INTRODUÇÃO A LATENT CLASS ANALYSIS (LCA):

ABORDAGEM DE MAXIMA VEROSSIMILHANCA

PROF. JORGE LUIS. BAZAN

https://jorgeluisbazan.weebly.com

TÓPICOS

1. Análise de Classe Latente
2. Estimação
3. Software
4. Exemplo
5. Referencias

1. Análise de Classe Latente (ACL)

- A ACL proporciona un enfoque flexible y poderoso para el análisis de datos categóricos (McCutcheon y Hagenaars, 1997).

- E um modelo de variáveis latentes que tenta descobrir ``classes'' (agrupamentos) latentes (uma variável discreta) que é subyacente a respostas observadas discretas.

- O modelo determina as probabilidades de pertencia a um dos agrupamentos em função das respostas a um conjunto de variáveis categóricas.

- E um tipo de análise de conglomerados (cluster ou agrupamentos).

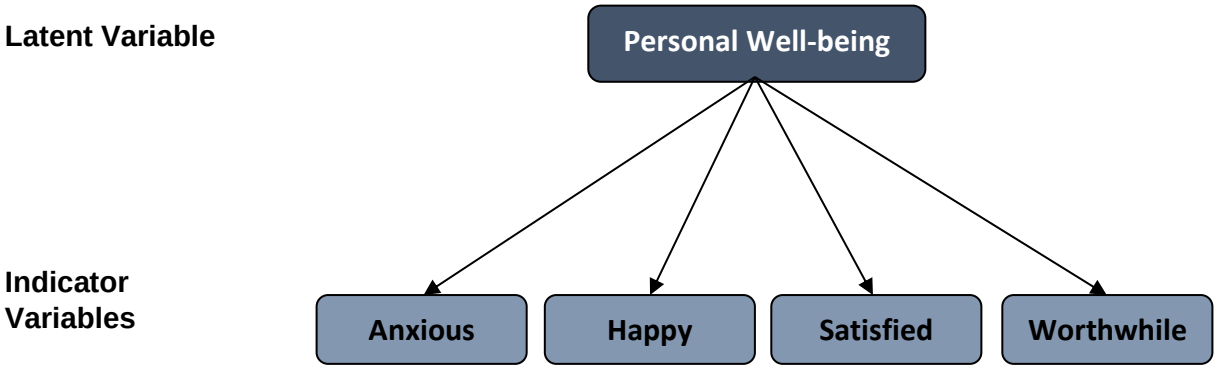
- Utiliza o algoritmo EM mas pode ser ajustado sobre abordagem bayesiana também (para más detalhes metodológicos, ver McCutcheon, 1997 y Linzer & Lewis, 2011).

- Em numerosos estudos, especialmente em investigação social, os investigadores estão interessados em variáveis latentes (variáveis que não podem ser medidas diretamente), por exemplo bem-estar pessoal ou qualidade de vida. Essas variáveis tendem a ser medidas por meio de uma série de variáveis indicadoras (observadas)

Exemplo

- Por exemplo, o Gabinete de Estatísticas Nacionais (Office for National Statistics ONS) utiliza quatro variáveis indicadoras para medir o bem-estar pessoal (uma variável latente) no Reino Unido[1]:

Figura 1. Variáveis indicadoras usadas para medir o bem-estar pessoal no Reino Unido



[1] Data for personal well-being official statistics are collected by the ONS as part of the Annual Population Survey (APS). See: Office for National Statistics (2016) for data.

https://analysisfunction.civilservice.gov.uk/wp-content/uploads/2017/02/A-Short-Guide-to-using-Latent-Class-Analysis-Final2.docx

Table 1: Labelling of threshold

Life satisfaction, worthwhile and happiness scores		Anxiety scores	
Response on an 11 point scale	Label	Response on an 11 point scale	Label
0 – 4	Low	0 – 1	Very low
5 – 6	Medium	2 – 3	Low
7 – 8	High	4 – 5	Medium
9 – 10	Very high	6 – 10	High

Source: Office for National Statistics
Personal well-being in the UK - Office for National Statistics (ons.gov.uk)

- Há quatro variáveis indicadoras (categóricas) observadas de quatro alternativas (não necessariamente na mesma escala) e estamos interessados em criar dois ou mais grupos para classificar os indivíduos.
- Subjacente a estas variáveis, queremos medir o ``Bem-estar`` em categorias e queremos determinar a probabilidade de cada indivíduo serem classificado nestas categorias.

