Die Methode der latenten Klassenanalyse

Pakete in R

Marcel Gumulak - SoSe 2024

Universität Bielefeld

Inhaltsverzeichnis

1. Einleitung (S. 3 – 7)
2. **poLCA**-Paket (S. 8 – 11)
3. **tidyLPA**-Paket (S. 12 – 14)
4. **depmixS4**-Paket (S. 15 – 17)
5. Weitere Pakete (**tidySEM, OpenMx, randomLCA**) (S. 18 – 19)

Einleitung: Informationen vorweg…

Zu den hier vorgestellten R-Paketen:

- Allgemeine kurze Vorstellung der Pakete
- Besonderheiten und Vor- und Nachteile der einzelnen Pakete
- Allgemeine Veranschaulichung anhand von R-File (LernRaum | GitHub)

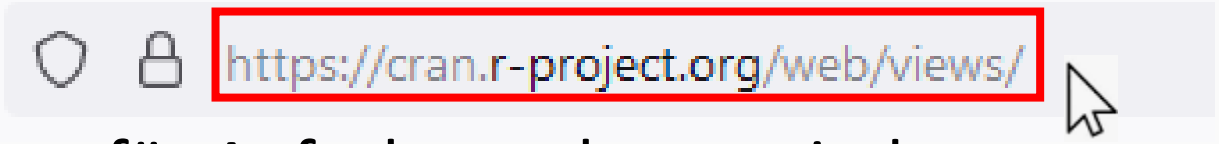
Hier nicht:

- Wie diese im Detail genutzt werden und was zu beachten ist
- Welche möglichen Funktionen in den Paketen existieren

→ **Interesse:** Längere Präsentation per folgenden Link (LernRaum | GitHub)

Einleitung: R-Pakete schnell finden

Nutzung von CRAN task views:



- Orientierungshilfe, welche Pakete für Aufgaben relevant sind
- **Aber:** Dort aufgeführte Pakete stellen nicht die „besten“ Pakete dar

Suche nach Latent Class Analysis R-Paketen:

1. CRAN task views Website aufrufen
2. **Topic:** Unter Psychometrics (alternativ auch: Clustering) schauen
3. **Unter Psychometrics:** Latent Class and Profile Analysis → Pakete

Einleitung: R-Paket Problematik

Die Welt wenn man LCA schnell und leicht in R machen könnte:

Einleitung: R-Paket Problematik (2)

poLCA is still undergoing active development. → **Große Erweiterungen angekündigt, aber kein Update bis heute**

The best way to do latent class analysis is by using Mplus, or if you are interested in some very specific LCA models you may need Latent Gold. Another decent option is to use PROC LCA in SAS. All the other ways and programs might be frustrating, but are helpful if your purposes happen to coincide with the specific R package.

First, there are all kinds of mixture models whose main purpose is to look for the classes in which the regression parameters differ. The meaning of the latent classes here is different as they are based not on the responses of respondents, but on the effects of one variables on other. These packages include flexmix, fpc, mmlcr, lcmm, and others. I do not discuss them.

→ **Modelle mit gleichem Namen, aber unterschiedlichem Modellansatz**

Der Bestseller:

Unfortunately, these approaches are currently only available in commercial software (Latent Gold, Mplus).

Einleitung: R-Paket Problematik (3)

	Classification by	
Design	Means (continuous data) / item probabilities (categorical data)	Regression parameters
Cross-sectional	Latent profile analysis Latent class analysis	Regression mixture models
Longitudinal	Repeated measures latent class analysis Latent transition analysis	Growth mixture models

Oft selbe Bezeichnung, aber unterschiedliche Modelle!

Übersicht – Paket: poLCA

Veröffentlicht:

Drew A. Linzer, Jeffrey B. Lewis am 14. Juni, 2011

Ziel:

Schätzung von latent-class Modellen & latent-class-Regressions-Modellen mit Kovariaten für dichotome und polytome Klassifikations-variablen (Indikatoren)

Verfügbar:

```
install.packages("poLCA")
library(poLCA)
```

poLCA: Vor- und Nachteile

Vorteile:

- Verknüpft simple Bedienung mit brauchbaren Methoden
  → Gut zum Einsteigen und für schnelle/simple Analysen
  → Bietet quasi alle grundlegend benötigten Funktionen
- Bietet neben einfachen latenten Klassenanalysemodellen die Möglichkeit Kovariaten durch Regressionen einzubauen
- Überschaubar strukturierte Ergebnisausgabe & Build-In Grafik
- Beinhaltet mehrere Modellprüfgrößen (AIC, BIC, χ^2, G^2, allg. Entropie)
- Nach Modellschätzung erweiterbar durch zusätzliche Funktionen

poLCA: Vor- und Nachteile (2)

Nachteile:

- Fehlen von konventionellen Funktionen (z.B. Ausgabe von class-means)
  → Nahezu keine (heutzutage zum Standard zählenden) Tests
- Ausschließlich veraltete Schätzmethode: One-Step Verfahren
  → Bei Verwendung von Kovariaten hängt die Klassifikation nicht nur von den Indikatoren, sondern auch den Kovariaten ab (im Vgl. zu Three-Step)
- Stark eingeschränkte Nutzungsmöglichkeiten
  → Schätzfunktion bietet gar keine Restriktionen und nur geringe Modifikation an
  → Kovariateneinfluss kann nur auf Indikatoren modelliert werden
  → Ausschließlich dichotome und polytome Indikatoren verwendbar

poLCA: Variablenvoraussetzung

Damit Funktionen des Pakets fehlerfrei genutzt werden können, müssen bestimmte Voraussetzungen bzw. Eigenschaften bezüglich der genutzten Variablen im übergebenen Datensatz erfüllt sein

1. Indikatoren müssen integer-Werte (aber nicht zwingend vom Typ integer im Datensatz) besitzen:
   - Ganzzahlig und ab dem Wert 1 beginnend
   - Nicht negativ und aufsteigend
2. Kovariaten haben keine (weiteren) Vorgaben (Handhabung wie bei üblichen Regressionen)

poLCA: Anwendungsbeispiel

Enthalten im Paket: `data(cheating, package = "poLCA")`

Quelle: Dayton CM (1998). Latent Class Scaling Analysis. Sage Publications, Thousand Oaks, CA.

Eigenschaften:

- dichotome Antworten von 319 Studenten auf Fragen zum Betrugsverhalten und deren Notendurchschnitt
- 4 Manifeste Variablen: LIEEXAM, LIEPAPER, FRAUD, COPYEXAM
- 1 Kovariate: GPA
- **Besonderheit:** Datensatz erfüllt Variablenvoraussetzungen bereits

	LIEEXAM	LIEPAPER	FRAUD	COPYEXAM	GPA
1	1	1	1	1	NA
2	1	1	1	1	NA
3	1	1	1	1	NA
4	1	1	1	1	NA
5	1	1	1	1	1
6	1	1	1	1	1
7	1	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1
10	1	1	1	1	1

Showing 1 to 11 of 319 entries, 5 total columns

```
> str(cheating) # Kombiniert head() & typeof()
'data.frame':   319 obs. of  5 variables:
 $ LIEEXAM : num  1 1 1 1 1 1 1 1 1 1 ...
 $ LIEPAPER: num  1 1 1 1 1 1 1 1 1 1 ...
 $ FRAUD   : num  1 1 1 1 1 1 1 1 1 1 ...
 $ COPYEXAM: num  1 1 1 1 1 1 1 1 1 1 ...
 $ GPA     : int  NA NA NA NA 1 1 1 1 1 1 ...
```

```
> summary(cheating) # Detailierte Zusammenfassung
    LIEEXAM         LIEPAPER         FRAUD          COPYEXAM         GPA       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :1.000   Median :1.000   Median :1.000   Median :1.000   Median :2.000  
 Mean   :1.107   Mean   :1.119   Mean   :1.066   Mean   :1.213   Mean   :2.327  
 3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:3.000  
 Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :5.000  
                                                                 NA's   :4      
```

Übersicht – Paket: tidyLPA

Veröffentlicht:

Rosenberg, Joshua & Beymer, Patrick & Anderson, Daniel & Schmidt, Jennifer am 10. Oktober, 2018

Ziel:

Schätzung von latent-profile Modellen für stetige (metrische) Klassifikationsvariablen (Indikatoren) unter Spezifikationen von ausgewählten Restriktionen und auf Basis von **mclust** bzw. **MPlus**

Verfügbar:

```
install.packages("tidyLPA")
library(tidyLPA)
```

tidyLPA: Vor- und Nachteile

Vorteile:

- Extrem simple Oberfläche für Schätzung von LPA-Modellen (Wrapper)
- Erlaubt die Spezifikation von ausgewählten Restriktionen
- Umfassende und moderne Fit-Statistiken verfügbar, Link (Weit unten)
- Mehrere Modelle können zeitgleich geschätzt und verglichen werden

Nachteile:

- Nur metrisch skalierte Variablen (→ Normalverteilung) verwendbar
- Keine großen zusätzlichen Funktionen (z.B. keine Kovariaten & FIML)

tidyLPA: Modellschätzung

Das Paket erlaubt es eines der folgenden 4 Modelle zu schätzen:

- Identisch fixierte Varianzen & auf 0 fixierte Kovarianzen (Model 1)
- Variierende Varianzen & auf 0 fixierte Kovarianzen (Model 2)
- Identisch fixierte Varianzen & Kovarianzen (Model 3)
- Variierende Varianzen & Kovarianzen (Model 6)

→ Die Modellparameterisierungen 4 & 5 sind ausschließlich unter Verwendung von MPlus verfügbar

tidyLPA: Anwendungsbeispiel

Enthalten im Paket: `data("pisaUSA15", package = tidyLPA)`

Quelle: OECD PISA Studie 2015, Link

Eigenschaften:

- Schülerfragebogendaten aus der PISA-Studie 2015 in den USA
- Manifeste Variablen (Bsp. auf Zeile 1:100, ohne NAs & instrumental_mot):

