上,对于信息矩阵的估计理论及具体实现方法进行了研究,并且提出了新的信息矩阵估计方法,即三明治信息矩阵 $\mathcal{J}_{Sw-\eta}$:

$$\mathcal{J}_{Sw-\eta}^1 = \mathcal{J}_{Obs-\eta} \mathcal{J}_{XPD-\eta}^1 \mathcal{J}_{Obs-\eta} \quad (16)$$

$\mathcal{J}_{Sw-\eta}$ 中间是 $\mathcal{J}_{XPD-\eta}$, 两边是 $\mathcal{J}_{Obs-\eta}$, 其形状类似三明治。

特别地, $\mathcal{J}_{XPD-\eta}$ 、 $\mathcal{J}_{Obs-\eta}$ 和 $\mathcal{J}_{Sw-\eta}$ 信息矩阵的主要不足在于采用了公式中结构参数 η 表达式,无法以属性掌握模式分布比例的形式直接计算 SE 的估计值。因此, Liu 等(2021)以结构参数 π 的表达替代了结构参数 η 的表达式,提出了改进的 Obs 和 Sw 矩阵,分别记作 $\mathcal{J}_{Obs-\pi}$ 和 $\mathcal{J}_{Sw-\pi}$:

$$\mathcal{J}_{Obs-\pi} = - \begin{bmatrix} \frac{\partial^2 \mathcal{A}(\hat{\gamma} | x)}{\partial \beta \partial \beta'} & \frac{\partial^2 \mathcal{A}(\hat{\gamma} | x)}{\partial \beta \partial \pi'} \\ \frac{\partial^2 \mathcal{A}(\hat{\gamma} | x)}{\partial \pi \partial \beta'} & \frac{\partial^2 \mathcal{A}(\hat{\gamma} | x)}{\partial \pi \partial \pi'} \end{bmatrix} \quad (17)$$

以及

$$\mathcal{J}_{Sw-\pi}^1 = \mathcal{J}_{Obs-\pi} \mathcal{J}_{XPD-\pi}^1 \mathcal{J}_{Obs-\pi} \quad (18)$$

3.4 信息矩阵估计方法的综合评价

不完整信息矩阵仅考虑项目参数,是完整信息矩阵的子矩阵(Philipp et al., 2018),计算量较小,在后续研究中运算速度较快。但是,由于没有考虑结构参数,求逆获得方差-协方差矩阵时会造成较大的误差,在模型参数的 SE 估计、DIF 检验、模型比较、Q 矩阵修正的研究中存在不够准确的问题。故不推荐研究者使用不完整信息矩阵。

期望信息矩阵 \mathcal{J}_E 基于模型参数真值,在实践中真值往往是未知的,计算量随着项目数量呈指数增长,在项目数较大时难以计算,所以说, \mathcal{J}_E 仅具有理论上的价值。XPD 矩阵的优势在于计算简单、省时,仅需通过计算一阶导数就可以获得,但在模型错

误设定时,可能会存在准确性较低的问题。与 XPD 矩阵相比,Obs 及 Sw 矩阵($\mathcal{I}_{Obs-\eta}$ 、 $\mathcal{I}_{Sw-\eta}$)用于统计检验时在大多数情况下具有更好的表现,但是 Obs 矩阵以及 Sw 矩阵涉及观察数据的对数似然函数关于所有模型参数的二阶偏导,公式推导复杂繁琐。需要指出的是,由于测验情境较为复杂,这三种方法各有优势。例如,刘彦楼等(2019)在 Q 矩阵错误设定的背景下研究模型选择时,认为当模型拟合良好或错误设定较多时推荐使用 $\mathcal{I}_{Sw-\eta}$,当 Q 矩阵中可能存在错误设定且样本量大于 500 时推荐使用 $\mathcal{I}_{XPD-\eta}$ 。因此,实践者在选择不同的信息矩阵时要结合测验情境以及可能面临的现实问题。改进的 Obs 及 Sw 矩阵($\mathcal{I}_{Obs-\pi}$ 、 $\mathcal{I}_{Sw-\pi}$),可以同时获得项目参数的 SE 和以属性掌握模式分布为直接表达式的结构参数的 SE,但它们的缺点也是计算量较大,用于后续统计检验时耗时较长。如何提高这两种信息矩阵的计算速度,值得研究者们进一步探讨。