- O ACL é utilizado para identificar padrões de respostas às variáveis indicadoras (variáveis categóricas) para criar um conjunto de classes latentes mutuamente exclusivas, ou seja, grupos de indivíduos ou outras unidades de análise.

- O ACL é utilizado para identificar padrões de respostas às variáveis indicadoras (variáveis categóricas) para criar um conjunto de classes latentes mutuamente exclusivas, ou seja, grupos de indivíduos ou outras unidades de análise.

- Os indivíduos na mesma classe latente terão padrões de resposta semelhantes às variáveis indicadoras, enquanto os indivíduos nas classes latentes tendem a ter padrões de resposta diferentes entre si.

- Em outras palavras, a LCA divide os respondentes em grupos homogêneos (classes latentes).

- De acordo com Pratt (2020), o ACL tem inúmeras vantagens sobre as técnicas tradicionais de análise de cluster, como análise hierárquica de cluster e agrupamento K-means. Algumas dessas vantagens incluem:

  a. É baseada em modelos, ao contrário de outros tipos de análise de cluster que tendem a ser baseadas na distância. Uma vantagem disso é que existem critérios mais formais para a escolha do modelo final quando se utiliza ACL (para mais informações ver Vermunt & Magidson, 2002).

  b. É relativamente fácil lidar com variáveis que possuem diferentes tipos de escala ou categorías (Vermunt & Magidson, 2002).

c. Nas técnicas tradicionais de análise de agrupamentos, as pessoas são designadas para agrupamentos numa base de tudo ou nada. Por outro lado, a ACL permite a adesão de uma pessoa a cada cluster até um certo grau, permitindo a adesão fracionária ao cluster (capturada por possibilidades posteriores usando probabilidades).

2. ESTIMAÇÂO

- A ACL trabalha partindo do pressuposto de que a distribuição multivariada observada é formada por uma mistura de distribuições resultantes de classes “não observadas” ou latentes.

- Usando conjuntos de indicadores observados, os modelos de ACL tentam melhor identificar essas classes. Os modelos de ACL utilizam uma função de máxima verossimilhança para estimar os parâmetros do modelo.

- A estimativa de máxima verossimilhança refere-se aos valores dos parâmetros que são mais consistentes dados os dados observados.

- Como os parâmetros representam os componentes mais fundamentais do modelo e determinam como ele se comporta, é crucial maximizar os dados para encontrar soluções/estimativas para os parâmetros que melhor representam os dados reais a partir dos quais o modelo é ajustado.

- Por exemplo, no caso de modelos compostos por variáveis contínuas, os parâmetros são as médias e os desvios padrão das distribuições que foram “misturadas” juntamente com as proporções da amostra que formam as distribuições.

- Nesse modelo, pode haver uma infinidade de valores de parâmetros (distribuições gaussianas) que podem resultar nos dados observados; no entanto, a estimativa de máxima verossimilhança identifica o valor do parâmetro que leva aos dados observados com a maior probabilidade.

- Como a participação na classe é latente, algoritmos de maximização de expectativas (EM) são usados para ajustar modelos de ACL.

- Os algoritmos EM fornecem uma estrutura para gerar estimativas de probabilidade na presença de dados faltantes; uma explicação mais detalhada desses algoritmos pode ser encontrada em uma revisão de Dempster e colegas. (1)

- Os parâmetros de mistura são otimizados iterativamente até que uma solução global para a estimativa de máxima verossimilhança seja identificada. As inferências das estimativas de máxima verossimilhança dos parâmetros são fundamentais para a determinação dos clusters e sua separação no modelo.

- Consequentemente, uma probabilidade de pertença de um indivíduo a todas as classes de um modelo pode ser gerada com base nos seus indicadores observados.

- Para uma classe g com proporcoes dadas por π , o modelo pode ser expressado como

$$f(y) = \sum_{i=1}^{g} \pi_i \, fi(y|x'B_i) \qquad [1]$$

Onde π_i é a probabilidade para a *i-ésima* classe e

$f_i(\cdot)$ é função de densidade de probabilidade condicional da resposta na i-ésima classe do modelo.

- Para estimar as probabilidades das classes latentes, é assumido um modelo multinomial com as seguintes probabilidades:

$$\pi_i = \frac{\exp(\gamma_i)}{\sum_{j=1}^{g} \exp(\gamma_i)} \qquad [2]$$

Em que γ_i é o preditor linear para a i-ésima classe latente.

Interpretando o Modelo Final

- Valores iniciais. Para encontrar a solução de máxima verossimilhança, o algoritmo EM deve iterar para frente e para trás até que a verossimilhança atinja seu valor máximo.