1.	broad_interest	Maß für allgemeines persönliche Interesse
2.	enjoyment	Maß für persönlichen Spaß
3.	instrumental_mot	Maß der pers. instrumentellen Motivation
4.	self-efficacy	Maß für Selbstwirksamkeit

Übersicht – Paket: depmixS4

Veröffentlicht:

Visser, Ingmar and Speekenbrink, Maarten am 5. August, 2010

Ziel:

Schätzung von latent-transition Modellen (Hidden Markov Modellen) unter Spezifikationen von Restriktionen und jegliche Art von Indikatoren
→ Latent-class und latent-profile Modelle als Unterart schätzbar

Verfügbar:

```
install.packages("depmixS4")
library(depmixS4)
```

Übersicht – Paket: depmixS4 (2)

	Classification by	
Design	Means (continuous data) / item probabilities (categorical data)	Regression parameters
Cross-sectional	Latent profile analysis	Regression mixture models
	Latent class analysis	
Longitudinal	Repeated measures latent class analysis	Growth mixture models
	Latent transition analysis	

Wir sind eigentlich hier!

depmixS4: Vor- und Nachteile

Vorteile:

- Ermöglicht die Schätzung von Modellen mit fast jeder Variablenart und Kovariaten unter Spezifikation der zugrundeliegenden Verteilung
- Erlaubt die Spezifikation von Restriktionen auf Parameterebene

Nachteile:

- Keine Fit-Statistiken, Tests und ordinalen Variablen möglich
- Andere Terminologie als Grundlage (z.B. Classes = States)
- (Fast) Gar keine zusätzlichen Funktionen (Grafiken, Entropie, etc.)

Weitere Pakete

Die meisten Pakete sind im Zusammenhang mit spezifischen Studien entstanden. Allgemeinere Alternativen in Paketform sind hingegen:

- **tidySEM** Kann als Verallgemeinerung von tidyLPA angesehen werden und ermöglicht LCA & LPA auf ausschließlich einem von mehreren Variablentypen (z.B. nur ordinal)

→ Wrapper-Funktionen für leichtere Verwendung auf Basis von OpenMX

→ Restriktionen wie in tidyLPA spezifizierbar & allgemein mehr Umfang im Paket

- **OpenMX** Paket bietet (lediglich) einen Rahmen für die Erstellung von LCA & LPA Modellen jeglicher Form & Variablentypen

→ **Aber:** Algorithmus oder Programm praktisch selbst schreiben (High Know-How)

Weitere Pakete (2)

Die meisten Pakete sind im Zusammenhang mit spezifischen Studien entstanden. Allgemeinere Alternativen in Paketform sind hingegen:

- **randomLCA** Ermöglicht es LCA mit Random Effects und bis zu 2 Ebenen zu modellieren, aber sehr eingeschränkt

➢ Bei Verletzungen der lokalen Unabhängigkeit wird ein normalverteilter Random Effect anstelle von zusätzlichen Klassen modelliert, aber benötigt auch mehr Parameter und damit geringere Freiheitsgrade

➢ Beinhaltet Bestrafungsterm bei der Likelihood der Wahrscheinlichkeiten für bessere Standardfehlerberechnung und Bootstrap Standardfehler-Funktion

➢ Random Effect- und Standardmodell nicht genestet (Unvergleichbar)

Quellenangabe

- Rudnev, M. (2016, December 28). *Ways to do Latent Class Analysis in R*. Elements of Cross-cultural Research. Link
- Mair, P., Rosseel, Y., Gruber, K. (2023, December 15). *Latent Class and Profile Analysis*. CRAN TASK VIEW: Psychometric Models and Methods. Link
- Bacher, J.; Pöge, A.; Wenzig, K. (2010): Clusteranalyse. Anwendungsorientierte Einführung in Klassifikationsverfahren. 3. Aufl. München.
- Bauer, J. (2022). A Primer to Latent Profile and Latent Class Analysis. In: Goller, M., Kyndt, E., Paloniemi, S., Damşa, C. (eds) Methods for Researching Professional Learning and Development. Professional and Practice-based Learning, vol 33. Springer, Cham. Link
- Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software*, *42*(10), 1–29. Link, R-Manual

Quellenangabe (2)

- Rosenberg et al., (2018). tidyLPA: An R Package to Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software. Journal of Open-Source Software, 3(30), 978, Link, R-Manual, Supplemental Material
- Visser, I., & Speekenbrink, M. (2010). depmixS4: An R Package for Hidden Markov Models. *Journal of Statistical Software*, *36*(7), 1–21. Link, R-Manual
- Van Lissa, C. J., Garnier-Villarreal, M., & Anadria, D. (2024). Recommended Practices in Latent Class Analysis Using the Open-Source R-Package tidySEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(3), 526–534. Link, Supplemental Material
- Neale, M.C., Hunter, M.D., Pritikin, J.N. *et al.* OpenMx 2.0: Extended Structural Equation and Statistical Modeling. *Psychometrika* 81, 535–549 (2016). Link, R-Manual, Website
- Beath, K. J. (2017). randomLCA: An R Package for Latent Class with Random Effects Analysis. *Journal of Statistical Software*, *81*(13), 1–25. Link, R-Manual