Latent class analysis

General introduction

DL Oberski& L Boeschoten

Website for the course

https://lauraboeschoten.github.io/LCA_GESIS/

Introductions

- What is your name?
- What is your discipline/background?
- Do you work in industry, academia, or nonprofit?
- On a scale of

  0 (random alpaca) – 10 (Hadley Wickham)

  How good are you with R?
- What do you hope to learn in this course?
- What is your favorite TV/Netflix/etc. show?

Groups

- You now have data about everybody
- I would like you to **create groups** of people
- You may do this ***however you like (!!)***
- You do not need to think too much about it, but you need to be able to explain your process

Learning goals

- Explain what makes an analysis an LCA
- Explain the different purposes of latent class analysis

What is latent class analysis?

- A specific statistical model
- A latent variable model
- A way to find groups in data
- An extreme case of missing data
- …?

“A statistical model can be called a latent class (LC) or mixture model if it assumes that some of its parameters differ across unobserved subgroups, latent classes, or mixture components.”

- Vermunt (2022)

A specific statistical model

DEFINITIONS

- "Indicators": Observed variables $Y_1, \ldots, Y_p$, collected in vector $\boldsymbol{y}$.
  - Indicators may be categorical or continuous (but usually categorical)
- "Latent classes": Unobserved *categorical* variable $X \in \{1, \ldots, K\}$

- Remarks:
  - The number of indicators is p
  - The number of classes is K

A specific statistical model

MODEL ASSUMPTIONS

1. **Mixture assumption**:

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} p(\boldsymbol{y} \mid X = k)\, p(X = k)$$

(*Remark*: using “Gelman notation” for the probability distributions *p*)

”The data arise from different groups, but we miss the groups”

“The chance to observe any *data (Y)* pattern is a result of collapsing over unobserved categories of latent variable (X)”

A specific statistical model

MODEL ASSUMPTIONS

2. Conditional independence assumption:

$$p(\boldsymbol{y} \mid X = k) = \prod_{j=1}^{p} p\left(y_j \mid X = k\right)$$

”All dependence in the data is caused by the latent variable”

“The observed indicators Y are conditionally independent, given the latent class X”

A specific statistical model

Complete model:

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} \prod_{j=1}^{p} p(y_j \mid X = k)\, p(X = k)$$

- Incorporates the **mixture** and **conditional independence** assumptions
- Many extensions are possible, such as including predictors of the latent classes (X|Z) or relaxing conditional independence

