

Lineaire structurele modellen

College 2: Dimensionele analyse

Harry Ganzeboom
MA-cursus
9 februari 2009

Agenda

- Practicum / Opdracht 1 doorspreken
- Dimensionele analyse
- Deze week is er een formele Opdracht 1
- Ineke Nagel geeft woensdag het practicum
- Deadline: zaterdagavond, 24:00 uur
- Volgende week: Carmines & Zeller, part 1+3.

LS2: Dimensionele analyse

2

Dimensionele analyse

- Dimensionele analyse hebben betrekking op variabelen ('indicatoren') waarin geen oorzaak-gevolg-structuur zit, maar die beschouwd worden als metingen van dezelfde 'dimensie'.
 - Hoeveel dimensies gaan er schuil in de voorliggende verzameling indicatoren?
 - Welke relatie (=correlatie) hebben deze dimensies met elkaar?
 - Hoe kan ik het beste een score op deze dimensie construeren?
 - Met welke kwaliteit meet deze score de onderliggende dimensie?

LS2: Dimensionele analyse

3

Factoranalyse - geschiedenis

- Factoranalyse is een in de psychologie ontwikkelde statistische technieken, nu meer dan 100 jaar oud.
- Ontstaansgrond is het intelligentie-onderzoek. Historisch belangrijk is de ruzie tussen Britse psychologen (Spearman) en Amerikaanse (Thurstone), die een verschillende theorie over de opbouw van intelligentie hadden:
 - Spearman: er is een algemene intelligentie (G-factor), die ten grondslag ligt aan verschillende soorten speciale intelligentie (zoals rekenkundige en taalkundige vaardigheid). Dit leidt tot een hiërarchische opbouw en zgn. 2e orde factoranalyse.
 - Thurstone: verschillende soorten intelligentie kunnen los van elkaar variëren. Dit leidt tot centrale betekenis van 'simple structure' en het 'varimax' model.
- In SPSS heeft het Amerikaanse model gewonnen. In LISREL kun je allebei doen.

LS2: Dimensionele analyse

4

Factoranalyse -- eigenaardigheden

- Doordat factoranalyse zich wat opzij van de rest van de statistiek heeft ontwikkeld (economen en biologen hebben zich er nooit mee bemoeid), zit het vol met eigenaardig jargon.
- Extractie, rotatie, eigenwaarden, componenten, factoren, communaliteit, lading: allemaal woorden die je elders niet of nauwelijks tegenkomt (of met een andere betekenis...).
- Toch hebben factoranalyses een sterke relatie met correlatie- en regressieanalyse.

LS2: Dimensionele analyse

5

PCA en PAF

- Onder het naam factor-analyse [in SPSS] verschuilen zich meerdere technieken van dimensionele analyse. Twee ervan zijn:
 - PCA: (principale) componenten analyse
 - PAF: principal axis factoring
- PAF en PCA zijn in aanleg heel verschillende technieken...
- ... maar de uitkomsten liggen vaak dicht tegen elkaar aan.
- Vandaar dat het verschil ons doorgaans niet zoveel kan schelen; het verschil is echter nuttig om betrouwbaarheidsanalyse en lisrel-modellen goed te kunnen interpreteren.

LS2: Dimensionele analyse

6

PCA (1)

- Component: hoe kan ik de scores op de indicatoren zo sommeren dat er een indexvariabel met maximale variantie ontstaat?
- $VAR(A+B) = VAR(A) + VAR(B) + 2 \cdot COVAR(A,B)$
- Het antwoord is: geef een groot gewicht aan de indicatoren die het hoogst met elkaar correleren.
- PCA is direct gericht op het construeren van een samenvattende componentscore.

LS2: Dimensionele analyse

7

PCA (2)

- Hoe komt PCA aan de opeenvolgende componenten?
 - Regresseer alle geobserveerde variabelen op de eerste component.
 - De residuen van deze regressies zijn niet gecorreleerd ('orthogonaal') ten opzichte van de eerste component.
 - Bereken nu de tweede component als de somscore van de residuen die in deze fase nog maximale variantie heeft.
 - Etcetera, tot het aantal componenten gelijk is aan het aantal indicatoren; deze hebben telkens een kleinere variantie.

LS2: Dimensionele analyse

8

PCA (3)

- Componentenmatrix: geeft correlaties tussen de componenten en de indicatoren.
- Als je deze correlaties kwadrateert (= verklaarde variantie) en per kolom sommeert, krijg je de (aflopende) eigenwaarden.
- Vandaar het stopcriterium eigenwaarde < 1.0 : de door een component verklaarde variantie zou kleiner zijn dan die van een enkele (gestandaardiseerde) indicator.
- Ook de knikplot ('screeplot') en het criterium knik-1 wordt duidelijk: vanaf de knik is het alleen maar ruis.

LS2: Dimensionele analyse

9

Rotatie en PCA

- Rotatie en zeker scheve rotatie is binnen PCA wat minder begrijpelijk.
- Rechte rotatie: het volgens 'simple structure' ('varimax') zodanig manipuleren van de somscores dat er eenvoudig interpreteerbare deelverzamelingen indicatoren worden opgeteld.
- Scheve rotatie: daarbij ook nog eens toestaan dat de berekende componenten met elkaar gecorreleerd zijn.
- In de praktijk gebruik je PCA vaak om een groepje indicatoren waaronder 1 dimensie schuilgaat, te identificeren en bereken je daaruit alleen de eerste componentscore.

