Centre
for
Social
Theory

PC vs. PAF

Een enigszins technische inleiding

Ronan Van Rossem
Universiteit Gent

e-doc

CST e-DOCUMENT NR. 2, JANUARI 2011

UNIVERSITEIT GENT
Department of Sociology
Korte Meer 3
9000 Gent
Belgium

Phone: +32(0)9 264.67.96
Fax: +32(0)9 264.69.75
Email: socio@ugent.be

PC vs. PAF

Een enigszins technische inleiding

Ronan Van Rossem

Universiteit Gent

CST e-DOCUMENT NR. 2, JANUARI 2011

Inhoud

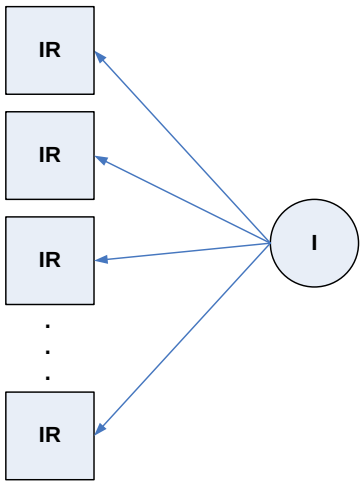
Figuren

Tabellen

Principes van hoofdcomponenten- en factoranalyse

Inleiding

Het wordt vaak aangeraden wanneer men onderzoek doet meerdere indicatoren voor eenzelfde onderliggende concept te gebruiken. Dit verhoogt de kwaliteit van de verzamelde informatie. De aanwezigheid van deze meerdere variabelen voor hetzelfde onderliggende concept in één enkele analyse kan echter wel voor problemen zorgen. Als men meerdere variabelen heeft die allemaal ongeveer hetzelfde meten zullen deze ook allemaal relatief sterk met elkaar gecorreleerd zijn. Situaties waarin je multipele indicatoren voor eenzelfde onderliggende dimensie heeft komen bv. vaak voor wanneer attitudes en opvattingen gemeten worden, bv. meningen over politieke issues, of bij het meten van aspiraties en gedrag (zoals criminaliteit, cultuurparticipatie, enz.). Dit betekent ook vaak dat indien men al deze variabelen samen in een analyse zou invoeren men multicollineariteitsproblemen kan ervaren en dat de interpretatie van de resultaten enorm vermoeilijkt. In deze gevallen zou men erbij gebaat zijn mocht men het aantal variabelen in de analyse kunnen terugbrengen tot één enkele of een paar variabelen die de onderliggende dimensies weergeven. Het aanmaken van schalen of indexen is één mogelijkheid om het aantal variabelen te reduceren. Men kan echter ook op puur statistische grond gaan proberen de onderliggende dimensies te identificeren, en het is hier dat hoofdcomponentenanalyse en factoranalyse hun intrede doen. Zoals weergegeven in Figuur 0-1 tracht factoranalyse een reeks geobserveerde manifeste variabelen te vervangen door een kleiner aantal latente variabelen. De manifeste variabelen dienen op het interval of rationiveau gemeten zijn, de latente factor is een interval variabele.



Figuur 0-1: Schematische voorstelling van factoranalyse

Hoofdcomponentenanalyse of “principal component analysis” (PCA) en factoranalyse (FA) zijn (verwante) technieken voor datareductie die in dergelijke situaties kunnen gebruikt worden. Zij laten ons toe een reeks geobserveerde of manifeste variabelen $\mathbf{X} = \{X_1, X_2, \ldots, X_K\}$ te vervangen door een reeks factoren—ook wel latente variabelen genoemd—die een lineaire combinatie zijn van de geobserveerde variabelen. Elk van deze factoren Y kan geschreven worden als:

$$Y_{il} = b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + \ldots + b_K X_{iK}$$

Het komt er op aan om *L* factoren dusdanig te kiezen dat het aantal factoren *L* niet alleen substantieel kleiner is dat het oorspronkelijke *K* aantal variabelen maar ook zo'n groot mogelijk deel van de gemeenschappelijke variantie in de oorspronkelijke variabelen verklaren. Deze technieken zoeken dus naar de patronen in de samenhang, in de correlatie, tussen de geobserveerde variabelen. In hoofdcomponenten- en factoranalyse wordt geen onderscheid gemaakt tussen afhankelijke en onafhankelijke variabelen. Het meetniveau van de geobserveerde variabelen is interval of ratio.

Hoofdcomponenten- en factoranalyse kunnen ook van pas komen wanneer men geconfronteerd wordt met een grote hoeveelheid data en men wil onderzoeken of er onderliggende structuren te vinden zijn in deze data. PC en factoranalyse kunnen helpen deze onderliggende structuren of dimensies bloot te leggen, en orde te scheppen in de datachaos.

Zoals vele van de correlationele statistische technieken vindt ook factoranalyse haar oorsprong in de psychologische studie van intelligentie. Charles Spearman observeerde dat de scores van leerlingen over verschillende vakken positief gecorreleerd waren. Hieruit deduceerde hij dat intelligentie uit 2 'factoren' bestond, een algemene intelligentiefactor *g* en een taakspecifieke factor *s* {Spearman 1904 # 299090}. De discussie rond deze algemene intelligentiefactor *g* is nog steeds aan de gang, maar het was voor Spearman wel de aanleiding om factoranalyse te ontwikkelen.

Voorbeeld 1: Redenen voor sociale participatie

Als voorbeeld wordt gebruik gemaakt van een reeks variabelen die peilen naar de redenen voor sociale participatie (zie Tabel 0-1) in een grote internationale steekproef, de *World Values Surveys* van 1990 (World Values Study Group, 1994). Deze vragen werden natuurlijk alleen maar gesteld van respondenten die aangaven dat ze vrijwilligerswerk in organisaties vervulden. Dit beperkt de steekproefgrootte *N* tot 6566. De vraag die men zich hierbij kan stellen is of de 14 concrete redenen die men kon aanhalen kunnen herleid worden tot enkele onderliggende redenen.