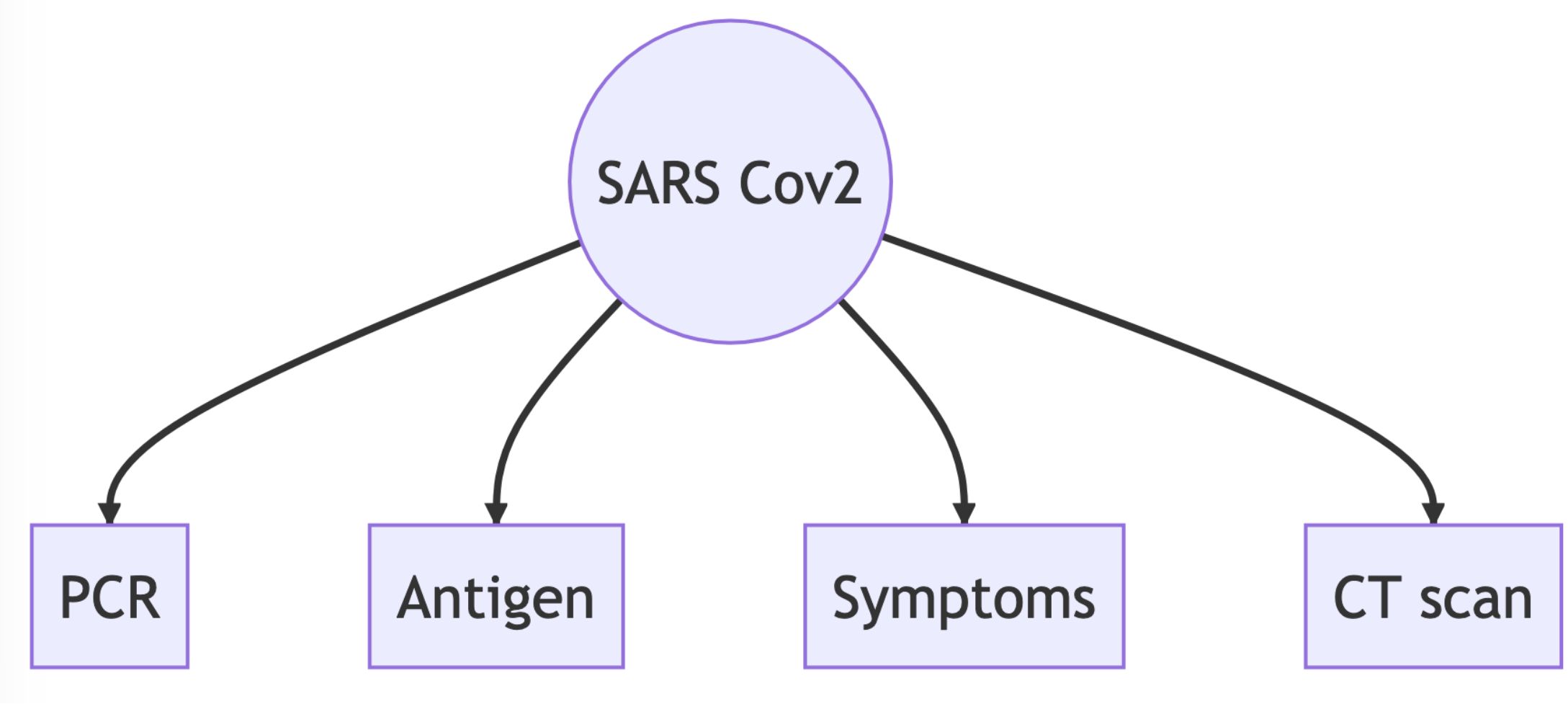
Some observations so far

- I *have* told you a statistical model
- I have *not* told you:
  1. What it means
  2. When it makes sense
  3. How it relates to other approaches with similar goals
  4. How to estimate the parameters of the model
  5. When this is even possible
  6. How you can do xyz in the context of an LCA

  Etc.

This is what the rest of the course is about.

A latent variable model



Note on graphical models

- An **ellipse** means a variable that is **unobserved** (i.e. completely missing);
- A **rectangle** means a variable that is **observed** (though it may have a few missings here and there);
- When there is ***no* arrow** between two variables, they are **conditionally independent** given everything else;
- The **direction** of the arrow indicates that we will be interested in the distribution of the “destination” given the “origins”.

American Journal of Epidemiology

Vol. 190, No. 8
https://doi.org/10.1093/aje/kwab093
Advance Access publication:
March 31, 2021

Practice of Epidemiology

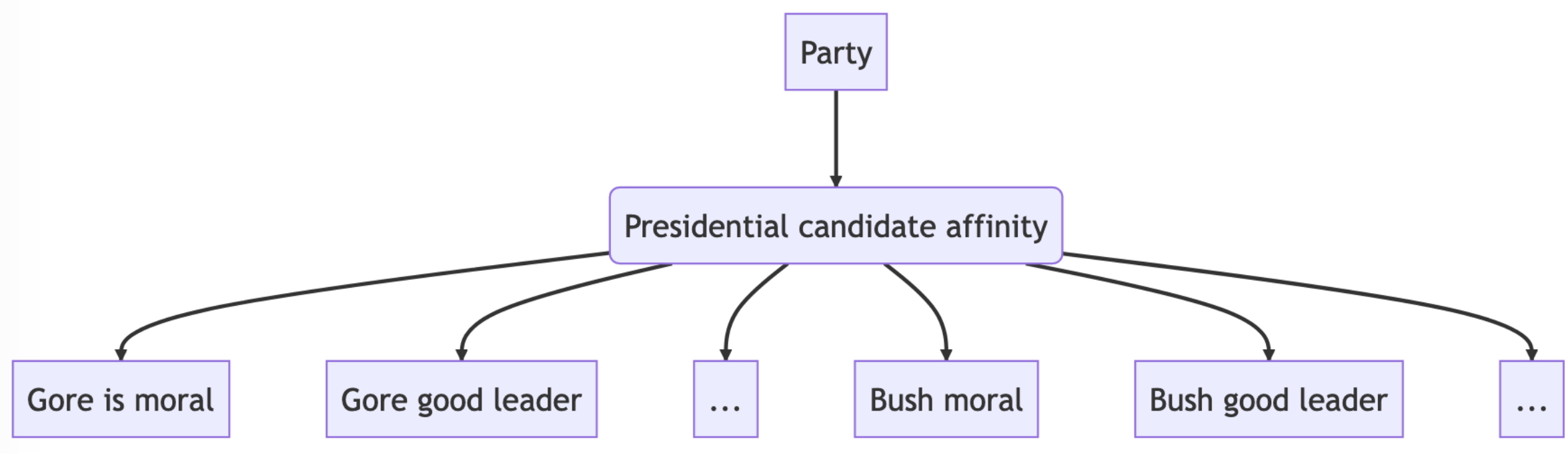
Diagnostic Accuracy Estimates for COVID-19 Real-Time Polymerase Chain Reaction and Lateral Flow Immunoassay Tests With Bayesian Latent-Class Models