表 1 CDM 中信息矩阵估计方法综合比较

方法	是否考虑结构参数 或结构参数符号	特点	属性掌握模式 分布的 SE
\mathcal{I}_{D09}	否	计算量小	无
\mathcal{I}_{D11}	否	计算量小	无
\mathcal{I}_E	η	需要模型真值 计算量指数增长	—
$\mathcal{I}_{XPD-\eta}$	η	计算量大	无
$\mathcal{I}_{Obs-\eta}$	η	涉及二阶偏导 计算量大	无
$\mathcal{I}_{XPD-\pi}$	π	计算量大	有
$\mathcal{I}_{Sw-\eta}$	η	计算量大	无
$\mathcal{I}_{Obs-\pi}$	π	涉及二阶偏导 计算量大	有
$\mathcal{I}_{Sw-\pi}$	π	计算量大	有

注: \mathcal{I}_{D09} 表示考虑全部项目参数的不完整信息矩阵; \mathcal{I}_{D11} 表示考虑单个项目参数的不完整信息矩阵; \mathcal{I}_E 表示期望(费舍)信息矩阵; $\mathcal{I}_{XPD-\eta}$ 表示刘彦楼等(2016)完整经验交叉相乘信息矩阵; $\mathcal{I}_{Obs-\eta}$ 表示刘彦楼等(2016)完整观察信息矩阵; $\mathcal{I}_{XPD-\pi}$ 表示 Philipp 等(2018)完整经验交叉相乘信息矩阵; $\mathcal{I}_{Sw-\eta}$ 表示 Liu, Xin 等(2019)完整三明治信息矩阵; $\mathcal{I}_{Obs-\pi}$ 和 $\mathcal{I}_{Sw-\pi}$ 表示 Liu 等(2021)改进的观察信息矩阵和三明治信息矩阵。 η 和 π 为结构参数。因为 \mathcal{I}_E 只有理论上的价值,无法在实践中应用,“—”表示不适用。

4 问题与展望

文章评述了方差—协方差矩阵在 CDM 的 5 种重要的统计检验中的作用,梳理了以往研究者提出的信息矩阵估计方法的发展思路和脉络。纵观国内外研究,方差—协方差矩阵的理论和实践探索已经比较丰富,但仍有一些地方需要进一步完善,未来的研究方向主要集中在以下几个方面。

4.1 开发新的方差—协方差矩阵估计方法

DeCarlo(2019)认为信息矩阵的完整表达形式

会受到边界值问题的影响并提出采用 PME (posterior mode estimation) 算法解决这一问题。PME 算法下的各种信息矩阵估计获得的方差—协方差矩阵的表现值得研究者们进一步探讨。Philipp 等(2018)认为在 logit 链接下,极大似然估计值更加渐进模型真值。因此,今后可以尝试基于不同模型、不同算法、不同链接函数探索影响信息矩阵估计方法准确性的因素。目前大多数研究中方差—协方差矩阵是通过信息矩阵求逆获得的,在一定程度上来说,计算较为复杂,不容易理解。故可以尝试将其他潜在的方法应用到 CDM 的研究中,例如 Oakes 方法(Chalmers, 2018)、数值微分法(Tian et al., 2013)、Jackknife 方法(Wainer & Wright, 1980)、补充的 EM 算法(supplemented expectation maximization, SEM; Cai, 2008)等。

