Análise de Dados Categorizados - Aula 17 - Variáveis Latentes

Márcia D Elia Branco

Universidade de São Paulo
Instituto de Matemática e Estatística
mdbranco@usp.br - sala 295-A -

Modelos de Variáveis Latentes

- Algumas quantidades de interesse não conseguem ser medidas diretamente, seja pela característica intrínseca dessa quantidade ou pela dificuldade prática de fazer essa medição.
- Exemplos nas ciências sociais: qualidade de vida, posição política, inteligência, habilidade, etc...
- Exemplos nas ciências ambientais: biomassa marinha, etc...
- No entanto, essas quantidades não são consideradas parâmetros (no sentido clássico) e sim variáveis aleatórias.
- No nosso curso, vimos o exemplo de uma variável latente continua como sendo o grau de fragilidade do coração, tendo como variável resposta uma binária $y = 1$ (doença cardíaca). Modelo latente probito.

Tabela de classificação dos métodos

		V. Manifesta	
		Métrica	Categórica
V. Latente	Métrica	A. Fatorial	A. Traço Latente
	Categórica	A. Perfil Latente	A. Classe Latente

Referência: Bartholomew, D., Knott, M. and Moustaki, I. (2011). Latent Variable Models and Factor Analysis.

Teoria da Resposta ao Item (TRI)

- A TRI pode ser vista como um caso especial de Análise de Traço Latente.
- Foi introduzida nos anos 50 na área de psicometria, como alternativa a Teoria Classica dos Testes.
- Vamos considerar um Teste com J itens, respondido por I indivíduos.
- Considera $x_{ij} = 1$ se o i-ésimo indivíduo responde corretamente ao item j e $x_{ij} = 0$ caso contrário.
- Definimos $p_{ij} = P(x_{ij} = 1)$. Diversos modelos foram propostos para modelar essa probabilidade.

Onde entra a variável latente?

Teoria da Resposta ao Item (TRI)

- Considere θ_i a habilidade do i-ésimo indivíduo no conteúdo em teste.
- O teste foi contruído com o objetivo de aprender sobre esta quantidade latente θ_i.
- Considere $x_{ij} \mid \theta_i \sim Be(p_{ij})$ independentes com

$$p_{ij} = F(\theta_i - \beta_j)$$

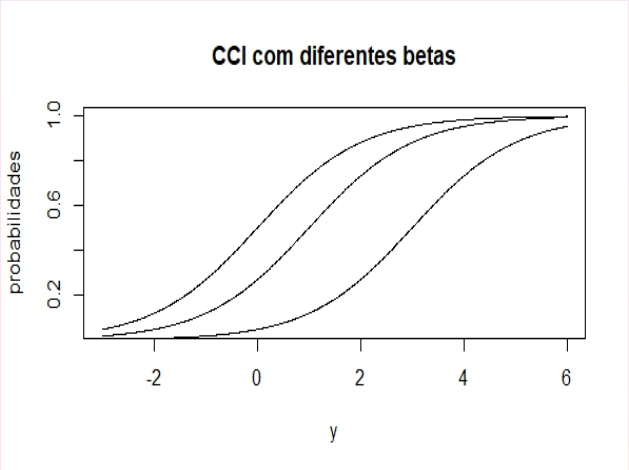
em que F é uma função distribuição acumulada de probabilidades e β_j são parâmetros associados aos itens.

- Usualmente considera-se $\theta_i \sim N(0, 1)$.

- Se F é a fda de uma logística, obtemos

$$p_{ij} = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}$$

- Conhecido como modelo de Rasch ou 1PL.
- Os parâmetros β_j são interpretados como dificuldade do item.



Teoria da Resposta ao Item (TRI)

- O modelo de Rash usualmente é estendido para um modelo de dois parâmetros.
- Considere $x_{ij} \mid \theta_i \sim Be(p_{ij})$ independentes com

$$p_{ij} = F(a_j(\theta_i - b_j))$$

em que F é uma função distribuição acumulada de probabilidades e a_j, b_j são parâmetros associados aos itens.