Tabel 0-1: Vragen en variabelen over redenen voor sociale participatie

Thinking about your reasons for doing voluntary work, please use the following five-point scale to indicate how important each of the reasons below have been in your own case. (WHERE 1 IS UNIMPORTANT AND 5 IS VERY IMPORTANT)

V 55 A) A sense of solidarity with the poor and disadvantaged
V 56 B) Compassion for those in need
V 57 C) An opportunity to repay something, give something back
V 58 D) A sense of duty, moral obligation
V 59 E) Identifying with people who were suffering
V 60 F) Time on my hands, wanted something worthwhile to do
V 61 G) Purely for personal satisfaction
V 62 H) Religious beliefs
V 63 I) To help give disadvantaged people hope and dignity
V 64 J) To make a contribution to my local community
V 65 K) To bring about social or political change
V 66 L) For social reasons, to meet people
V 67 M) To gain new skills and useful experience
V 68 N) I did not want to, but could not refuse

In totaal waren er *K* = 14 (V55 – V68) vragen over de redenen voor sociale participatie. Elke vraag werd gescoord op een vijfpuntsschaal waarbij een waarde van 1 betekende dat deze reden voor

vrijwilligerswerk onbelangrijk was voor de respondent, en een waarde van 5 dat ze een zeer belangrijke reden was om vrijwilligerswerk te verrichten[1].

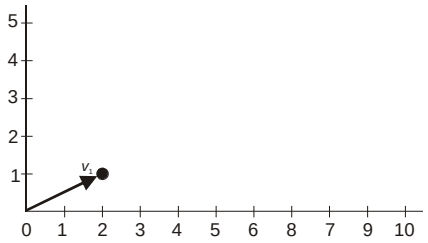
Tabel 0-2: Correlatietabel voor voorbeeld 1

	V55	V56	V57	V58	V59	V60	V61	V62	V63	V64	V65	V66	V67	V68
V55	1.000	0.675	0.308	0.338	0.585	0.187	0.054	0.315	0.629	0.185	0.299	0.089	0.096	0.075
V56	0.675	1.000	0.417	0.390	0.629	0.179	0.097	0.296	0.600	0.208	0.229	0.116	0.138	0.140
V57	0.308	0.417	1.000	0.399	0.395	0.212	0.130	0.165	0.313	0.198	0.179	0.171	0.182	0.179
V58	0.338	0.390	0.399	1.000	0.453	0.195	0.116	0.188	0.350	0.241	0.254	0.134	0.156	0.173
V59	0.585	0.629	0.395	0.453	1.000	0.215	0.112	0.313	0.585	0.180	0.288	0.122	0.159	0.161
V60	0.187	0.179	0.212	0.195	0.215	1.000	0.267	0.127	0.209	0.192	0.117	0.320	0.308	0.098
V61	0.054	0.097	0.130	0.116	0.112	0.267	1.000	0.066	0.077	0.050	0.090	0.249	0.231	0.133
V62	0.315	0.296	0.165	0.188	0.313	0.127	0.066	1.000	0.380	0.207	0.062	0.025	-.018	0.101
V63	0.629	0.600	0.313	0.350	0.585	0.209	0.077	0.380	1.000	0.260	0.312	0.115	0.163	0.114
V64	0.185	0.208	0.198	0.241	0.180	0.192	0.050	0.207	0.260	1.000	0.273	0.221	0.168	0.075
V65	0.299	0.229	0.179	0.254	0.288	0.117	0.090	0.062	0.312	0.273	1.000	0.190	0.235	0.126
V66	0.089	0.116	0.171	0.134	0.122	0.320	0.249	0.025	0.115	0.221	0.190	1.000	0.522	0.129
V67	0.096	0.138	0.182	0.156	0.159	0.308	0.231	-.018	0.163	0.168	0.235	0.522	1.000	0.162
V68	0.075	0.140	0.179	0.173	0.161	0.098	0.133	0.101	0.114	0.075	0.126	0.129	0.162	1.000

Tabel 0-2 toont de correlatiematrix voor de 14 geobserveerde variabelen. Veel structuur kan men daar direct niet in vinden. Alleen tussen variabelen V55 (solidariteit met benadeelden), V56 (medelijden), V59 (identificatie met hen die lijden) en V63 (benadeelden hoop en waarde geven) bestaan er sterke correlatie ($r > 0.500$) terwijl de andere correlaties meestal onder de 0.300 blijven.

Eigenwaarden en eigenvectoren

Vooraleer we verder kunnen ingaan op PC en factoranalyse moet er eerst een inleiding gegeven worden over eigenwaarden en eigenvectoren. Deze vormen de wiskundige basis voor PC en factoranalyse.



Figuur 0-1: Grafische voorstelling van een vector

Een vector kan men grafisch voorstellen als een pijl, waarvan de staart gevormd wordt door de oorsprong van het assenstelsel, en de kop door het punt met als coördinaten de elementen van de vector. Neem bv., de vector

$$\mathbf{v_1} = \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

[1] Strikt genomen gaat het hier om ordinale variabelen waarvoor deze technieken niet mogen toegepast worden. Men doet echter vaak of het hier om variabelen op intervalniveau gaat omdat men er van uit gaat dat ze allen indicatoren zijn van één of meerdere onderliggende continue schalen.