4.2 基于方差—协方差矩阵的新统计检验方法开发

方差—协方差矩阵在 CDM 统计检验方法开发中的潜力仍有待进一步开发。下面以三个简单的具体例子进行说明:(1)研究者使用提出方差—协方差矩阵计算结构参数的 SE,并以此构建 z 统计量用于探索属性层级关系(Liu et al., 2021);同理,当获得所有项目参数的 SE 后,也可以使用关于项目参数的 z 统计量对其显著性进行检验,以此达到 CDM 的参数水平上模型选择的目的。(2)Wald 统计量用于 DIF 以及模型比较的表现还依赖于模型(王卓然等, 2014),研究者们也开发了很多能用于多级计分的 CDM,如 P-DINA 模型(涂冬波等, 2010)等,在多级计分模型中考察不同类型的信息矩阵构建的 Wald 统计量在 DIF 检验和项目水平模型比较中的表现也很有意义。(3)在 Q 矩阵估计或修正方法中,以往研究者在构建 Wald 统计量时使用的大多是不完整的信息矩阵(\mathcal{I}_{D09} 或 \mathcal{I}_{D11})计算的方差—协方差矩阵;研究发现完整信息矩阵的表现优于不完整信息矩阵。基于此,可以使用 $\mathcal{I}_{XPD-\pi}$ 、 $\mathcal{I}_{Obs-\pi}$ 及 $\mathcal{I}_{Sw-\pi}$ 矩阵构建 Wald 统计量用于 Q 矩阵估计或修正。基于这类新的 Wald 统计量用于 Q 矩阵估计或者修正的表现有很大可能会优于使用不完整信息矩阵构建的 Wald 统计量。

4.3 探索方差—协方差矩阵的实践应用

在心理统计与测量模型中,点估计值相同的两个模型参数可能由于 SE 和 CI 的不同而具有不同的估计精度。但在 CDM 的研究中报告 SE 及 CI 的研究仍然较少。部分原因在于缺乏简便易用的估计软件。举例而言,当前在 R 语言中有三个软件包可以输出 SE:CDM 软件包(George et al., 2016; Robitzsch & George, 2019)、GDINA 软件包(Ma & de la Torre,

2020b)以及 *dcmiffo* 软件包(Liu & Xin, 2017)。但在这三个软件包中, *CDM* 软件包输出的 *SE* 是用不完整信息矩阵估计的, 是错误的; *dcmiffo* 软件包可以输出 $\mathcal{I}_{XPD-\eta}$ 、 $\mathcal{I}_{Obs-\eta}$ 、 $\mathcal{I}_{Sw-\eta}$ 计算的 *SE*, 但是其结构参数采用的是公式的表达式, 无法直接输出属性掌握模式分布的 *SE*; *GDINA* 软件包可以输出不完整信息矩阵的 *SE*, 也可以输出使用 $\mathcal{I}_{XPD-\pi}$ 计算且采用公式的结构参数, 但是并没有包含 *Obs* 及 *Sw* 矩阵。因此, 就目前而言, 推荐使用 *dcmiffo* 软件包中的 $\mathcal{I}_{Obs-\eta}$ 和 $\mathcal{I}_{Sw-\eta}$ 估计项目参数的 *SE*, 使用 *GDINA* 软件包中的 $\mathcal{I}_{XPD-\pi}$ 估计结构参数的 *SE*。所以, 结合 R 软件更好的估计 CDM 模型参数的 *SE* 和 *CI* 具有重要的实践意义。例如, 刘彦楼(2022)开发出并行自助法用于估计 CDM 中模型参数的 *SE* 及 *CI*。

