

## Enige statistische aspecten van de factoranalyse \*

door G. de Leve

### Summary

Some statistical aspects of factor analysis.

*In this report attention is paid to the hypotheses of Factor Analysis. The author makes a distinction between Component Analysis and Factor Analysis. Some statistical objectives of Factor Analysis are formulated and a summary of some known statistical techniques of Factor Analysis is given.*

### 1. Inleiding

In dit korte overzicht zullen wij enkele statistische aspecten van de factoranalyse bespreken. Alvorens wij deze technieken afzonderlijk behandelen, willen wij eerst enige aandacht besteden aan een probleem, waarbij de factoranalyse veelvuldig wordt toegepast. De toepasbaarheid van de factoranalyse blijft echter geenszins tot dit voorbeeld of zelfs tot het gebied, waaraan dit ontleend is, beperkt.

Stel dat wij gelijktijdig aan  $N$  personen een serie van  $n$  verschillende psychologische tests afnemen. Met behulp van deze testresultaten kunnen de correlaties tussen de verschillende tests berekend worden. Wanneer nu blijkt, dat deze testvariabelen hoog gecorreleerd zijn, dan rijst direct de vraag of er misschien factoren zijn (b.v. karaktereigenschappen), die meer dan één test beïnvloeden. Men zou dan de reactie van de persoon op een test kunnen beschouwen als een functie van een aantal wellicht psychologisch interpreteerbare factoren, waarvan er één of meer bij meer dan één test een rol spelen. In formule:

$$(1) \quad Z_j = Z_j(X_1, \dots, X_p) \quad (j = 1, \dots, n)$$

waarbij  $p$  het aantal factoren aangeeft,

$Z_j$  de functie is behorende bij de  $j$ -de test en  
 $X_i$  de  $i$ -de factorvariabele voorstelt.

Als vervolgens  $Z_{jk}$  de score is, welke de  $k$ -de persoon in de  $j$ -de test behaalt en  $X_{ik}$  de waarde van de  $i$ -de factor bij de  $k$ -de persoon, dan geldt voor de scores  $Z_{jk}$ :

$$(2) \quad Z_{jk} = Z_j(X_{1k}, \dots, X_{pk}) \quad (j = 1, \dots, n; k = 1, \dots, N).$$

\*) Rapport S 210 van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam. De afdeling staat onder leiding van Prof. Dr. D. van Dantzig.

De vragen, waarop men een antwoord tracht te vinden, kunnen als volgt worden samengevat:

- 1e. Hoeveel factoren  $X_j$  spelen bij meer dan één test een rol.
- 2e. Hoe is de gedaante van de functie

$$Z_j = Z_j(X_1, \dots, X_p) \quad (j = 1, \dots, n),$$

- 3e. Welke waarden nemen de variabelen  $X_j$  aan voor de  $k$ -de persoon

$$(i = 1, \dots, m; k = 1, \dots, N).$$

De technieken, die wij later zullen bespreken, gaan uit van de *veronderstelling*, dat de functies  $Z_j$  *lineair* zijn, m.a.w. dat voor (1) geschreven kan worden:

$$(3) \quad Z_j = a_{j1}X_1 + \dots + a_{jp}X_p \quad (j = 1, \dots, n),$$

waarbij de constanten  $a_{ji}$  ook de waarde nul mogen aannemen.

De betrekking (2) gaat dan over in:

$$(4) \quad Z_{jk} = a_{j1}X_{1k} + \dots + a_{jp}X_{pk} \quad (j = 1, \dots, n; k = 1, \dots, N).$$

De factoren, die bij meer dan één test een rol spelen, zullen wij in het vervolg aangeven met de letter  $F$  en *gemeenschappelijke factoren* noemen. Voor het geval er  $m$  gemeenschappelijke factoren zijn, gaat de betrekking (4) over in:

$$(5) \quad Z_{jk} = a_{j1}F_{1k} + a_{j2}F_{2k} + \dots + a_{jm}F_{mk} + a_j S_{jk}$$

waarbij  $a_j S_{jk}$  dat deel van de score aangeeft, dat niet „verklaard” kan worden door de bijdragen van de gemeenschappelijke factoren. De variabelen  $S_j$  worden in het vervolg *specifieke factoren* genoemd.