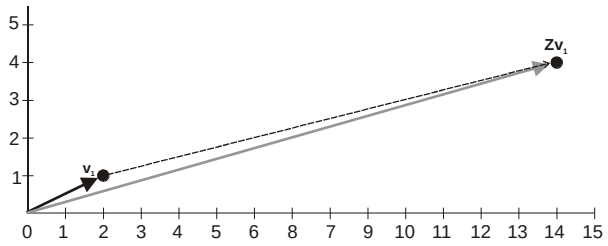
deze kan grafisch voorgesteld worden zoals getoond in Figuur 0-1: als een pijl die de oorsprong $\begin{pmatrix}0\\0\end{pmatrix}$ verbindt met het punt $\begin{pmatrix}2\\1\end{pmatrix}$. Als men deze vector nu vermenigvuldigt men een vierkante matrix **Z**, dan beeldt deze de oorspronkelijke vector $\mathbf{v_1}$ af op een andere vector in dezelfde ruimte. Neem nu dat:

$$\mathbf{Z} = \begin{pmatrix}3 & 8\\1 & 2\end{pmatrix},$$

dan beeldt deze *v_1 af op de vector:

$$\mathbf{Zv_1} = \begin{pmatrix}3 & 8\\1 & 2\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix} = \begin{pmatrix}14\\4\end{pmatrix}.$$

De grafische voorstelling hiervan wordt weergegeven in Figuur 0-2 waar de stippellijn de transformatiefunctie weergeeft.



Figuur 0-2: Grafische weergave van de afbeelding van vector $\mathbf{v_1}$ op $\mathbf{Zv_1}$

De transformatiematrix **Z** kan dus elk punt uit deze ruimte afbeelden op een ander punt in deze ruimte. Neem nu de vectoren $\mathbf{v_2} = \begin{pmatrix}4\\7\end{pmatrix}$ en $\mathbf{v_3} = \begin{pmatrix}1.75\\-4.55\end{pmatrix}$, deze worden door **Z** afgebeeld op, respectievelijk:

$$\mathbf{Zv_2} = \begin{pmatrix}3 & 8\\1 & 2\end{pmatrix}\begin{pmatrix}4\\7\end{pmatrix} = \begin{pmatrix}68\\18\end{pmatrix}$$

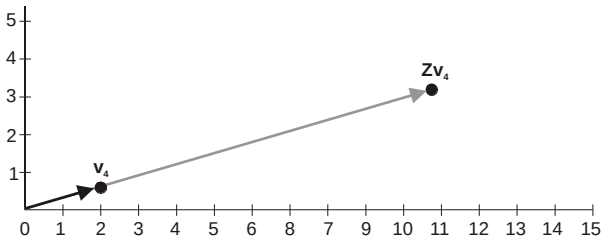
en

$$\mathbf{Zv_3} = \begin{pmatrix}3 & 8\\1 & 2\end{pmatrix}\begin{pmatrix}1.75\\-4.55\end{pmatrix} = \begin{pmatrix}-31.15\\-7.35\end{pmatrix}.$$

Maar als men nu de vector $\mathbf{v_4} = \begin{pmatrix}2\\0.593\end{pmatrix}$ vermenigvuldigt met de matrix **Z**, dan wordt deze afgebeeld op:

$$\mathbf{Zv_4} = \begin{pmatrix}3 & 8\\1 & 2\end{pmatrix}\begin{pmatrix}2\\0.593\end{pmatrix} = \begin{pmatrix}10.745\\3.186\end{pmatrix}.$$

Op het eerste zicht lijkt dit niets bijzonders, maar als men dit grafisch weergeeft dan stelt men toch iets bijzonders vast.



Figuur 0-3: Grafische weergave van de afbeelding van $\mathbf{v_4}$ op $\mathbf{Zv_4}$

Namelijk de afbeelding $\mathbf{Zv_4}$ ligt in het verlengde van de oorspronkelijke vector $\mathbf{v_4}$. Dit wil zeggen dat men de transformatie (of matrixproduct) $\mathbf{Zv_4}$ ook kan schrijven als het product van een scalaire waarde λ met de oorspronkelijke vector $\mathbf{v_4}$:

$$\mathbf{Zv_4} = \lambda\mathbf{v_4}$$

In dit geval is $\lambda = 5.372$, en inderdaad:

$$\lambda\mathbf{v_4} = 5.372\begin{pmatrix} 2 \\ 0.593 \end{pmatrix} = \begin{pmatrix} 10.745 \\ 3.186 \end{pmatrix}.$$

Wanneer een vector door een transformatiematrix afgebeeld wordt op een veelvoud van zichzelf, dan is deze vector een eigenvector van de transformatiematrix, en is het veelvoud (λ) de corresponderende eigenwaarde. In dit voorbeeld is $\mathbf{v_4}$ dus een eigenvector van $\mathbf{Z}$, en is de corresponderende eigenwaarde $\lambda = 5.372$. Let wel, met elke eigenwaarde stemmen een oneindig aantal eigenvectoren overeen, in dit geval zullen alle veelvouden van $\mathbf{v_4}$ ook eigenvectoren met eigenwaarde 5.372 zijn. Bv. in het geval van $2 \times \mathbf{v_4}$:

$$\mathbf{Z}\ 2\mathbf{v}_4 = \begin{pmatrix} 3 & 8 \\ 1 & 2 \end{pmatrix} 2\begin{pmatrix} 2 \\ 0.593 \end{pmatrix} = \begin{pmatrix} 3 & 8 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 4 \\ 1.186 \end{pmatrix} = \begin{pmatrix} 21.489 \\ 6.372 \end{pmatrix} = 5.372\begin{pmatrix} 4 \\ 1.186 \end{pmatrix} = \lambda\ 2v_4 \quad .$$

Eigenwaarden en eigenvectoren berekenen

Men heeft een vierkante $K \times K$ matrix $\mathbf{A}$ met rang K – d.w.z. $|\mathbf{A}| \neq 0$ – en waarvan de elementen aangeduid worden als a_{rc}, waar r de rij van het element is en c de kolom:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{KK} \end{pmatrix}$$