- Usualmente considera-se $\theta_i \sim N(0, 1)$.

- Se F é a fda de uma logística, obtemos

$$p_{ij} = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}$$

  Conhecido como modelo 2PL.

- Os parâmetros b_j são interpretados como dificuldade do item e os a_j como discriminação.

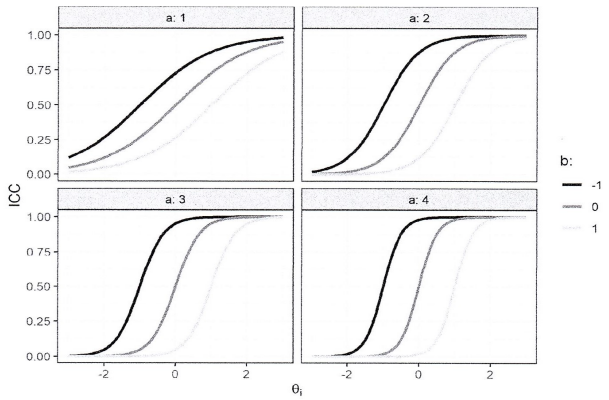


Figura 2.1: *Alguns exemplos de ICC do modelo 2PL para diferentes valores de a_j e b_j.*

Teoria da Resposta ao Item (TRI)

- Em TRI temos dois objetivos principais: (1) estimar os parâmetros dos itens; (2) estimar as habilidades.
- Em geral, essas estimativas são realizadas em duas etapas. A estimação dos parâmetros dos itens é denominada calibração.
- Muitas alternativas ao modelo $2PL$ foram propostas em TRI. Em particular, flexibilizações da função característica do item (CCI). Para tal, considera-se diferentes possibilidades para F.
- Por exemplo, para ligações assimétricas ver Bazán, Branco e Bolfarine (2006) em *Bayesian Analysis*. Introduzindo um novo parâmetro de item, interpretado como penalização.

- Um problema associado ao modelo multidimensional da TRI, mas um pouco diferente, é o modelo multi-unidimensional.
- Vamos considerar que um indíviduo responda T diferentes testes e para cada teste uma habilidade θ_t é requerida.
- Assim, $x_{ijt} = 1$ se o i-ésimo indivíduo responde corretamente o j-ésimo item do t-ésimo teste e $x_{ijt} = 0$ caso contrário.
- O modelo de 2 parâmetros neste caso é

$$p_{ijt} = F(a_{jt}(\theta_{it} - b_{jt}))$$

- $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iT})$ é o vetor de habilidades do indivíduo i .

Teoria da Resposta ao Item (TRI)

- Assumir independência entre as habilidades dos mesmo indivíduo não é rasoável.
- Assim, considera-se

$$\theta_i \sim N_T(0, \Sigma)$$

- O desafio esta em modelar a estrutura de correlação, representada por Σ.
- O aluno de mestrado Pedro Araújo trabalhou neste problema com uma solução bayesiana usando ideias de estatística espacial.

TRI: multi-unidimensional

- Motivado por modelos de estatística espacial, uma nova estrutura latente foi consideranda.
- Para cada teste t considera-se um par ordenado (z_{1t}, z_{2t}) que irá indicar uma posição espacial para esse teste.
- A seguir uma figura ilustrativa da distância de 4 testes.

