Hoorcollege 4: PCA en FA

PCA en FA

Principale Componenten Analyse en Factor Analyse zijn analyses waarbij het reduceren van data het doel is. Beiden zijn zij ontwikkeld in samenspraak met de psychologie, in tegenstelling met tot de biologie. Bij *datareductie* wordt een grote set variabelen verkleind tot een veel kleinere set onderliggende dimensies. Dit kan nuttig zijn om overzicht te creëren en om te veel overlap te voorkomen tussen de dimensies. Een voorbeeld hiervan is alle persoonlijkheidseigenschappen reduceren tot de Big Five. Deze variabelen moeten van interval niveau zijn. Slechts bij de PCA mogen de variabelen dichotoom zijn.
Deze week worden PCA en *exploratieve* factoranalyse (EFA) besproken, waarbij er geen exacte vraag wordt gesteld. Van te voren heeft de onderzoeker wel bepaalde verwachtingen, echter bevinden deze verwachtingen zich in vroege fase van het onderzoek. Het onderzoek dient de techniek gedeeltelijk open te laten. Het verschil tussen Principale Componenten Analyse en Factor Analyse is dat *Factor Analyse* een expliciet model heeft. Hiermee wordt bedoeld dat het model zo in elkaar zit dat de scores op de geobserveerde variabelen worden verklaard in dit expliciete model door niet direct observeerbare, ofwel: latente variabelen. Bij Factor Analyse gaan we met een gedetailleerde hypothese bekijken of het idee wat we over de data hebben ook daadwerkelijk klopt. Bij de *Principale Componenten Analyse* wordt alleen een a-theoretische 'herschrijving' maakt van variabelen tot componenten. Bij Principale Componenten Analyse gaan we kijken naar de data zonder verwachtingen. Er wordt dus alleen geobserveerd. PCA heeft geen beschrijving voor error. Bij PCA hoopt men dat '*errors'* verdwijnen naar hogere dimensies, terwijl de Factoranalyse een expliciet model voor error heeft.

Functie van PCA

PCA kunnen we op zowel een algebraïsche manier als een geometrische manier bekijken. Als we kijken naar de algebraïsche manier is een *principale component* een lineaire combinatie van variabelen. De eerste component moet zoveel mogelijk variantie verklaren van de variabelen. Zo komt de eerste component het dichtst in de buurt met het beschrijven van de variabelen. Ieder component dat hierop volgt probeert ook zoveel mogelijk variantie te verklaren, maar is totaal niet gecorreleerd aan de voorafgaande component (*orthogonaliteit*). Hierdoor zullen een aantal componenten een overgroot deel van de variantie verklaren en kunnen de belangrijkste componenten geselecteerd worden. De data zijn nu gereduceerd.

Als we PCA op de geometrische manier bekijken zijn de componenten gelijk aan *vectoren*. Hoe meer deze naar rechts of naar boven gelegen is, hoe hoger de score op een van de componenten. De opvolgende vector is niet gecorreleerd aan de voorafgaande vector en staat daarom loodrecht op de eerste vector. Het is lastig om dit model te maken voor meer dan 2 variabelen, omdat er dan een ruimtelijke structuur ontstaat.

Communaliteit en componentlading

C_{ij} is de *componentlading*. Dit is de correlatie van variabele X_i met component j. Wanneer de componentlading gelijk is aan 0, hebben de variabele en het component niets met elkaar gemeen. Als je de componentlading kwadrateert (Cij^2) krijg je de proportie variantie van variabele X_i verklaard door component j.

De 'uniekheid' van een variabele wordt aangegeven met de *communaliteit.* Hoe lager de

communaliteit, hoe unieker de variabele. Het is de proportie verklaarde variantie per variabele. Verder is het de som van gekwadrateerde componentladingen. In deze formule is k het maximaal aantal componenten, h_i^2 is de communaliteit, Cij^2 is de componentlading in het kwadraat.

$$h_i^{\,2} = \sum_{i=1}^{k} c_{ij}^2$$

Eigenwaarde

De *eigenwaarde* van een variabele is de som van de gekwadrateerde componentladingen per component. Het is de hoeveelheid verklaarde variantie van alle variabelen bij elkaar. Als dit wordt gedeeld door het aantal variabelen krijgt men de proportie verklaarde variantie.

Criteria voor het aantal componenten

Om te bepalen tot hoeveel componenten men het best kan reduceren, zijn er een aantal richtlijnen, namelijk:

- De eigenwaarde moet groter zijn dan 1. Je kunt alle componenten gebruiken die een eigenwaarde hebben die groter is dan 1.
- Een grafiek met daarin een lijn die componenten aangeeft en hun bijbehorende eigenwaarde. We zien dat de lijn erg snel afneemt met het toenemen van het aantal componenten. Als we een grens willen stellen voor het aantal componenten die gebruikt kunnen worden voor de analyse, moet er gekeken worden naar de knik in de lijn. Dit is een vage omschrijving, maar meestal geeft dit wel het juiste antwoord. Vaak kan er ook 1 component meer of minder worden gebruikt.
- Interpreteerbaarheid: alle oplossingen bekijken en dan de oplossing kiezen waar je het beste een verhaal van kunt maken, is de meest begrijpelijke/praktische oplossing. Dit is een hele vage methode.

Interpretatie

Componentladingen kunnen worden gebruikt voor de interpretatie van een PCA-oplossing. Als dit algebraïsch wordt gedaan, moeten de ladingen worden onderstreept met de absolute waarde boven een grenswaarde (meestal wordt de grenswaarde 0.40 gehanteerd, maar dit kan verschillen). Daarna moet worden bepaald voor variabelen met hoge ladingen op hetzelfde component of deze wat gemeenschappelijk hebben, en of dit gemeenschappelijke deze variabelen onderscheidt van variabelen die niet op de component laden.

Als er meetkundig wordt geïnterpreteerd maakt men een grafiek waarin de variabelen als vectoren in de componentenruimte staan. De lijn loopt vanuit de oorsprong naar het punt van componentenlading. Er wordt dan gekeken naar de verschillende lengtes (hoe langer de vector, hoe beter de variabele verklaard wordt) en naar de hoek (hoe scherper de hoek tussen de vectoren, hoe hoger de correlatie tussen de variabelen).

Rotatie

Wanneer de hoek tussen de verschillende vectoren klein is, zullen de vectoren meer met elkaar correleren. Een hoek kleiner dan 90 graden geeft een correlatie aan. Als de hoek gelijk is aan 90

graden, zullen de componenten die zijn uitgebeeld via de vectoren onderling niet correleren. Bij rotatie wordt er anders gekeken naar de oplossing: het assenstelsel wordt veranderd. Zo krijg je een *'simple structure'* (de meest ideale situatie) waarbij een interpretatie makkelijker wordt. VARIMAX is de meest gebruikte rotatie, waarbij nieuwe assen worden gekozen op zo'n manier dat varianties van gekwadrateerde factorladingen per factor zo hoog mogelijk zijn. Voorbeelden met SPSS zijn te zien in de slides vanaf dia 19.