Polychronis Kostoulas*, Paolo Eusebi, and Sonja Hartnack

Table 2. Medians and 95% Probability Intervals for the Sensitivity and Specificity of the Real-Time Reverse-Transcriptase Polymerase Chain Reaction and the Lateral Flow Immunoassay Tests Detecting Immunoglobulin G or Immunoglobulin M Antibodies Against Coronavirus Disease 2019, Using Bayesian Latent-Class Models

Model	Median	PrI	Week 1		Week 2		Week 3	
			Median	PrI	Median	PrI	Median	PrI
A[a]								
$Se_{RT\text{-}PCR}$[b]	0.68	0.63, 0.73						
$Sp_{RT\text{-}PCR}$[b]	0.99	0.98, 1.00						
$Se_{IgG/M}$			0.32	0.23, 0.41	0.75	0.67, 0.83	0.93	0.88, 0.97
$Sp_{IgG/M}$			0.97	0.92, 1.00	0.98	0.95, 1.00	0.98	0.94, 1.00

A latent variable model

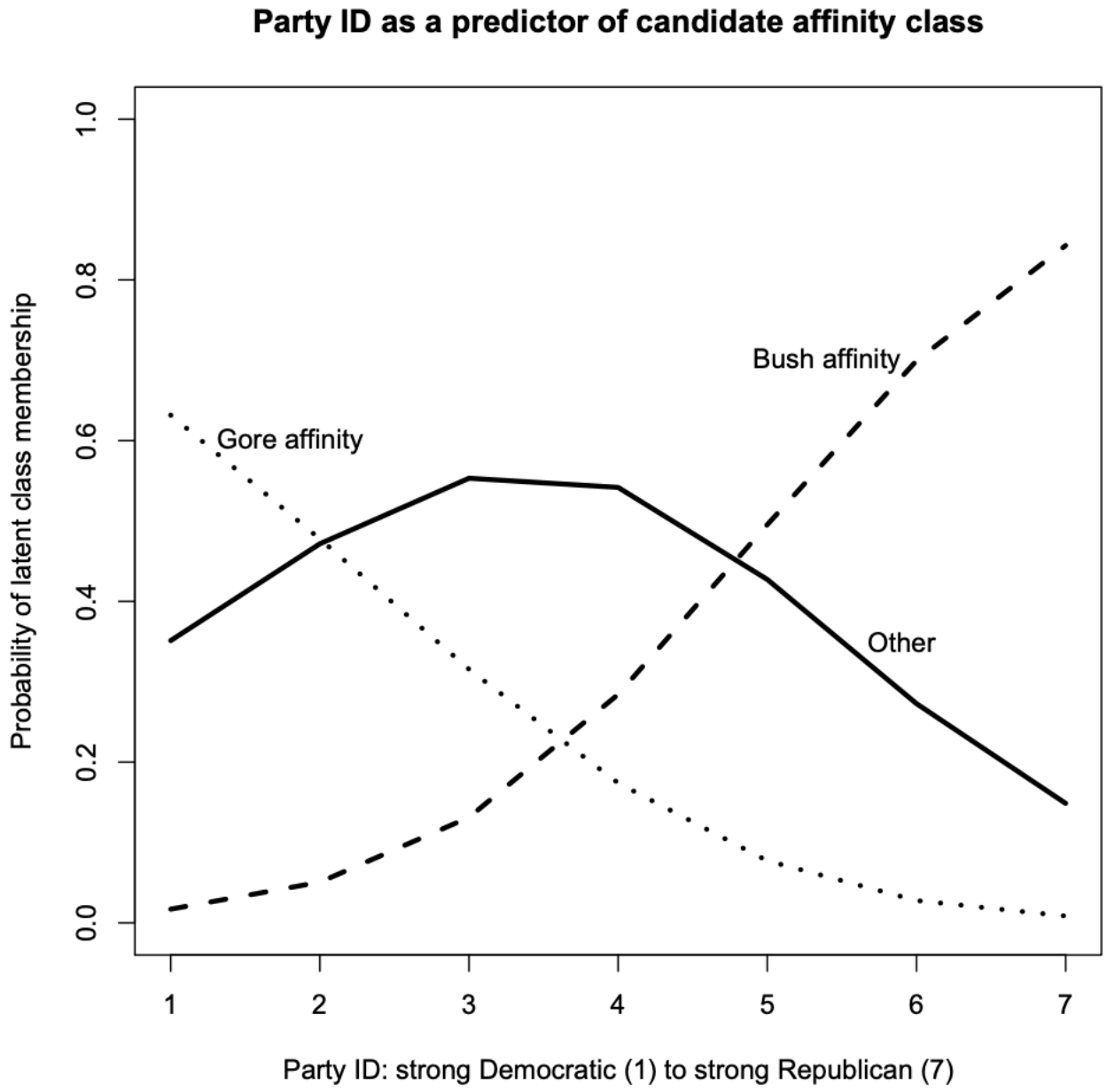


US National election study 2000

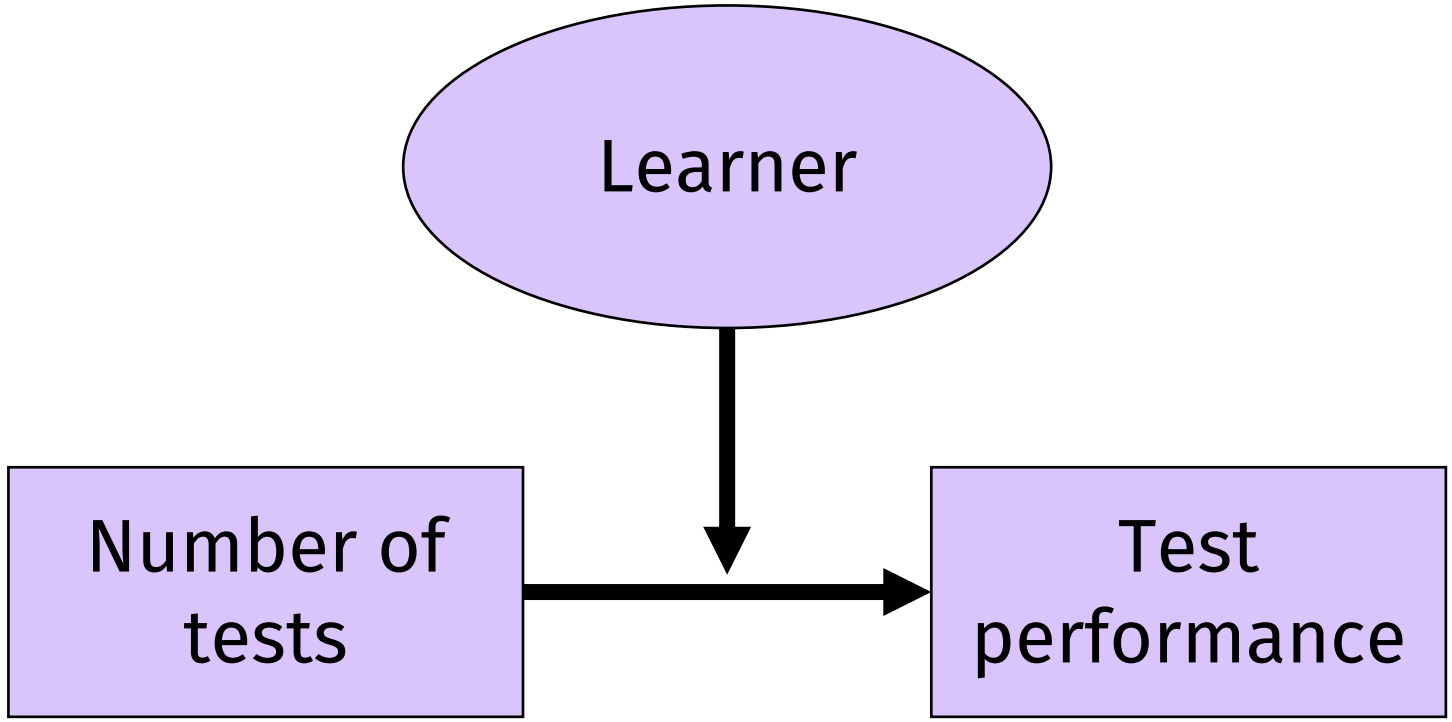
- “Respondents to the 2000 American National Election Study public opinion poll were asked to evaluate how well a series of traits:
  - moral,
  - caring,
  - knowledgable,
  - good leader,
  - dishonest, and
  - intelligent
- described presidential candidates Al Gore and George W. Bush.
- Each question had four possible choices: (1) extremely well; (2) quite well; (3) not too well; and (4) not well at all.”