参考文献

- 高一珠, 陈孚, 辛涛, 詹沛达, 姜宇. (2017). 心理测量学模型在学习进阶中的应用: 理论、途径和突破. *心理科学进展*, 25(9), 1623–1630.
- 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, 北京.
- 刘彦楼. (2022). 认知诊断模型的标准误与置信区间估计: 并行自助法. *心理学报*, 54(6), 1–22.
- 刘彦楼, 辛涛, 李令青, 田伟, 刘笑笑. (2016). 改进的认知诊断模型项目功能差异检验方法——基于观察信息矩阵的 Wald 统计量. *心理学报*, 48(5), 588–598.
- 刘彦楼, 张倩萌, 郑宗军, 尹昊. (2019). 认知诊断模型中项目水平模型比较统计量的健壮性. *心理科学*, 42(5), 1251–1259.
- 涂冬波, 蔡艳, 戴海琦, 丁树良. (2010). 一种多级评分的认知诊断模型: P-DINA 模型的开发. *心理学报*, 42(10), 1011–1020.
- 魏丹, 张丹慧, 刘红云. (2020). 基于多维题组反应模型的项目功能差异检验探究. *心理科学*, 43(1), 206–214.
- 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.
- 汪文义, 朱黎君, 叶宝娟, 方小婷. (2020). Bootstrap 区间估计在认知诊断模型误设中的应用. *心理科学*, 43(6), 1498–1505.
- 王卓然, 郭磊, 边玉芳. (2014). 认知诊断测验中的项目功能差异检测方法比较. *心理学报*, 46(12), 1923–1932.
- Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, 61(2), 309–329.
- Chalmers, R. P. (2018). Numerical approximation of the observed information matrix with Oakes' identity. *British Journal of Mathematical and Statistical Psychology*, 71(3), 415–436.
- Chiu, C. – Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37(8), 598–618.
- 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.
- de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, 50(4), 355–373.
- DeCarlo, T. (2019). Insights from reparameterized DINA and beyond. In M. von Davier & Y. – S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 549–572). Springer.
- George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Uenlue, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, 74(2), 1–24.
- Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, 51(1), 98–125.
- Hou, L., Terzi, R., & de la Torre, J. (2020). Wald test formulations in DIF detection of CDM data with the proportional reasoning test. *International Journal of Assessment Tools in Education*, 7(2), 145–158.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoaka's rule-space approach. *Journal of Educational Measurement*, 41(3), 205–237.
- Liu, Y., Andersson, B., Xin, T., Zhang, H., & Wang, L. (2019). Improved Wald statistics for item-level model comparison in diagnostic classification models. *Applied Psychological Measurement*, 43(5), 402–414.
- Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, 41(1), 3–26.
- Liu, Y., & Xin, T. (2017). *Dcmiffo: Information matrix for diagnostic classification models*. R package version 0.1.6. <https://CRAN.R-project.org/package=dcmiffo>
- Liu, Y., Xin, T., Andersson, B., & Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology*, 72(1), 18–37.
- Liu, Y., Xin, T., & Jiang, Y. (2021). Structural parameter standard error estimation method in diagnostic classification models: Estimation and application. *Multivariate Behavioral Research*, 57(5), 784–803.
- Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A

- comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology*, 10, 1137.
- Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, 69(3), 253–275.
- Ma, W., & de la Torre, J. (2019). Category – level model selection for the sequential G – DINA model. *Journal of Educational and Behavioral Statistics*, 44(1), 45–77.
- Ma, W., & de la Torre, J. (2020a). An empirical Q – matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology*, 73(1), 142–163.
- Ma, W., & de la Torre, J. (2020b). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, 93(14), 1–26.
- Ma, W., Ragip, T., & de la Torre, J. (2021). Detecting differential item functioning using multiple – group cognitive diagnosis models. *Applied Psychological Measurement*, 45(1), 37–53.
- Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the Estimation of Standard Errors in Cognitive Diagnosis Models. *Journal of Educational and Behavioral Statistics*, 43(1), 88–115.
- Robitzsch, A., & George, A. C. (2019). The R package CDM. In M. von Davier & Y. – S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 549–572). Springer. https://doi.org/10.1007/9783-030055844_26
- Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgment test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, 19, 506–532.
- Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 79(2), 317–339.
- Tian, W., Cai, L., Thissen, D., & Xin, T. (2013). Numerical differentiation methods for computing error covariance matrices in item response theory modeling: An evaluation and a new proposal. *Educational and Psychological Measurement*, 73(3), 412–439.
- von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional ‘diagnostic’ classification models—a commentary. *Psychometrika*, 79(2), 340–346.
- Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, 45(3), 373–391.

The Variance – covariance Matrix based Statistical Inferences in Cognitive Diagnostic Models

Wu Qiongqiong¹, Zhao Yue¹, Liu Yanlou²

(1. School of Psychology, Qufu Normal University, Jining 273165;

2. Academy of Big Data for Education, Qufu Normal University, Jining 273165)

Abstract: The variance – covariance matrix for the maximum likelihood estimates of model parameters in cognitive diagnostic models plays a key role in statistical inference. We (a) extensively analyze the applications of variance – covariance matrix in various fields of cognitive diagnosis, including the estimation of models parameter standard error, detection of differential item functioning, item – level comparison of saturated and reduced models, the Q – matrix estimation or validation and attribute hierarchy exploration, (b) provide possible explanations for previous study results and suggests how to advance the application of variance – covariance matrix, (c) introduce 13 information matrix estimation methods proposed in literature, which are classified according to whether structural parameters are considered, (d) briefly comment the advantages or disadvantages of these methods. We conclude by discussing future directions for variance – covariance matrix research.

Key words: cognitive diagnosis models; variance – covariance matrix; information matrix; standard error; the Q – matrix validation