De eigenwaarden of karakteristieke wortels λ van de matrix $\mathbf{A}$ zijn de oplossingen van de volgende vergelijking:

$$\left|\mathbf{A}-\lambda\mathbf{I}\right| = \begin{vmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K1} & a_{K2} & \cdots & a_{KK}-\lambda \end{vmatrix} = 0 \qquad (01)$$

het zijn dus de oplossingen voor de vergelijking gevormd door de determinant van de matrix **A** waarbij van de diagonale cellen een onbekende waarde λ afgetrokken wordt gelijk is aan nul. De waarden van λ waarvoor de determinant van$|\mathbf{A} - \lambda\mathbf{I}|$ nul is worden de eigenwaarden van de matrix **A** genoemd. **I** is hierbij steeds de $K\times K$ identiteitsmatrix, een matrix met de waarde 1 op de hoofddiagonaal, en een waarde 0 in alle niet-diagonale cellen. De matrix **A** heeft evenveel eigenwaarden l als het aantal rijen of kolommen. Indien **A** een $K\times K$ matrix is, zal het K eigenwaarden hebben λ_1 tot en met λ_K. Het kan natuurlijk zo zijn dat een bepaalde eigenwaarde meerdere malen voorkomt.

Een $K\times K$ matrix **A** heeft dan ook K eigenvectoren **V**, één voor elke eigenwaarde. Deze eigenvectoren zijn de vectoren waarvoor geldt dat:

$$\mathbf{AV} = \lambda\mathbf{V} \qquad (02)$$

Wanneer een eigenvector **V** door de transformatiematrix **A** gaat wordt die afgebeeld op λ maal zichzelf. Het product **AV** projecteert de vector **V** in het verlengde van zichzelf. Als **V*** = b**V** waarbij b een scalaire constante is en **V** een eigenvector is van de matrix **A**, dan zal **V*** ook een eigenvector zijn van matrix **A** geassocieerd met de zelfde eigenwaarde λ. Immers:

$$\mathbf{AV^*} = \mathbf{A}(b\mathbf{V}) = b(\mathbf{AV}) = b(\lambda\mathbf{V}) = \lambda(b\mathbf{V}) = \lambda\mathbf{V^*}.$$

Elke eigenwaarde heeft dan ook een oneindige reeks geassocieerde eigenvectoren. De dimensies van deze eigenvectoren is $K\times 1$. Omdat er een oneindig aantal eigenvectoren geassocieerd zijn met elke eigenwaarde berekent men meestal de genormaliseerde eigenvector, dit is de eigenvector met lengte 1. De lengte van een vector is de vierkantswortel van de som van de kwadraten van de elementen van de vector.

Voorbeeld

Als voorbeeld veronderstellen we dat **A** een 3×3 matrix is, met als elementen:

$$\mathbf{A} = \begin{pmatrix} 6 & 2 & 1 \\ 4 & 5 & 9 \\ 2 & 4 & 9 \end{pmatrix}$$

De rang van deze matrix is ook 3, daar deze matrix niet-singulier is met een determinant $|\mathbf{A}| = 24$. Dit betekent ook dat deze matrix 3 eigenwaarden zal hebben verschillende van 0, daar de determinant van een matrix gelijk is aan het product van de eigenwaarden. In dit geval betekent dit dat het product van de 3 eigenwaarden van **A** gelijk moet zijn aan de determinant van **A** of 24.

De identiteitsmatrix **I** is hier ook een 3×3 matrix met 1-en op de diagonaal en 0-en in de niet-diagonale cellen:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Om de eigenwaarden van **A** te berekenen moeten we eerst de matrix **A** – λ**I** aanmaken:

$$\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} 6-\lambda & 2 & 1 \\ 4 & 5-\lambda & 9 \\ 2 & 4 & 9-\lambda \end{pmatrix},$$

welke gelijk is aan de oorspronkelijke matrix **A** waarvan bij de diagonale cellen een onbekende waarde λ afgetrokken wordt. De eigenwaarden van **A** zijn nu deze waarden van λ waarvoor de determinant |**A** – λ**I**| = 0. In het voorbeeld zijn dit de oplossing van de volgende vergelijking:

$$|\mathbf{A} - \lambda\mathbf{I}| = 24 - 83\lambda + 20\lambda^2 - \lambda^3 = 0$$

De waarden voor λ waarvoor deze vergelijking gelijk is aan 0 zijn de eigenwaarden van **A**.

De oplossingen voor deze vergelijking zijn[2]:

$$\begin{bmatrix} \left(\frac{854}{27}+\frac{1}{9}\,i\cdot\sqrt{301515}\right)^{\left(\frac{1}{3}\right)} + \frac{151}{\left[9\left(\frac{854}{27}+\frac{1}{9}\,i\cdot\sqrt{301515}\right)^{\left(\frac{1}{3}\right)}\right]} + \frac{20}{3} \\ \frac{-1}{2}\left(\frac{854}{27}+\frac{1}{9}\,i\cdot\sqrt{301515}\right)^{\left(\frac{1}{3}\right)} - \frac{151}{\left[18\left(\frac{854}{27}+\frac{1}{9}\,i\cdot\sqrt{301515}\right)^{\left(\frac{1}{3}\right)}\right]} + \frac{20}{3} + \frac{1}{2}\,i\cdot\sqrt{3}\left[\left(\frac{854}{27}+\frac{1}{9}\,i\cdot\sqrt{301515}\right)^{\left(\frac{1}{3}\right)} - \frac{151}{\left[9\left(\frac{854}{27}+\frac{1}{9}\,i\cdot\sqrt{301515}\right)^{\left(\frac{1}{3}\right)}\right]}\right] \\ \frac{-1}{2}\left(\frac{854}{27}+\frac{1}{9}\,i\cdot\sqrt{301515}\right)^{\left(\frac{1}{3}\right)} - \frac{151}{\left[18\left(\frac{854}{27}+\frac{1}{9}\,i\cdot\sqrt{301515}\right)^{\left(\frac{1}{3}\right)}\right]} + \frac{20}{3} - \frac{1}{2}\,i\cdot\sqrt{3}\left[\left(\frac{854}{27}+\frac{1}{9}\,i\cdot\sqrt{301515}\right)^{\left(\frac{1}{3}\right)} - \frac{151}{\left[9\left(\frac{854}{27}+\frac{1}{9}\,i\cdot\sqrt{301515}\right)^{\left(\frac{1}{3}\right)}\right]}\right] \end{bmatrix}$$