- Na ACL, os algoritmos EM utilizados pelos modelos são algoritmos “gananciosos”, portanto são sensíveis à inicialização que pode encontrar um “máximo local”, que pode não ser a verdadeira máxima verossimilhança por si só. Para evitar isso, vários valores iniciais aleatórios devem ser usados.

- Se cada um levar ao mesmo valor de máxima verossimilhança, pode-se ter certeza de que o verdadeiro máximo foi encontrado.

- Para demonstrar a consistência do desempenho do modelo, os modelos com melhor ajuste (log-verossimilhança máxima) devem ser replicados aumentando o número de replicas para garantir que o valor máximo esteja provavelmente correto.

- Para um modelo, a falta de convergência nas replicas para obter o máximo da log-verossimilhança sugeriria que os dados não suportam o número de classes contidas no modelo.(2)

1. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological). 1977; 39(1):1-38.

2. Berlin KS, Williams NA, Parra GR. An introduction to latent variable mixture modeling (part 1): overview and cross-sectional latent class and latent profile analyses. J Pediatr Psychol. 2014;39(2):174-87.

3. SOFTWARE

- A ACL sobre abordagem de máxima verossimilhança. pode ser executado em muitos programas de software, como

  - SAS®3, R (R Foundation for Statistical Computing, 2011),

  - STATA3 (StataCorp LP, 2015),

  - Mplus (Muthén & Muthén, 2011)

  - Latent Gold (Vermunt & Magidson, 2013), entre outros.

- ACL pode ser feito usando o pacote no R poLCA (Linzer & Lewis, 2013; Linzer & Lewis, 2011).

- A expressão para definir uma ACL no poLCA, segue a seguinte especificação:

```
f <- cbind(Y1, Y2, Y3) ~ 1
```

Neste caso Y1, Y2 e Y3 são variaveis categoricas que serão inclusas na analise. O simbolo "~ 1" diz ao pacote estimar o modelo de classe latente básico. Se consideramos "~ Y4" iso indica um modelo de regressao de classe latente, neste caso as variáveis categóricas serão usadas para explicar Y4.

- Para executar o poLCA, o comando básico utilizado é o seguinte:

```
poLCA (formula, data, nclass = 2, maxiter = 50000,
graphs = FALSE, na.rm = TRUE, nrep =  10, verbose =
TRUE)
```

- A seguir são descritos os componentes dentro da função

- formula: a definição da formula ‘f’ especificada acima

- data: o nome do data frame que sera usado na ACL

- nclass: o número de classes latentes a serem calculadas no modelo. O padrão são 2 classes latentes. poLCA assume um conjunto de classes latentes cada vez que é executado. Portanto, para obter múltiplos modelos, cada um assumindo

um número diferente de classes latentes, o comando deve ser executado um número de vezes de cada vez especificando um número diferente de classes latentes a serem assumidas.

- maxiter: este é o número máximo de iterações para convergência. Se a convergência não for alcançada antes de atingir este número de iterações, uma mensagem de erro aparecerá e a análise será encerrada.

- graphs: isso especifica se um gráfico mostrando as estimativas dos parâmetros deve ser produzido. O padrão é FALSE. Demora muito para R executar análises com 4 ou mais classes latentes em grandes conjuntos de dados. Portanto, pode ser mais rápido executar a análise sem produzir gráficos e, se necessário, apenas produzir um gráfico para o modelo mais adequado depois de os modelos resultantes terem sido comparados.

- na.rm: isso específica como o poLCA trata casos com valores ausentes. Se especificado como TRUE, esses casos serão removidos por meio de exclusão listwise antes da estimativa do modelo. Se especificado como FALSE, os casos com valores ausentes serão retidos. O padrão (default) é TRUE. Linzer e Lewis (2011) sugerem que não é necessário excluir casos com valores faltantes antes de estimar o modelo porque o poLCA exclui do cálculo os casos com valores faltantes.

- nrep: esta opção é usada para especificar o número de vezes que o modelo deve ser estimado usando diferentes valores iniciais. É preferível definir nrep como maior que 1 para garantir que o algoritmo encontre um máximo global em vez de local da função de log-verossimilhança

- verbose: isso indica se os resultados do modelo devem ser exibidos na tela ou não. O padrão é TRUE.

- Existem várias opções adicionais que podem ser incluídas neste comando. Para uma lista completa dessas opções, consulte Linzer e Lewis (2011).

- A ACL descrita acima é uma ACL exploratória. Esta é provavelmente a forma mais simples de ACL, existem muitas extensões para isso que criam uma infinidade de usos para a ACL.