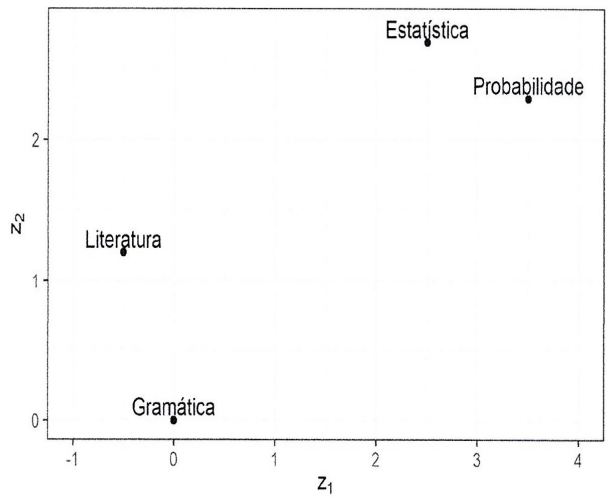


Figura 3.1: *Exemplo hipotético da relação latente entre 4 testes.*

TRI: multi-unidimensional

A matriz de correlações Σ considera a seguinte estrutura

$$Corr(y_{it}, y_{is}) = \exp\left\{-\frac{1}{\phi}||z_t - z_s||\right\}$$

$\phi > 0$ é um hiperparâmetro que será estimado.

- Valores muito baixos de ϕ estão associados a baixa correlação entre os teste.
- Testes cujas distância $||z_t - z_s||$ são altas, tem baixa correlação.
- Para as novas variáveis latentes $z_t = (z_{1t}, z_{2t})$ é considerada uma distribuição $N_2(0, I_2)$.

TRI: multi-unidimensional

- O modelo 2PL foi usado para definir as Curvas Características dos Itens.
- Foi realizado um longo estudo de sensibilidade para escolha da distribuição *a priori* para ϕ.
- A estimação dos parâmetros foi realizado usando-se o STAN.
- Vamos ilustrar o modelo com uma aplicação feita as provas do ENEM de 2018. Amostra de $n = 3000$ estudantes.
- São 4 habilidades testadas por 4 provas: Linguagem e Suas Tecnologias (LC), Matemática e Suas Tecnologias (MT), Ciências da Natureza e Suas Tecnologias (CN) e Ciências Humanas e Suas Tecnologias (CH). 45 itens por prova.
- Os resultados são apresentados em uma artigo submetido (ver artigo!).

TRI: multi-unidimensional

- De um modo geral as habilidades medidas pelas provas tem valores altos de correlação, em torno de 80 %.
- A menor distância obida foi entre CH-LC, próximo de 1 (mediana a posteriori).
- As maiores distâncias ocorrem MT e as outras provas. Com destaque para uma distância por volta de 3 (mediana a posteriori) entre MT e LC.
- As estimativas para o parâmetro ϕ foram $Med = 7.9$ com intervalo de credibilidade $IC(0.95) = (4.19, 13.04)$.

Modelo estatístico de ponto ideal

- O segundo exemplo, considera um modelo latente proposto por Barberá (2015).
- O objetivo é classificar os indivíduos usuários de redes sociais segundo a sua posição política.
- Para tal, é construída uma variável latente a qual denomina *ponto ideal* (θ_i) associada a cada usuário.
- θ_i é um valor real. Valores positivos colocam os usuários em determinado espectro político (direita) e valores negativos no lado oposto (esquerda).
- Os dados utilizados para predizer essas quantidades são as conexões feitas pelos usuários na rede social Twitter.
- $x_{ij} = 1$, se o usuário i segue o influenciador j e zero, caso contrário.

Modelo estatístico de ponto ideal

Modelo considerado:

$$logit(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right)$$

Em que

$$logit(p_{ij}) = \alpha_j + \beta_i - \gamma||\theta_i - \phi_j||^2$$

α_j y ϕ_j são as variáveis latentes associadas aos influenciadores

β_i y θ_i são as variáveis latentes associadas aos cidadãos

$\gamma > 0$ é uma constante de normalização.

Ver slides da apresentação do trabalho da Camila.

Referências

[1] Araújo, Pedro M. (2022). Visualizando a relação entre testes em modelos de TRI multi-unidimensionais. Dissertação de mestrado, IME-USP.

[2] Bartholomew, D., Knott, M. and Moustaki, I. (2011). Latent Variable Models and Factor Analysis: A Unified Approach. Wiley Series.

[3] Bazán, Branco and Bolfarine (2006). A skew item response model. *Bayesian Analysis.*

[4] Morais, Camila L. (2021). Diga-me quem segues e lhe direi quem és: Estimação de ideologia feminista no Twitter usando ponto ideal bayesiano. TCC - IME - USP .