LS2: Dimensionele analyse

10

PCA en indexconstructie

- Als je via PCA en (scheve) rotatie een groepje indicatoren hebt gevonden die een enkele dimensie vormen, ligt het voor de hand om de analyse te herhalen met alleen die indicatoren.
- Je kunt de somscore dan zien als een 'optimaal' geconstrueerde indexvariabele.

LS2: Dimensionele analyse

11

PCA en betrouwbaarheidsanalyse

- Er is een sterk verband tussen PCA en betrouwbaarheidsanalyse via het interne consistentiemodel (Cronbach's alpha).
- Merk op dat de tabel (what happens if item deleted?) van Cronbach's alpha ook een kolom heeft over de variantie van de somscore.
- Cronbach's alpha veronderstelt (we zullen nog zien hoe) dat alle indicatoren gelijkmatig gecorreleerd zijn (en alle gewichten aan elkaar gelijk zijn).
- Dit is een bijzondere variant van het PCA-model: het geval waarin alle indicatoren gelijke gewichten krijgen.

LS2: Dimensionele analyse

12

Missing values

- Een wat ongemakkelijke zaak bij dimensionele analyse (zowel PCA als PAF) is hoe missing values verwerkt worden (deze treden vaak op!):
 - Listwise ('complete cases'): je krijgt alleen een score voor cases zonder MV.
 - Mean substitution: je krijgt een score voor alle cases, waarbij alle MV zijn vervangen door het gemiddelde.
- Beide opties zijn onbevredigend. Beter werkt het SPSS statement COMPUTE ... MEAN, dat uitgaat van de beschikbare informatie. Dit statement is erg handig, maar sluit niet direct aan bij de resultaten van component- en betrouwbaarheidsanalyse.

LS2: Dimensionele analyse

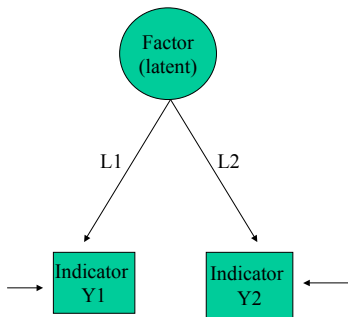
13

PAF (1)

- (Echte) factoranalyse heeft als vraagstelling: hoe kan ik de correlaties tussen de indicatoren verklaren vanuit een causaal model met een of of meerdere latente variabelen?
- Dit idee sluit rechtstreeks aan bij het 'klassieke meetmodel': indicator = true score + residu.

LS2: Dimensionele analyse

14



LS2: Dimensionele analyse

15

De vermenigvuldigregel

- Een PAF model is een gewoon causaal model en voldoet aan de gewone regels voor indirecte en confounding effecten.
- In het bijzonder:
 - $r(Y1, Y2) = L1 * L2$
- Als $L1 = L2$, geldt: $L1 = L2 = \sqrt{r}$

LS2: Dimensionele analyse

16

PAF (2)

- De 'patternmatrix' geeft de correlaties (directe effecten) aan tussen latente dimensies en indicatoren. (De 'structure matrix' geeft de totale effecten – deze zijn doorgaans niet interessant).
- De effecten ('ladingen') zijn hoog voor twee variabelen A en B als A en B hoog gecorreleerd zijn; PAF en PCA komen in dit opzicht vaak op hetzelfde neer.

LS2: Dimensionele analyse

17

PAF (3)

- Bij twee indicatoren is de PAF vraagstelling alleen oplosbaar als je veronderstelt dat de beide meetrelaties L1 en L2 identiek zijn.
- Bij drie indicatoren is PAF precies geïdentificeerd.
- Bij meer indicatoren is PAF overgedetermineerd en kunnen we de aanwezigheid van meer dan een latente dimensie onderzoeken.
- Anders dan bij PCA is meerdimensionaliteit en correlatie tussen dimensies ('scheve rotatie') bij PAF geen moeizaam idee.

LS2: Dimensionele analyse

18

PAF (4)

- Ook 'simple structure' is vanuit PAF geen noodzakelijk idee.
 - Een indicator die wordt beïnvloed door meerdere latente dimensies is heel goed interpreteerbaar.
 - Hiërarchische opbouw van dimensies (het Britse model structuur van IQ) is goed te begrijpen.
- Alles bij elkaar genomen is PAF wat gemakkelijker interpreteerbaar dan PCA.

Maar ...

- PAF is rekenkundig niet helemaal stabiel en gevoelig voor modelafwijkingen. Als je model niet goed bij de data past, loopt de techniek af en toe vast.
- PAF leidt niet tot een indexscore. Hoewel je in SPSS wel een 'factorscore' kunt opvragen, gedraagt deze zich niet helemaal volgens het model.
- Om deze redenen is PAF minder populair dan PCA.

PAF en LISREL

- PAF is een explorerende variant van het algemene LISREL model.
- In beide gevallen probeer je een correlatiematrix van indicatoren te verklaren via een beperkt aantal latente variabelen:
 - In LISREL kies je de locatie van de ladingen (je kiest daarmee de 'rotatie'); in PAF laat je dit aan het (blinde) varimax (=simple structure) criterium over.
 - In LISREL kunnen de latente dimensies een causale structuur hebben, in PAF hebben ze alleen een correlatieve structuur.