of, iets eenvoudiger:

$$\lambda = \begin{pmatrix} 14.322 \\ 0.312 \\ 5.366 \end{pmatrix},$$

waar de eigenwaarden λ_1, λ_2 en λ_3 de elementen van de vector λ zijn. Dat het product van de eigenwaarden gelijk is aan de determinant van **A** blijkt ook hier[3]:

$$\prod_{k=1}^{3} \lambda_k = \lambda_1\lambda_2\lambda_3 = 14.322 \times 0.312 \times 5.366 = 24 = |\mathbf{A}|$$

De eigenvectoren voor de eerste eigenwaarde λ_1 = 14.322 kunnen we berekenen door volgende vergelijking op te lossen:

$$\mathbf{A}\cdot\mathbf{V_1} = 14.322 \times \mathbf{V_1}$$

[2] Met dank (sic) aan Mathcad 7.0 (MathSoft, 1997) voor deze oplossing.
[3] Op enkele afrondingsfouten na.

waarbij $\mathbf{V_1}$ de eigenvector is geassocieerd met de eerste eigenwaarde. Dit levert volgend stelsel van vergelijkingen op:

$$\begin{pmatrix} 6 & 2 & 1 \\ 4 & 5 & 9 \\ 2 & 4 & 9 \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = 14.322 \times \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

of

$$\begin{cases} 6v_1 + 2v_2 + 1v_3 = 14.322v_1 \\ 4v_1 + 5v_2 + 9v_3 = 14.322v_2 \\ 2v_1 + 4v_2 + 9v_3 = 14.322v_3 \end{cases}$$

Dit stelsel van vergelijkingen heeft een oneindig aantal oplossingen. Daarom berekent men gewoonlijk de genormaliseerde eigenvector, d.w.z. de eigenvector met lengte 1. De lengte van een vector **b**

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{pmatrix}$$

is gelijk aan de vierkantswortel van de som van de kwadraten van de coëfficiënten:

$$|\mathbf{b}| = \sqrt{\sum_{i=1}^{K} b_i^2}$$

of in matrixvorm:

$$(|\mathbf{b}|)^2 = \mathbf{b'b}$$

Met deze bijkomende beperking vindt men voor $\mathbf{V_1}$:

$$\mathbf{V_1} = \begin{pmatrix} 0.251 \\ 0.726 \\ 0.640 \end{pmatrix}$$

Dat de lengte van deze vector gelijk is aan 1, blijkt uit:

$$\mathbf{v_1'v_1} = \sum_{i=1}^{3} v_{1i}^{\ 2} = 1$$

Men kan zien dat deze vector $\mathbf{V_1}$ een eigenvector is daar:

$$\mathbf{A} \cdot \mathbf{V_1} = \begin{pmatrix} 6 & 2 & 1 \\ 4 & 5 & 9 \\ 2 & 4 & 9 \end{pmatrix} \cdot \begin{pmatrix} 0.251 \\ 0.726 \\ 0.640 \end{pmatrix} = \begin{pmatrix} 3.600 \\ 10.397 \\ 9.168 \end{pmatrix} = 14.322 \times \begin{pmatrix} 0.251 \\ 0.726 \\ 0.640 \end{pmatrix} = \lambda_1 \mathbf{V_1}$$

Voor de tweede eigenwaarde λ_2 = 0.312, de corresponderende genormaliseerde eigenvector $\mathbf{V_2}$ is:

$$\mathbf{V_2} = \begin{pmatrix} 0.254 \\ -0.899 \\ 0.356 \end{pmatrix}$$

en voor de derde eigenwaarde λ_3 =5.366"

$$\mathbf{V_3} = \begin{pmatrix} -0.909 \\ 0.084 \\ 0.408 \end{pmatrix}$$

Als men deze vectoren post-vermenigvuldigt met matrix **A** krijgt men respectievelijk:

$$\mathbf{A} \cdot \mathbf{V_2} = \begin{pmatrix} 0.079 \\ -0.281 \\ 0.111 \end{pmatrix} \text{ en } \mathbf{A} \cdot \mathbf{V_3} = \begin{pmatrix} -4.879 \\ 0.453 \\ 2.187 \end{pmatrix}$$

wat respectievelijk gelijk is aan λ_2·**V2** en λ_3·**V3**.

Als men bv. de vector **B**:

$$\mathbf{B} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

dan is de uitkomt van de post-vermenigvuldiging met **A**:

$$\mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} 9 \\ 18 \\ 15 \end{pmatrix}$$

Hieruit kan men besluiten dan **B** geen eigenvector is, **B** wordt door **A** niet geprojecteerd op een veelvoud van zichzelf. De onderstaande vector **C** daarentegen is wel een eigenvector of **A**, daar **AC** = λ_3**C**:

$$\mathbf{C} = \begin{pmatrix} -3.637 \\ 0.338 \\ 1.630 \end{pmatrix}$$

$$\mathbf{A} \cdot \mathbf{C} = \begin{pmatrix} 6 & 2 & 1 \\ 4 & 5 & 9 \\ 2 & 4 & 9 \end{pmatrix} \cdot \begin{pmatrix} -3.637 \\ 0.338 \\ 1.630 \end{pmatrix} = \begin{pmatrix} -19.517 \\ 1.811 \\ 8.748 \end{pmatrix} = 5.366 \times \begin{pmatrix} -3.637 \\ 0.338 \\ 1.630 \end{pmatrix} = \lambda_3 \mathbf{C}$$

C is dus wel een eigenvector geassocieerd met eigenwaarde λ_3 = 5.366, maar het is geen genormaliseerde eigenwaarde daar de lengte van **C:**

$$|\mathbf{C}| = -19.517^2 + 1.811^2 + 8.748^2 = 460.72$$

is niet gelijk aan 1.