Results

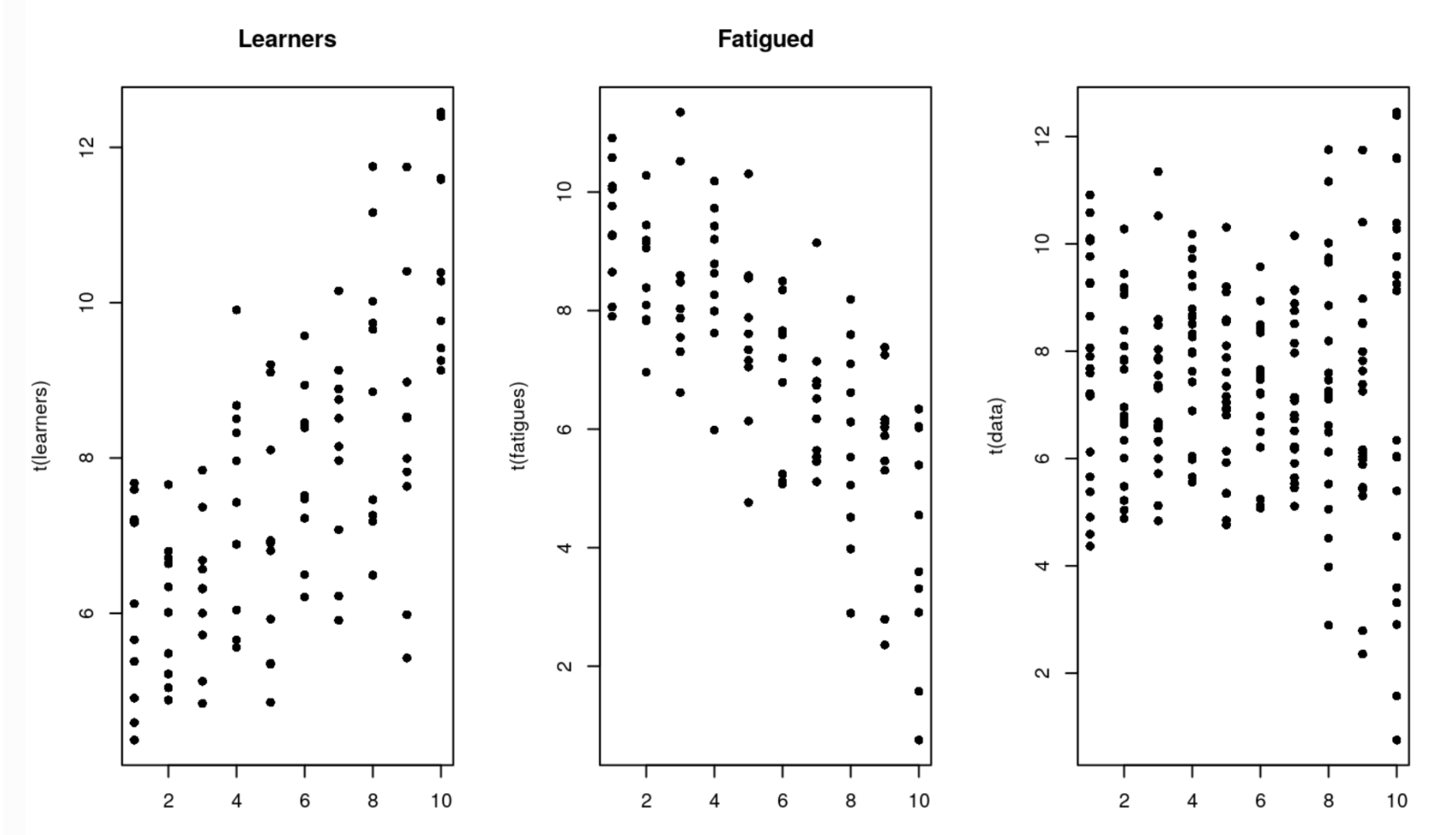
(Linzer & Lewis)



A latent variable model



Number of tests taken vs. Test performance



A way to find groups in data

Latent class analysis in the national* news!