4. UMA APLICAÇÂO

- Seguiremos o exemplo em Pratt (2020):

Pratt, B. (2020). Latent Class Analysis.
https://pop.princeton.edu/sites/g/files/toruqf496/files/documents/2020Jan_LatentClassAnalysis.pdf

5. REFERENCIAS

- Biemer, P. P., 2001. Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. *Journal of Official Statistics*, 17(2), p.295-320.

- Biemer, P. P. and Wiesen, C., 2002. Measurement Error evaluation of self-reported drug use: a latent class analysis of the US Household National Survey on Drug Abuse. *Journal of the Royal Statistical Society A,* 165(1), p.97-119.

- Biemer, P. P., 2011. Latent Class Analysis of Survey Error. New Jersey: Wiley.

- Forster M.R., 2000. Key Concepts in Model Selection: Performance and Generalizability.*Journal of Mathematical Psychology*, 44, 205-231.

- Chanfreau, J., Cullinane,C., Calcutt, E. and McManus, S. , 2014. Wellbeing in Wales Secondary analysis of the National Survey for Wales 2012-13. [pdf]. Welsh Government Social Research. Available at: http://gov.wales/docs/caecd/research/2014/140430-national-survey-wellbeing-wales-2012-13-en.pdf [Accessed 25 January 2017].

- Kreuter, F., Yan, T. and Tourangeau, R., 2008. Good item or bad—can latent class analysis tell?: the utility of latent class analysis for the evaluation of survey questions. *Journal of the Royal Statistical Society A*. 171(3).

- Lanza, S.T., Collins, L.M., Lemmon, D.R. and Schafer, J.L. 2007. PROC LCA: A SAS Procedure for Latent Class Analysis. *Structural Equation Modeling*, 14(4), p.671–694.

- Lanza, S.T. and Rhoades, B.L., 2013. Latent Class Analysis: An Alternative Perspective on Subgroup Analysis in Prevention and Treatment. *Prevention Science*, 14(2), p.157-168.

- Lin T.H. and Dayton C.M., 1997. Model Selection Information Criteria for Non-Nested Latent
- Class Models. *Journal of Educational and Behavioral Statistics*, *22*(3), 249-264.

- Linzer, D. A. and Lewis J. B., 2013. "poLCA: Polytomous Variable Latent Class Analysis." R package version 1.4. http://dlinzer.github.com/poLCA.

- Linzer, D. A. and Lewis J. B., 2011. "poLCA: an R Package for Polytomous Variable Latent Class Analysis." *Journal of Statistical Software*. 42(10): 1-29. http://www.jstatsoft.org/v42/i10

- McCutcheon A.L. and Hagenaars, J.A.,1997. Simultaneous Latent Class Models for Comparative Social Research. In: Langeheine, R. and Rost, J. (eds) Applications of Latent Trait and Latent Class Models. New York: Waxmann. Pgs. 266-277.

- Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

- Office for National Statistics., 2016. Personal well-being in the UK: Oct 2015 to Sept 2016 [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/oct2015tosept2016 [Accessed 25 January 2017].

- Office for National Statistics., 2016. Annual Population Survey [online]. Available through: UK Data Service Discover

https://discover.ukdataservice.ac.uk/series/?sn=200002 [Accessed 25 January 2017].

- Office for National Statistics., 2017. Surveys using the 4 Office for National Statistics personal well-being questions [online]. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/surveysusingthe4officefornationalstatisticspersonalwellbeingquestions [Accessed 25 January 2017].

- R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. 3.0.2. [online]. Available at: http://www.R-project.org [Accessed 19 January 2016].

- SAS Institute Inc. 2010. SAS/STAT™ 9.22 User's Guide. Cary, NC: SAS Institute Inc. [online] Available at: https://support.sas.com/documentation/cdl/en/statug/63347/PDF/default/statug.pdf [Accessed 19 January 2016].

- StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.

- The Methodology Centre. 2015. SAS Procedures for Latent Class Analysis & Latent Transition Analysis. [online]. Available at: https://methodology.psu.edu/downloads/proclcalta [Accessed 18 January 2016].

- Vermunt, J. K., & Magidson, J., 2002. Latent class cluster analysis. In: J. A. Hagenaars, & A. L. McCutcheon eds. *Applied latent class analysis.* New York: Cambridge University Press, pp.89-106.

- Vermunt, J. K. and Magidson, J., 2013. Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Belmont Massachussetts: Statistical Innovations Inc. [online]. Available at: http://www.statisticalinnovations.com/latent-gold-5-1/ [Accessed 19 January 2016].