Indien er geen specifieke factoren worden ondersteld, gaat (5) over in:

$$(6) \quad Z_{jk} = a_{j1}F_{1k} + a_{j2}F_{2k} + \dots + a_{jm}F_{mk} \quad \left( \begin{array}{l} j = 1, \dots, n \\ k = 1, \dots, N \end{array} \right).$$

Het getal  $Z_{jk}$  in betrekkingen (5) en (6) stelt eigenlijk niet de gevonden testscore van de  $j$ -de test bij de  $k$ -de persoon voor, maar de mathematische verwachting daarvan. De gevonden testscore zal hier ongetwijfeld van verschillen, daar er altijd bijkomstige omstandigheden zijn, die het resultaat van de test beïnvloeden.

Aangezien het model aangegeven in (5) en (6) berust op de veronderstelling, dat de waarden  $F_{ik}$  en  $S_{jk}$  constant zijn, zal men bij verwerking van de

gevonden testcores in het model extra termen moeten toevoegen om de stochastische afwijkingen van de verwachting te kunnen verklaren.

Voor de gevonden testcores gelden dan in plaats van (5) en (6) de volgende betrekkingen:

$$(7) \quad Z_{jk} = a_{j1}F_{1k} + a_{j2}F_{2k} + \dots + a_{jm}F_{mk} + a_j S_{jk} + e_{jk} \\ (j = 1, \dots, n ; k = 1, \dots, N)$$

en

$$(8) \quad Z_{jk} = a_{j1}F_{1k} + \dots + a_{jm}F_{mk} + e_{jk} \\ (j = 1, \dots, n ; k = 1, \dots, N).$$

Daar het echter bij de te bespreken technieken, waarbij de aanwezigheid van specifieke factoren wordt verondersteld, niet mogelijk is een onderscheid te maken tussen deze factoren en de storingstermen, zal men toch uitgaan van het model aangegeven in (5), waarbij dan de waarden  $S_{jk}$  niet meer constant behoeven te zijn.

De technieken, die worden toegepast op het model aangegeven in (8) vormen de *Componentenanalyse*.

De technieken, die uitgaan van het model aangegeven in (5) behoren tot de *Zuivere Factoranalyse*.

De componentenanalyse en de zuivere factoranalyse vormen tezamen de factoranalyse. De verschillende interpretatie van de laatste termen (resp.  $e_{jk}$  en  $a_j S_{jk}$ ) als stochastische afwijkingen specifieke factoren (de laatste desgewenst als som van specifieke factoren en stochastische afwijkingen) leidt tot verschil in de mogelijkheden van analyse. In het eerste geval kan men trachten door het verrichten van duplo-bepalingen (indien deze mogelijk zijn) iets over de verdeling der stochastische afwijkingen te weten te komen en op grond hiervan het aantal factoren  $m$  te toetsen. In het laatste geval hebben duplo-bepalingen weinig zin, daar de specifieke factoren bij beide bepalingen dezelfde waarden aannemen en de variantie der duplo-bepalingen slechts op de, feitelijk in de specifieke factoren opgenomen, stochastische afwijkingen betrekking heeft. Voor de bepaling van het aantal gemeenschappelijke factoren moet men dan andere criteria gebruiken. Men kan het verschil ook als volgt karakteriseren. Afgezien van de stochastische afwijkingen  $e_{jk}$  zijn de  $Z_{jk}$  van (8) *algebraïsch* lineair afhankelijk (de lineaire betrekkingen worden verkregen door eliminatie van de factoren  $F_i$  en zij gelden dan voor iedere  $k$ ), terwijl dit voor (5) niet geldt.