INDEPENDENT

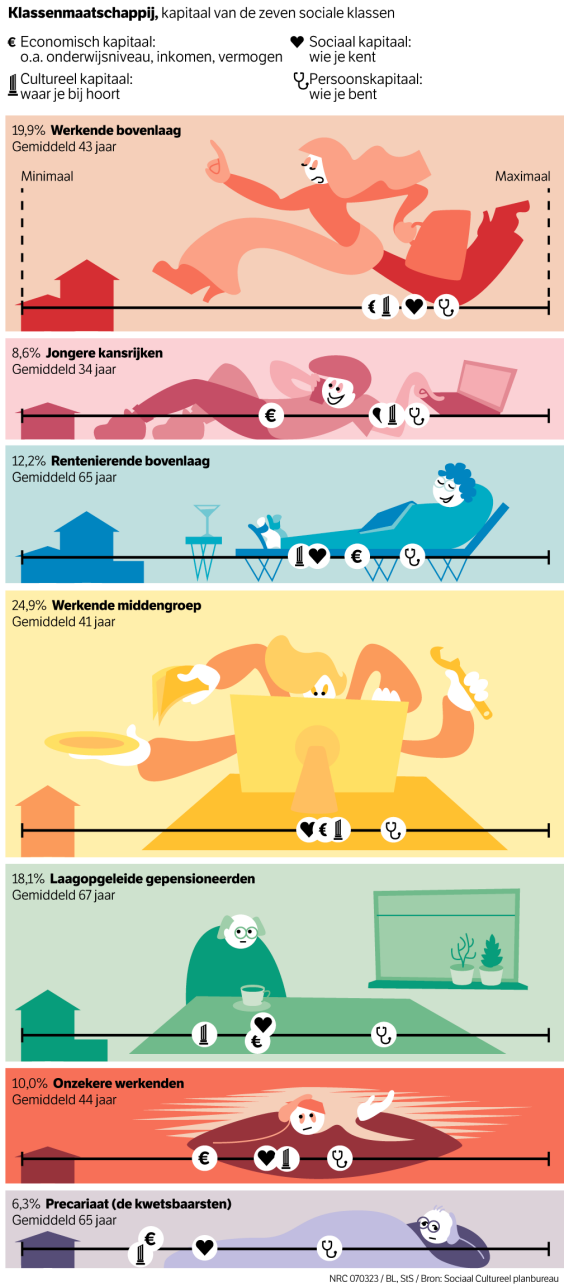
NEWS SPORT VOICES CULTURE LIFESTYLE TRAVEL PREMIUM

News > UK > Home News

Britain now has 7 social classes - and working class is a dwindling breed

Savage, Mike (2015). *Social Class in the 21st Century*. London: Penguin.


PRECARIAT: The poorest and most deprived class in Britain. With low levels of economic, cultural and social capital, everyday lives of members of this class are precarious.
TRADITIONAL WORKING CLASS: Contains more older members than other classes but also scores low on all forms of the three capitals. They are not the poorest group.
EMERGENT SERVICE WORKERS: Young and often found in urban areas, this new class has low economic capital but has high levels of 'emerging' cultural capital and high social capital.
TECHNICAL MIDDLE CLASS: A less culturally engaged new class with high economic capital. Small in numbers, they have relatively few social contacts.
NEW AFFLUENT WORKERS: Generally young and active, members have medium levels of economic capital and higher levels of cultural and social capital.
ESTABLISHED MIDDLE CLASS: Not quite elite but members of this class have high levels of all three capitals. They are a gregarious and culturally engaged class.
ELITE: This is the most privileged class in Great Britain who have high levels of all three capitals. Their high amount of economic capital sets them apart from everyone else.



DutchNews.nl

2.3°

Thursday 09 March 2023

The Netherlands has seven social classes, SCP social policy unit concludes

Features

Living in NL: How do you know when it's time to go 'home'?

Vrooman et al. (2023). *Eigentijdse ongelijkheid*. Den Haag: Sociaal en Cultureel planbureau.

https://www.scp.nl/publicaties/publicaties/2023/03/07/eigentijdse-ongelijkheid

Discussion in pairs

Please discuss the following questions with your partner:

1. What could be the *utility* of the groupings created by Savage and others?
2. Everyone appears to find precisely seven (7) social classes. Just thinking about the problem *intuitively*, how would you go about determining this number?
3. Can you come with a way of validating the existence of the classes, outside of the LCA?

In case you need it:

https://www.scp.nl/publicaties/publicaties/2023/03/07/eigentijdse-ongelijkheid

Diagnoses of carcinoma (sample data)

Description

Dichotomous ratings by seven pathologists of 118 slides for the presence or absence of carcinoma in the uterine cervix. Pathologists are labeled `A` through `G`. There were 20 different observed response patterns. This data set appears in Agresti (2002, p. 542) as Table 13.1.

Usage

```
data(carcinoma)
```

Format

A data frame with 118 observations on 7 variables representing pathologist ratings with 1 denoting "no" and 2 denoting "yes".

Source

Agresti, Alan. 2002. *Categorical Data Analysis, second edition*. Hoboken: John Wiley \& Sons.

An extreme case of missing data

```
> dat <- carcinoma %>% mutate_all(as.factor)

> df_freq <- table(dat[, 1:3]) |>
                                    as.data.frame()

> df_freq$X <- latentFactor(NROW(df_freq), 2)
```

An extreme case of missing data

```
> df_freq
  A B C Freq    X
1 1 1 1   36 <NA>
2 2 1 1    2 <NA>
3 1 2 1   16 <NA>
4 2 2 1   19 <NA>
5 1 1 2    0 <NA>
6 2 1 2    1 <NA>
7 1 2 2    0 <NA>
8 2 2 2   44 <NA>
```

Notice that X is **completely missing**!

An extreme case of missing data

```
> library(cvam)
> fit <- cvam(~ A + B + C + X + A:X + B:X + C:X,
              data = df_freq,
              freq = Freq,
              control = list(startValJitter = 0.1))
> summary(fit_cvam)
```

- We are making a (Poisson) regression model that involves a **completely missing** variable!

- Is this really going to work?????

Yes.

	coef	SE
(Intercept)	-1.47196	0.9852
A1	-0.72735	0.7110
B1	-0.72849	0.2517
C1	1.31520	0.6786
X1	0.04025	1.0177
A1:X1	2.26678	0.7154
B1:X1	1.13332	0.2520
C1:X1	1.75226	0.6796

	Models for means		Regression models	
	Latent		*Latent*	
	Continuous	Discrete	Continuous	Discrete
Observed				
Continuous	Factor analysis	**Latent profile analysis**	Random effects	Regression mixture
Discrete	Item response theory	**Latent class analysis**	Logistic ran. eff.	Logistic reg. mix.