Belangrijk is dat bij symmetrische matrices de eigenwaarden steeds reële getallen zijn ($\lambda_i \in \mathbb{R}$), terwijl bij niet-symmetrische matrices de eigenwaarden ook complexe getallen kunnen zijn. Neem bv. matrices **K** en **L**, beide zijn vierkante 4×4 matrices, maar waar **K** symmetrisch is, is **L** niet-symmetrisch.

$$\mathbf{K} = \begin{pmatrix} 2 & 6 & 1 & 0 \\ 6 & 4 & 8 & 2 \\ 1 & 8 & 3 & 4 \\ 0 & 2 & 4 & 9 \end{pmatrix} \text{ en } \mathbf{L} = \begin{pmatrix} 2 & 8 & 9 & 7 \\ 6 & 4 & 0 & 5 \\ 1 & 8 & 3 & 1 \\ 0 & 2 & 4 & 9 \end{pmatrix}$$

De eigenwaarden voor deze matrices zijn respectievelijk:

$$\lambda_{\mathbf{K}} = \begin{pmatrix} -6.256 \\ 0.642 \\ 7.706 \\ 15.908 \end{pmatrix} \text{ en } \lambda_{\mathbf{L}} = \begin{pmatrix} 16.098 \\ -1.582 + 4.348i \\ -1.528 - 4.348i \\ 5.067 \end{pmatrix}.$$

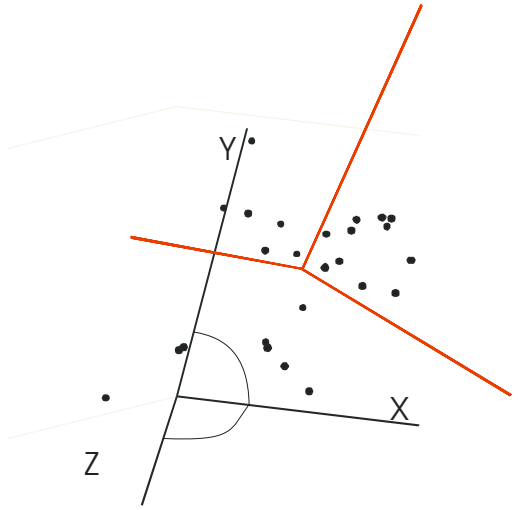
Waar de eigenwaarden voor **K** alle reële getallen zijn, zijn de tweede en derde eigenwaarde voor **L** complexe getallen, d.w.z., getallen die zowel een reëel als een imaginair deel hebben. Het imaginaire deel wordt aangeduid door de wiskundige grootheid, de imaginaire eenheid *i*, die als uniek kenmerk heeft dat $i^2 = -1$.

Hoofdcomponenten- vs. factoranalyse

Als men een verzameling van *K* variabelen heeft: $\mathbf{X} = \{X_1, X_2, \ldots, X_K\}$ dan kan men elk van die variabelen beschouwen als één van de *K* dimensies of assen van een graph. In het algemeen zijn deze verschillende dimensies niet onafhankelijk van elkaar of niet-orthogonaal ten opzichte van elkaar als de variabelen gecorreleerd zijn. Grafisch kan men zich de correlatie tussen twee variabelen voorstellen als de hoek tussen de assen voor deze variabelen. De correlatie is niets anders dan de cosinus van de hoek tussen de twee assen. Bv. twee variabelen die een correlatie van 0.30 hebben zullen corresponderen met twee assen met een hoek van 72.5°, een correlatie van 0.50 correspondeert met een hoek van 60°, en een correlatie van bv. -0.40 met een hoek van 113.6°. Een correlatie van 1 correspondeert met een hoek van 0° tussen de twee assen, d.w.z. met twee assen die volledig samenvallen, terwijl een correlatie van 0 correspondeert met een hoek van 90°, dus met twee assen die loodrecht op elkaar staan en orthogonaal zijn.

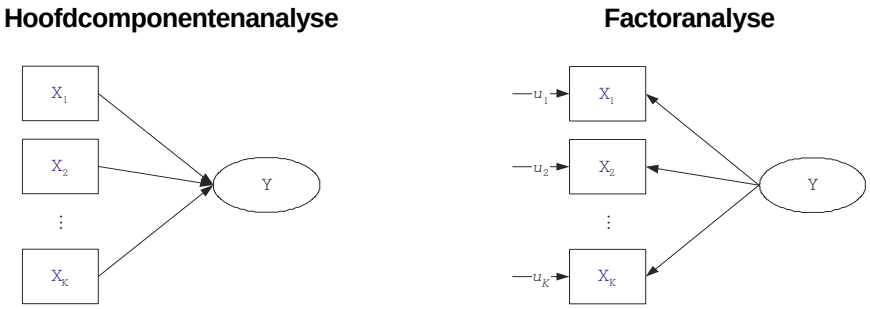
De waarden van een observatie *i* op deze verzameling variabelen $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$ kan men dan beschouwen als de coördinaten van een punt in deze *K*-dimensionele ruimte. Wat hoofdcomponentenanalyse en factoranalyse nu trachten te doen is een nieuw en efficiënter othogonaal assenstelsel met *L* dimensies te vinden waarbij $L \leq K$ is, en waarmee men ook alle datapunten kan lokaliseren. In Figuur 0-1 wordt dit voorgesteld. Het oorspronkelijke assenstelsel X, Y en Z is gecorreleerd. Bv. de hoek tussen de X- en de Y-as, $\alpha < 90°$ wat betekent dat de variabelen *X* en *Y* gecorreleerd zijn. Bij hoofdcomponenten- en factoranalyse vervangt men dit oorspronkelijke assenstelsel door een nieuw (rood in Figuur 0-1) waarvan de verschillende assen wel orthogonaal zijn en waarvan de eerste dimensies het grootste deel van de spreiding der

punten beschrijven. Om dit efficiënter assenstelsel te vinden gebruikt men eigenwaarden en eigenvectoren.