Aan de variabelen  $F_i$  en  $S_j$  in (5) en (8) worden de volgende beperkingen, — die de algemeenheid niet schaden — opgelegd:

$$(9) \quad \begin{array}{ll} \sum_{k=1}^N F_{ik} = 0, & \sum_{k=1}^N S_{jk} = 0, \\ \frac{1}{N} \sum_{k=1}^N F_{ik}^2 = 1, & \frac{1}{N} \sum_{k=1}^N S_{jk}^2 = 1. \end{array}$$

Men kan gemakkelijk inzien, dat de gegevens verschaft door de testcores onvoldoende zijn om de waarden  $a_{ji}$ ,  $a_j$ ,  $F_{ik}$  en  $S_{jk}$  ondubbelzinnig vast te leggen. Deze grootheden zijn binnen het model zoals het nu is *niet* identificeerbaar. Immers ook de waarden, welke de factorvariabelen  $F_i$  en  $S_j$  aannemen voor de verschillende personen zijn onbekend. Men zal nu aan de variabelen  $F_i$  en  $S_j$  extra eigenschappen moeten toekennen om identificeerbaarheid te verkrijgen en deze eigenschappen zijn in de te bespreken technieken statistisch van karakter.

Deze weg wordt gevolgd, omdat directe meting der factoren niet mogelijk is. Was dit wel zo, dan zou men de zoveel eenvoudigere regressie-analyse kunnen toepassen.

Het identificeerbaar maken, dat nu dus wel door het opleggen van extra eigenschappen moet geschieden, geeft geen enkele garantie omtrent praktische interpreteerbaarheid der uiteindelijk verkregen ondubbelzinnige oplossing. Het is uiteraard zeer wel mogelijk, dat één of meer der factoren of lineaire combinaties daarvan, gerepresenteerd door de gevonden getalwaarden (schattingen), een praktische interpretatie bezitten, doch het is al te optimistisch hierop zonder meer te rekenen. Alleen indien een factor, waarvoor men een praktische interpretatie meent te hebben gevonden, achteraf toch direct meetbaar blijkt te zijn, kan men tot verificatie van het resultaat overgaan. Dit is echter een uitzonderlijke situatie.

In verschillende handboeken over factoranalyse worden niettemin aanvullende methoden beschreven met behulp waarvan men, uitgaande van de op bovenstaande wijze verkregen oplossing, toch het gestelde doel — het schrijven van de testvariabelen als lineaire combinaties van variabelen, behorende bij factoren met psychologisch interpreteerbare betekenis — meent te kunnen bereiken.

Wanneer men in een  $m + n$  dimensionale ruimte op de assen van een orthogonaal coördinatensysteem de variabelen  $F_i$  en  $S_j$  uitzet (één punt voor iedere  $k$ ), dan leiden deze aanvullende methoden tot een coördinatentransformatie in de door de  $m$  gemeenschappelijke factoren voortgebrachte deelruimte. Op grond van bepaalde a priori opvattingen over het verband tussen de variabelen  $Z_j$  en de factorvariabelen komt men dan tot een nieuwe ondubbelzinnige schrijfwijze, waaraan men dan een psychologische interpretatie geeft. Tot een directe verificatie van deze interpretatie leiden deze methoden

echter niet. Wij zullen ze hier, in verband met hun zeer specialistische karakter, niet bespreken.

Wij kunnen dus onze  $n \times N$  waarnemingen plaatsen in één van de volgende modellen:

$$(8) \quad Z_{jk} = \sum_{i=1}^m a_{ji} F_{ik} + e_{jk} \quad (\text{Componentenanalyse})$$

$$(5) \quad Z_{jk} = \sum_{i=1}^m a_{ji} F_{ik} + a_j S_{jk} \quad (\text{Zuivere Factoranalyse})$$

De statistische doelstellingen van de factoranalyse kunnen nu als volgt worden geformuleerd:

- 1e. Het opstellen van een hypothese omtrent het minimale aantal gemeenschappelijke factoren en het toetsen van deze hypothese.
- 2e. Het schatten van, en het geven van betrouwbaarheidsintervallen voor, de constanten  $a_{ji}$  en  $a_j$ .
- 3e. Het schatten van, en het geven van betrouwbaarheidsintervallen voor, de factorwaarden  $F_{ik}$ .

De technieken, die gebruikt worden zowel in de Componentenanalyse als in de Zuivere Factoranalyse, verschillen van elkaar doordat de extra eigenschappen, welke men aan de variabelen  $F_i$  en  $S_j$  heeft opgelegd, voor iedere techniek niet dezelfde zijn.

