DINA モデルの再定式化と EM アルゴリズムの直接的導出

山口一大（法政大学・学術振興会）

キーワード：認知診断モデル，DINA モデル，EM アルゴリズム

問題と目的

認知診断モデルは新しい教育測定モデルであり，テストから解答者の認知能力の強み弱みを推定することができる（山口・岡田, 2017）。なかでも，DINA モデル（deterministic input noisy and gate model）は最も基本的な認知診断モデルであり，研究が盛んに行われている。de la Torre（2009）は DINA モデルにおけるパラメタの最尤推定値を得るアルゴリズムを示した。しかしながら，de la Torre（2009）は対数周辺尤度を直接微分して，EM アルゴリズムを構成しており，EM アルゴリズムがどのように構成されるのか明らかではない。これは，MAP 推定値を得るアルゴリズムへの拡張が行いにくく問題がある。そこで本研究では，新たに潜在変数を明示的に導入することで，DINA モデルの再定式化，完全データの尤度を示し，より直接的に EM アルゴリズムを導出する。

DINA モデル

認知診断モデルでは，問題正答に必要な認知要素としてアトリビュートが設定される。アトリビュートは一般に 0 であれば未習得，1 であれば習得を表す潜在変数であり，複数のアトリビュートの習得・未習得の組合せを考える。このアトリビュート習得パタンを$\boldsymbol{\alpha}_l = [\alpha_{l1}, \cdots, \alpha_{lk}, \cdots, \alpha_{lK}]^{\mathrm{T}}$とする。$k(= 1, \cdots, K)$はアトリビュートの番号であり，$(l = 1, \cdots, L = 2^K)$はアトリビュート習得パタンを示す番号である。一方，Q 行列は項目とアトリビュートの間の関連を示す行列であり，$q_{jk} = 1$であれば，項目$j(= 1, \cdots, J)$にアトリビュートkが必要，そうでなければ 0 を示すように，分析者が作成する。これらを用いて，$\eta_{lj} = \prod_k \alpha_{lk}^{q_{jk}}$という，理想反応が定義される。理想反応はアトリビュート習得パタンlが項目jに必要なアトリビュートをすべて習得しているならば 1，そうでなければ 0 である

このη$_{lj}$を用い，DINA モデルの項目反応関数は

$$\mathrm{P}(x_{ij}, z_{il} | s_j, g_j, \pi_l) = \left[\pi_l \left((1-s_j)^{\eta_{lj}} g_j^{1-\eta_{lj}}\right)^{x_{ij}} \left(s_j^{\eta_{lj}}(1-g_j)^{1-\eta_{lj}}\right)^{1-x_{ij}}\right]^{z_{il}}$$

と定義される。ここで，x_{ij}は解答者$i(= 1, \cdots, I)$の項目jへの反応で，正答ならば 1，誤答ならば 0 である。z_{il}は解答者iがアトリビュート習得パタンlであれば 1 を示し，そうでなければ 0 を示す潜在変数であり，$\sum_l z_{il} = 1$を満足する。さらに，２つの項目パラメタは，$s_j = \mathrm{P}(x_{lj} = 0 | \eta_{lj} = 1)$と$g_j = \mathrm{P}(x_{lj} = 1 | \eta_{lj} = 0)$であり，$s_j$は項目$j$に正答に必要なアトリビュートを全て習得しているパタンlに属する人が誤答する確率で，g_jは項目jに正答に必要なアトリビュートが一つ以上足りないパタンlに属する人が誤答する確率と定義される。π_lはアトリビュート習得パタンの混合比率を表すパラメタで$\sum_l \pi_l = 1$である。

以上から，完全データの尤度は

$$\mathrm{P}(\mathbf{X}, \mathbf{Z} | \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\pi}) = \prod_i \prod_j \prod_l \mathrm{P}(x_{ij}, z_{il} | s_j, g_j, \pi_l)$$

となる。重要なのは，個人のアトリビュート習得パタンを指し示す変数z_{il}を明示的に組み込んだことである。これにより，EM アルゴリズムで期待値計算をすべき変数がわかりやすく示され，EM アルゴリズムを容易に導出可能となる。

EM アルゴリズム

EM アルゴリズムは，現在のパラメタのもとで，所属クラスを示す，z_{il}の期待値の計算し（E-step），対数完全尤度の$\mathbf{Z}$の事後分布での期待値をパラメタで最大化すればよい（M-step）。

E-step：，z_{il}の期待値の計算

$$\gamma(z_{il}) = \mathbb{E}_{\mathrm{P}(\boldsymbol{z_{il}} | \boldsymbol{x_i}, \Theta^{\mathrm{old}})}[z_{il}]$$

$$= \frac{\pi_l \prod_j \left((1-s_j)^{\eta_{lj}} g_j^{1-\eta_{lj}}\right)^{x_{ij}} \left(s_j^{\eta_{lj}}(1-g_j)^{1-\eta_{lj}}\right)^{1-x_{ij}}}{\sum_l \left(\pi_l \prod_j \left((1-s_j)^{\eta_{lj}} g_j^{1-\eta_{lj}}\right)^{x_{ij}} \left(s_j^{\eta_{lj}}(1-g_j)^{1-\eta_{lj}}\right)^{1-x_{ij}}\right)}$$

ただし，$\Theta^{\mathrm{old}} = [\boldsymbol{s}^{\mathrm{old}}, \boldsymbol{g}^{\mathrm{old}}, \boldsymbol{\pi}^{\mathrm{old}}]$で，常識右辺のパラメタは現在のパラメタ値とするが，表記の簡略化のために，“old”の添字を省略した。

M-step：パラメタを更新する。

$$\pi_l^{\mathrm{new}} = \frac{\sum_i \gamma(z_{il})}{I}$$

$$s_j^{\mathrm{new}} = \frac{\sum_i \sum_l \gamma(z_{il})\, \eta_{lj}(1-x_{ij})}{\sum_i \sum_l \gamma(z_{il})\, \eta_{lj}}$$

$$g_j^{\mathrm{new}} = \frac{\sum_i \sum_l \gamma(z_{il})\, (1-\eta_{lj}) x_{ij}}{\sum_i \sum_l \gamma(z_{il})\, (1-\eta_{lj})}$$

以上の 2 ステップを収束するまで反復する。このように，z_{il}を導入することで，EM アルゴリズムの導出が容易に行なうことができた。

引用文献

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

山口一大・岡田謙介 (2017). 近年の認知診断モデルの展開. *44*, 181-198.

付 記：本研究は特別研究員奨励費 18J01312 の助成を受けた。