Table 1 Names of different kinds of latent variable models.

Latent class analysis is used to...

- “Discover groups” (*DANGER*)
- Estimate the quality of indicators of a latent variable
- Reduce dimensions, smooth tables with many cells
- Account for unobserved heterogeneity, e.g. in regression
- Test substantive hypotheses about (un)observed variables

History across disciplines

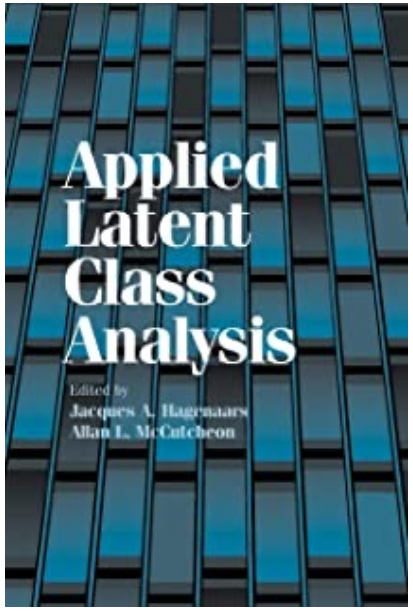
- **Social science**:
  - Lazarsfeld (1950) – introduction of “latent class model”
  - Goodman (1974) – a way to actually estimate such models (!)
  - Dayton and Macready (1988) – including covariates
  - Hagenaars (1988, 1990) – local dependence, categorical SEM
  - Heinen (1996) – restricted LCMs for IRT
  - Magidson and Vermunt (2001) – multiple latent variables
  - Van de Pol and Langeheine (1990); Collins & Lanza (2010) - longitudinal models
  - Vermunt (2003) – multilevel
  - Bolck, Croon, and Hagenaars (2004); Vermunt (2010) - three-step LCM

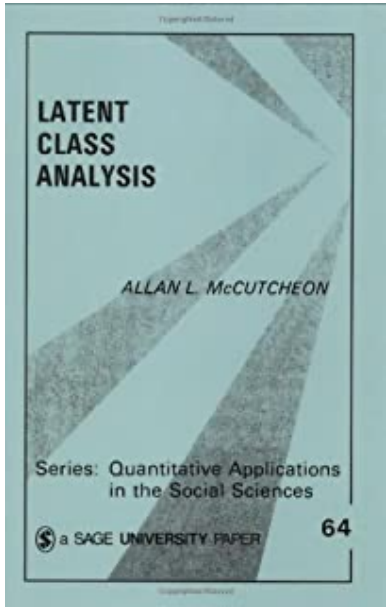
History across disciplines

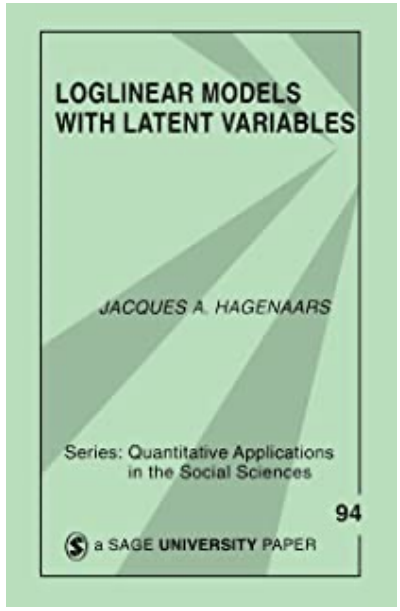
Epidemiology

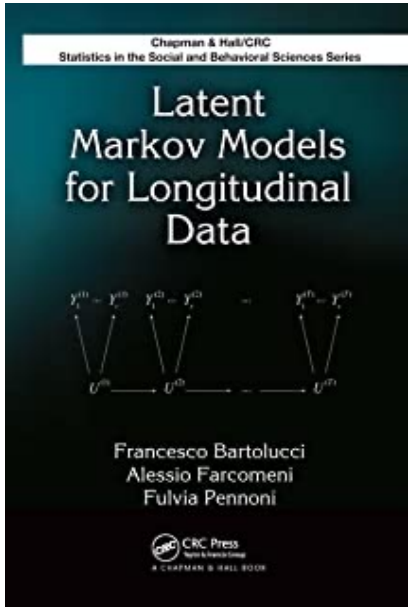
- Young MA. Evaluating diagnostic criteria: a latent class paradigm. J Psychiatr Res. 1982;17(3):285–296.
- Rindskopf & Rindskopf. The value of latent class analysis in medical diagnosis. Stat Med. 1986;5(1):21–27.
- Espeland MA, Handelman SL. Using latent class models to characterize and assess relative error in discrete measurements. Biometrics. 1989;45(2):587–599.
- Formann AK, Kohlmann T. Latent class analysis in medical research. Stat Methods Med Res. 1996;5(2):179–211.
- Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. Stat Methods Med Res. 1998;7(4):354–370.
- Walter SD. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. Epidemiology. 1999;10(1):67–72.
- Garrett ES, Eaton WW, Zeger S. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. Stat Med. 2002;21(9): 1289–1307
- Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Stat Med. 2009;28(3):441–461.