## 2. Componentenanalyse

Zoals wij reeds eerder hebben vastgesteld, dienen aan de variabelen  $F_i$  extra eigenschappen te worden toegevoegd om tot een ondubbelzinnige uitkomst te komen.

Het opleggen van de extra eigenschappen geschiedt op de meest natuurlijke wijze in de populatie, waaruit de proefpersonen een steekproef vormen. Om tot eenvoudige schattingsmethoden te komen worden deze eigenschappen veelal onveranderd opgelegd binnen de steekproef. Het onderscheid hier-tussen wordt in vele boeken en artikelen over factoranalyse verwaarloosd.

De technieken behorende tot de Componentenanalyse gaan dus uit van het volgende model:

$$(8) \quad Z_{jk} = \sum_{i=1}^m a_{ji} F_{ik} + e_{jk}$$

De oudste methode, welke wij zullen bespreken is afkomstig van H. Hotelling [1, 2]. In het begin van zijn beschouwing neemt hij aan, dat er eventueel gemeenschappelijke factoren zijn als variabelen  $Z_j$  en negeert hij het bestaan van meetfouten.

Zijn model ziet er dan als volgt uit:

$$(9) \quad Z_{jk} = \sum_{i=1}^n a_{ji} F_{ik} \quad (j = 1, \dots, n; k = 1, \dots, N).$$

Alvorens aan te geven welke extra eigenschappen Hotelling oplegt aan de variabelen  $F_i$ , zullen wij zijn methode eerst meetkundig toelichten.

Om te beginnen worden de  $Z_{jk}$  gestandaardiseerd over  $k$ , d.w.z. verminderd met  $Z_j = \frac{1}{N} \sum_{k=1}^N Z_{jk}$  en gedeeld door de spreiding  $\sqrt{\frac{1}{N} \sum_{k=1}^N (Z_{jk} - Z_j)^2}$ .

Daartoe voeren wij een orthogonaal assenstelsel in met op de assen de variabelen  $Z_j$  uitgezet. Met de  $k$ -de persoon correspondeert dan een punt in deze door  $n$  assen voortgebrachte ruimte en wel met de coördinaten

$$Z_{jk} \quad (j = 1, \dots, n).$$

De testresultaten kunnen dan worden aangegeven met een puntenwolk van  $N$  punten. Door de standaardisering voor iedere waarde van  $j$  ligt het zwaartepunt van deze puntenwolk in de oorsprong. Wij voeren nu een nieuw assenstelsel in, waarvan de assen, met het zwaartepunt van de puntenwolk als nieuwe oorsprong, als volgt worden vastgelegd. De eerste as wordt zo gekozen, dat zij ligt in de richting van de grootste spreiding. Vervolgens projecteren wij de puntenwolk op een hypervlak loodrecht op de eerste as. De in dit hypervlak overblijvende spreiding is dan zo klein mogelijk. Vervolgens kiezen wij de tweede as in de richting van de grootste spreiding van deze geprojecteerde puntenwolk. Wanneer wij steeds na het vaststellen van een richting van het nieuwe assenstelsel de puntenwolk projecteren op een hypervlak, dat loodrecht staat op de reeds verkregen assen, dan vinden wij op deze wijze de richtingen van de  $n$  assen van ons nieuwe coördinatenstelsel. Op de assen van het nieuwe coördinatenstelsel worden nu de variabelen  $F_i$  uitgezet. Dit betekent, dat de eerste  $m$  gemeenschappelijke factoren gezamenlijk de grootste bijdrage leveren, die een  $m$ -tal gemeenschappelijke factoren geven kan. Dit is nu de eigenschap, toegekend aan de factoren  $F_1, F_2, \dots$  in volgorde, waardoor ondubbelzinnige schattingen verkregen worden.

Ook heeft Hotelling een toets ontwikkeld om na te gaan of de bijdragen van de laatste factoren te onderscheiden zijn van meetfouten. Hiervoor is het noodzakelijk, dat de waarnemingen in duplo of in multiplo worden verricht.