Figuur 0-1: Grafische voorstelling van doelstellingen hoofdcomponenten- en factoranalyse

Hoofdcomponentenanalyse en factoranalyse zijn twee gerelateerde technieken voor variabelenreductie. Technisch gezien zijn deze twee technieken vrij gelijklopend en het verschil tussen hen is dan ook in de eerste plaats filosofisch. Men gebruikt dan ook vaak de term factoranalyse om naar beide methoden te verwijzen. Het onderscheid tussen de twee technieken wordt geïllustreerd in Figuur 0-2.



Figuur 0-2: Hoofdcomponentenanalyse vs. factoranalyse

Hoofdcomponentenanalyse (PCA) zoekt heel eenvoudig naar de lineaire combinaties van de waargenomen variabelen die zo'n groot mogelijk deel van de totale variantie van die variabelen verklaren. Men probeert dus een latente (niet gemeten) variabele *Y* te schatten als:

$$Y = b_1 X_1 + \ldots + b_k X_k + \ldots + b_K X_K \,, \qquad (03)$$

waarbij de b_k-s wegingscoëfficiënten zijn en de X_k-s scores op de waargenomen variabelen. Bij PCA is er geen onderliggend theoretisch model van de relaties tussen de variabelen, maar probeert men gewoon het aantal variabelen te reduceren door de *K* waargenomen **X** variabelen ter vervangen door een kleiner aantal $P \leq K$ latente variabelen **Y** die zoveel mogelijk van de oorspronkelijke variantie in **X** verklaren. Zoals getoond in Figuur 0-2 is in deze benadering de latente variabele *Y* niets anders dan een combinatie van de waargenomen variabelen $X_1, \ldots, X_K$.

Bij factoranalyse vertrekt men van een radicaal ander gezichtspunt. In tegenstelling tot PCA ziet men hier de latente variabele niet als een combinatie van de waargenomen variabelen, maar de waargenomen variabelen **X** als expressies van onderliggende latente dimensies **Y**. Het is daarom ook dat in Figuur 0-2 de pijlen bij factoranalyse in de tegenovergesteld richting lopen, van de latente variabele Y naar de waargenomen variabelen X_1, …, X_K, dan bij PCA. Deze latente dimensies zijn theoretische concepten waarvan de waargenomen variabelen indicatoren zouden zijn. In het geval van een enkele onderliggende dimensie, zoals in Figuur 0-2, kan men elk van de waargenomen variabelen dan ook schrijven als een functie van deze onderliggende dimensie en een unieke residuele term:

$$\begin{cases} X_1 = b_1 Y + u_1 \\ \vdots \\ X_k = b_k Y + u_k \\ \vdots \\ X_K = b_K Y + u_K \end{cases} \qquad (04)$$

waarbij b_1, …, b_K de wegingscoëfficienten zijn om de waargenomen variabelen X_1, …, X_K te schatten op basis van de latente variabele Y, en de u_1, …, u_K de residuele termen voor variabelen X_1, …, X_K, dit is het deel van de waargenomen variabelen dat niet veroorzaakt wordt door de latente variabele. Het belangrijkste praktische verschil tussen PCA en factoranalyse is dat bij deze laatste er men vanuit gaat dat de variantie van de waargenomen variabelen kan opgesplitst worden in twee componenten: een gemeenschappelijk deel van de variantie dat verklaard wordt door de latente variabelen en dat men in de literatuur deel de communaliteit van de variabele noemt, en een deel unieke variantie die niet kan toegeschreven worden aan de latente variabelen.

De extractie van de hoofdcomponenten en factoren

Inleiding

In dit hoofdstuk worden de twee meest gebruikte technieken uitgelegd, namelijk de hoofdcomponentenanalyse of “principal components” (PC) en hoofdassen-factoranalyse of “principal axis factoring” (PAF). De hoofdcomponenten is de meest basis van deze technieken, en hoofdassen-factoranalyse is een factoranalysetechniek die gebaseerd is op de hoofdcomponentenanalyse. Vooraleer men echter dergelijke analyses kan aanvatten dient men eerst na te gaan of de gebruikte set van variabelen wel geschikt zijn voor hoofdcomponenten- of factoranalyse. Eerst worden dan ook enkele veel gebruikte toetsen voor de geschiktheid van variabelen voor deze analyses voorgesteld, en daarna wordt in detail beschreven hoe hoofdcomponenten en factoren (middels hoofdassen-factoranalyse) geëxtraheerd worden.

Toetsen of de variabelen wel geschikt zijn voor hoofdcomponenten- of factoranalyse

Al te vaak wordt een hoofdcomponenten- of factoranalyse uitgevoerd zonder zich af te vragen of de data daar wel geschikt voor zijn. Niet alle datastructuren lenen zich voor factoranalyse. Om factoranalyse mogelijk te maken dienen de correlaties tussen de variabelen sterk genoeg zijn om onderliggende dimensies te onderscheiden. Aangezien het doel van deze technieken is het aantal variabelen te reduceren, heeft het weinig zin om een factoranalyse uit te voeren indien deze tot evenveel factoren zou leiden als het oorspronkelijk aantal variabelen. Twee vaak gebruikte toetsen om na te gaan of een correlatiematrix geschikt is voor factoranalyse zijn de Bartlett toets voor sfeericiteit en de Kaiser-Meyer-Olkin toets voor toereikendheid van de steekproef.