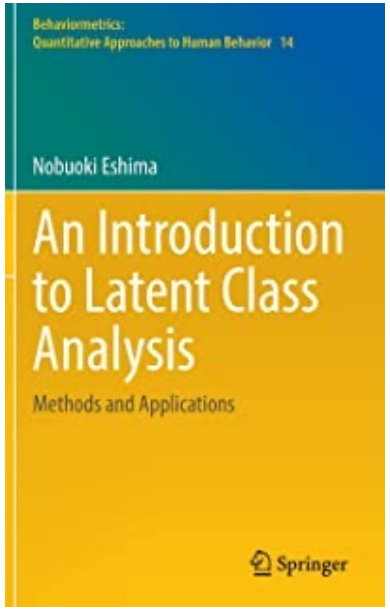
Books


Applied
Latent
Class
Analysis
Edited by
Jacques A. Hagenaars
Allan L. McCutcheon

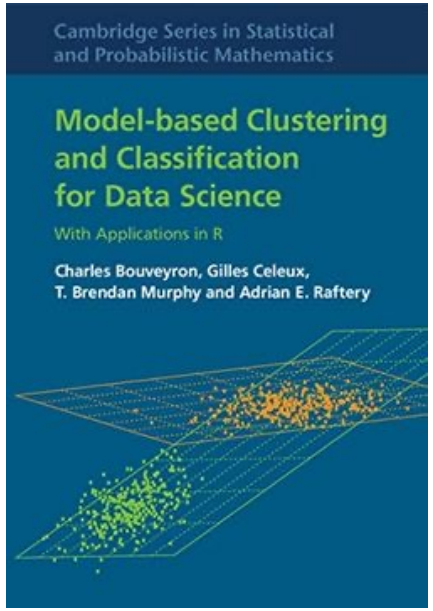

LATENT
CLASS
ANALYSIS
ALLAN L. McCUTCHEON
Series: Quantitative Applications
in the Social Sciences
a SAGE UNIVERSITY PAPER
64


LOGLINEAR MODELS
WITH LATENT VARIABLES
JACQUES A. HAGENAARS
Series: Quantitative Applications
in the Social Sciences
94
a SAGE UNIVERSITY PAPER


Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series
Latent
Markov Models
for Longitudinal
Data
Francesco Bartolucci
Alessio Farcomeni
Fulvia Pennoni
CRC Press


Behaviormetrics:
Quantitative Approaches to Human Behavior 14
Nobuoki Eshima
An Introduction
to Latent Class
Analysis
Methods and Applications
Springer


WILEY SERIES IN PROBABILITY AND STATISTICS
Latent Class and
Latent Transition Analysis
With Applications in the Social, Behavioral,
and Health Sciences
Linda M. Collins
Stephanie T. Lanza
WILEY


Cambridge Series in Statistical
and Probabilistic Mathematics
Model-based Clustering
and Classification
for Data Science
With Applications in R
Charles Bouveyron, Gilles Celeux,
T. Brendan Murphy and Adrian E. Raftery

Software

Latent GOLD (https://www.statisticalinnovations.com/)

- Most advanced features in terms of LCM
- Commercial → free academic license since 2025!!!
- Windows only

Mplus (https://www.statmodel.com/)

- Most popular by far
- Also does (mixture) SEM, very good for such models
- Commercial
- Windows+Mac+Linux

LEM (https://jeroenvermunt.nl/)

- Very powerful, but limited in model size (and not updated since 1997)
- Free ("as in beer", *not* open source)
- Windows only

Software: R

- Definitely *not* user-friendly
- Need to use different packages for different things
- Need to understand better what is going on
- A lot more hands-on
- Free (“as in speech”, i.e. open source *and* “as in beer”)

Oh-so-rewarding!

R packages in the course

`poLCA`

- Probably the most robust and most popular R package for LCA
- Limited to “plain vanilla” LCA, categorical indicators

`mclust`

- Great for continuous indicators (“latent profile analysis”)
- Nice ecosystem from model-based clustering community
- Functionality and use following book of Bouveyron et al.

R packages in the course (cont.)

`flexmix`

- Can fit certain extensions, such as random effects (multilevel) models, growth mixture models etc.
- Limited to single categorical latent variable, but flexible (as the name implies) in terms of what differs over classes

`Cvam → does not exist in the latest version of R`

- Loglinear approach to latent class and missing data
- Very powerful, can fit almost any LCM involving categorical latent and observed variables (quite similar to LEM)
- Sadly limited to small models due to its approach (like LEM)

R packages for specific models

`LMest`

- Specific package for latent Markov models
- Functionality following the book of Bartolucci et al.