De Principal Component Analysis van Hotelling beantwoordt dus in sterke mate aan onze statistische doelstellingen.

P. Whittle [3] heeft, voortbouwende op denkbeelden van G. Young

een andere techniek voorgesteld, welke het best gezien kan worden als een reactie op die technieken, waarin verondersteld wordt, dat de factorvariabelen normaal verdeeld zijn (een onderstelling, die voor de methode van Hotelling ook niet noodzakelijk is). Whittle is van mening dat in de praktijk aan deze veronderstelling niet altijd voldaan is en bovendien is zij vaak overbodig.

Hij ziet dan ook  $F_{ik}$  niet als een waarde, welke de variabele  $F_i$  aanneemt, maar als een parameter van de persoon, terwijl  $a_{ji}$  een parameter is van de test. De beide parameters komen in het model gelijkwaardig voor en waarom zouden zij dan ook niet gelijkwaardig worden behandeld?

Ter onderscheiding van de andere technieken zullen wij het model van Whittle als volgt aangeven:

$$(10) \quad Z_{jk} = \sum_{i=1}^m a_{ji} b_{ik} + e_{jk}.$$

Zijn techniek komt dan neer op de minimalisering van de volgende vorm:

$$(11) \quad U = \sum_{j=1}^N \sum_{k=1}^N \frac{(Z_{jk} - \sum_{i=1}^m a_{ji} b_{ik})^2}{S^2 e_j}.$$

Met betrekking tot de parameterwaarden  $a_{ji}$  en  $b_{ik}$ , waarbij  $S^2 e_j$  een schatting is van de variantie van de meetfoutvariabele  $e_j$ . Hiertoe moeten de waarnemingen b.v. in duplo worden verricht, hetgeen echter niet altijd mogelijk is (zo kan men b.v. een psychologische test niet „onafhankelijk” herhalen met dezelfde proefpersoon).

Hierdoor verkrijgt men een aantal betrekkingen waaraan de op een voorgeschreven wijze genormaliseerde constanten  $a_{ji}$  en  $b_{ik}$  moeten voldoen.

Verder zijn Whittle en Young in staat de hypothese omtrent het minimum aantal gemeenschappelijke factoren te toetsen. Ook kunnen zij betrouwbaarheidsintervallen geven voor de parameterwaarden.

De techniek van Whittle en Young past men dus direct toe op het waarnemingsmateriaal, terwijl door Hotelling eerst de variabelen worden gestandaardiseerd. Daardoor worden volgens Whittle de meetfouten niet gelijkwaardig behandeld. De door Whittle uitgevoerde standaardisering van de parameters heeft dit bezwaar niet.

Ook D. N. Lawley [4] heeft een techniek ontwikkeld, welke men zou kunnen rangschikken onder de Componentenanalyse. Aangezien deze methode veel overeenkomst vertoont met een andere techniek van Lawley, welke wij bij de zuivere Factoranalyse zullen bespreken, wordt slechts met de vermelding volstaan.

### 3. Zuivere Factoranalyse

De eerste methode, welke wij zullen bespreken is afkomstig van D. N. Lawley [5]. Zoals boven is vermeld, ziet het model in de zuivere Factoranalyse er als volgt uit:

$$(5) \quad Z_{jk} = \sum_{i=1}^m a_{ji} F_{ik} + a_j S_{jk}.$$

Lawley gaat er bij zijn methode vanuit, dat de variabelen  $F_i$  en  $S_j$  alle normaal en onafhankelijk verdeeld zijn op de populatie der proefpersonen. Verder neemt hij aan, dat deze verdelingen alle  $\sigma = 1$  hebben (hetgeen mogelijk is, daar de spreidingen in de coëfficiënten  $a_{ji}$  en  $a_j$  opgenomen kunnen worden. De verdeling van de steekproefcovarianties van de  $Z_j$ 's worden dan gegeven door de Wishart-verdeling. De kansdichtheid van deze verdeling kan zo geschreven worden, dat hierin de constanten  $a_{ji}$  en  $a_j$  als onbekende parameters voorkomen. Deze constanten worden nu zo gekozen, dat de kansdichtheid voor de gevonden waarden van de steekproefcovarianties maximaal wordt (methode der grootste aannemelijkheid). Wederom vinden wij dan een aantal betrekkingen, waaraan de constanten  $a_{ji}$  en  $a_j$  dienen te voldoen. Deze vergelijkingen geven echter geen unieke oplossing. Lawley vervangt deze betrekkingen dan ook door een ander stelsel vergelijkingen, waarvan niet duidelijk is, welke extra voorwaarden het vertegenwoordigt, doch dat een ondubbelzinnige oplossing bezit, die ook voldoet aan het eerste stelsel vergelijkingen. Later zullen wij bij de bespreking van Rao's canonical factor analysis nog een tweede oplossing ontmoeten van het eerste stelsel vergelijkingen. Vervolgens heeft Lawley in de loop der jaren verschillende bijzondere gevallen van zijn methoden bekeken.

Lawley berekent geen factorwaarden, maar behandelt wel een toets voor de hypothese omtrent het minimum aantal gemeenschappelijke factoren. Bovendien geeft hij betrouwbaarheidsintervallen voor de constanten  $a_{ji}$  en  $a_j$ .

Tenslotte zullen wij enige aandacht besteden aan een methode, welke door C. R. Rao [6] is voorgesteld en door hem *Canonical Factor Analysis* wordt genoemd.

Om aan te geven welke extra voorwaarden Rao aan de variabelen  $F_i$  oplegt, denken wij eerst de variabelen  $Z_j$  geschreven als lineaire combinaties van de variabelen  $X_j$  en  $S_j$  en wel als volgt:

$$(12) \quad Z_j = X_j + a_j S_j \quad (j = 1, \dots, n).$$

Vervolgens voeren wij twee variabelen  $U$  en  $V$  in, waarvoor geldt:



$$(13) \quad U = \sum_{j=1}^n l_j Z_j$$

$$(14) \quad V = \sum_{j=1}^n q_j X_j$$

en waarbij zowel  $l_j$  als  $q_j$  constanten zijn.

Het principe van de methode is nu, dat de correlatie tussen  $U$  en  $V$  zo groot mogelijk gemaakt wordt. Deze correlatiecoëfficiënt is een functie van de  $l_j$  en  $q_j$  en bezit  $m$  extreme waarden. Bij ieder van deze extreme waarden behoort een stelsel  $(l_j, q_j)$ , dus ook een variabele  $V$ . Deze  $m$  variabelen neemt Rao nu als de gemeenschappelijke factoren  $F_j$ . De berekeningen verlopen volgens een ingewikkelde iteratiemethode, met behulp waarvan de coëfficiënten  $a_{ji}$  en  $a_j$  geschat worden. Vervolgens kunnen dan ook schattingen van de waarden der factoren verkregen worden. Men kan bewijzen, dat de constanten  $a_{ji}$  en  $a_j$  ook moeten voldoen aan de eerste betrekkingen, die Lawley heeft gevonden, zodat de oplossing van Rao ook meest aannemelijke schattingen geeft, hetgeen deze techniek bijzonder aantrekkelijk maakt.

Ook geeft Rao een toets voor de hypothese omtrent het minimum aantal gemeenschappelijke factoren.

### Literatuur

- [1] H. Hotelling, Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology 1933, blz. 417 e.v.
- [2] Zie ook voor korte samenvatting: M. G. Kendall, Factor Analysis, Journal of the Royal Statistical Society B, 12 (1950), blz. 60 e.v.
- [3] P. Whittle, On principal components and least square methods of Factor Analysis, Skandinavisk Aktuarie tidskrift, 35 (1953), blz. 223-229.
- [4] D. N. Lawley, Further investigations in Factor Analysis, Proceedings of the Royal Society of Edinburgh, A 61, blz. 176 e.v.
- [5] D. N. Lawley, the estimation of Factor Loadings by the method of maximum likelihood, Proceedings Royal Society of Edinburgh A 61 (1940), blz. 64-83.
- [6] C. R. Rao, Estimation and tests of significance in Factor Analysis, Psychometrika, vol. 20, no. 2, blz. 93-111.