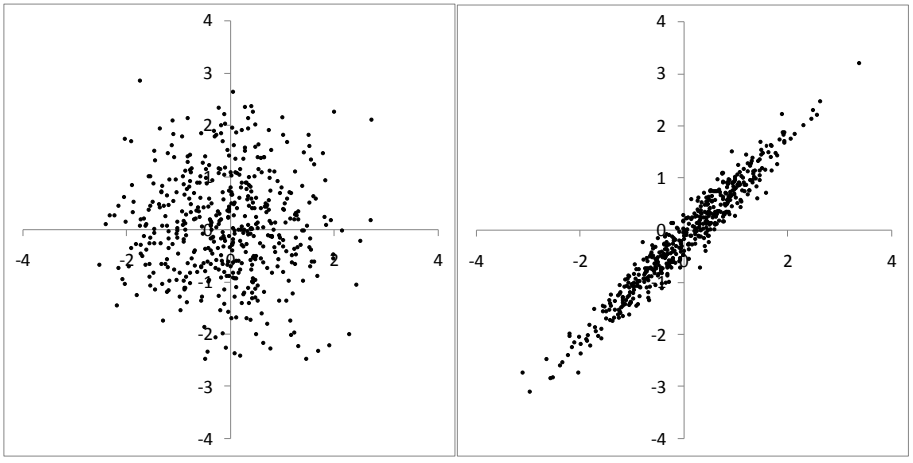
De Bartlett toets voor sfeericiteit

Hoofdcomponenten- en factoranalyse zijn technieken om naar onderliggende dimensies binnen data te zoeken, en dit op basis van de covariantie of correlatie tussen de variabelen. Indien de gebruikte variabelen allen onafhankelijke van elkaar zijn, en dus hun onderlinge correlaties allemaal gelijk zijn aan 0, heeft het weinig zin om hier een factoranalyse op uit te voeren, daar de factoren die men zou bekomen gelijk zouden zijn aan de oorspronkelijke variabelen. Bartletts “test of sphericity”, toets voor sfeericiteit (of bolvormigheid) toetst of het wel de moeite is om een hoofdcomponenten- of factoranalyse uit te voeren op een gegeven set van variabelen binnen een bepaalde steekproef (Bartlett, 1950; Dziuban & Shirkey, 1974). Deze test gaat na of deze steekproef getrokken zou kunnen zijn uit een populatie waarin de correlaties tussen de variabelen allen gelijk zijn aan 0.

Indien in de populatie alle variabelen onafhankelijk van elkaar zijn, dan moet de populatie correlatiematrix ρ een identiteitsmatrix zijn:

$$\rho = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} = \mathbf{I},$$

d.w.z., een matrix met allemaal enen op de hoofddiagonaal en 0 elders. Bartletts toets voor sfeericiteit toetst of de waargenomen correlatiematrix **R** kan getrokken zijn uit een populatie waarin $\rho = \mathbf{I}$. Men noemt deze toets voor de onafhankelijkheid van variabelen een toets van sfeericiteit of bolvormigheid omdat wanneer een reeks variabelen onderling onafhankelijk zijn, en dus een orthogonaal assenstelsel vormen de data-puntenwolk een sfeer met een homogene dichtheid. Dit wordt geïllustreerd in Figuur 0-1, waar aan de ene kant een spreidingsdiagram getoond wordt voor twee onafhankelijke of orthogonale variabelen, en aan de andere kant een spreidingsdiagram voor twee sterk gecorreleerde of niet-orthogonale variabelen. Deze figuur toont dat het spreidingsdiagram voor de twee orthogonale variabelen een duidelijk meer cirkelvorming patroon opleveren dan dit voor de niet-orthogonale variabelen.



Orthogonaal Niet-orthogonaal

Figuur 0-1: Grafische voorstelling van puntenwolken van orthogonale en niet-orthogonale datareeksen:

De hypothesen voor deze toets zijn dus:

$$H_0: \rho = \mathbf{I}$$

$$H_1: \rho \neq \mathbf{I}$$

De nulhypothese stelt dus dat de populatiecorrelatiematrix een identiteitsmatrix is, terwijl de alternatieve hypothese stelt dat dit niet het geval is.

De toetsstatistiek volgt een χ^2-verdeling met df = (1/2) $K(K - 1)$ vrijheidsgraden, waar K het aantal variabelen in de analyse is. De toetsstatistiek is:

$$\chi^2 = -\left(N - 1 - \frac{2K+5}{6}\right)\ln\ |\mathbf{R}|$$

(05)

waar *N* de steekproefomvang is en **R** de geobserveerde correlatiematrix, en |**R**| dus de determinant van deze geobserveerde correlatiematrix (Dziuban & Shirkey, 1974; SPSS, 2006). Aangezien de determinant van **R** steeds kleiner is dan 1, is ln(|**R**|) per definitie steeds negatief, wat maakt dat de toetsstatistiek steeds positief is. Dit betekent dus dat hoe groter de steekproef is, hoe groter de kans is dat een correlatiematrix goed genoeg is voor factoranalyse.

Wanneer deze toetsstatistiek kleiner of gelijk is aan de kritieke waarde χ^2_α aanvaarden we de nulhypothese dat deze correlatiematrix getrokken werd uit een populatie waarin deze variabelen onafhankelijk van elkaar zijn, en dat het dus niet de moeite waard is om een hoofdcomponenten- of factoranalyse uit te voeren. Is de toetsstatistiek groter dan de kritieke waarde dan verwerpen we de nulhypothese en aanvaarden we de alternatieve hypothese, en kan er inderdaad een analyse uitgevoerd worden.

Een equivalente formule (Jobson, 1992) voor deze toetsstatistiek is:

$$\chi^2 = -\left(N - \frac{2K+11}{6}\right)\sum_{i=1}^{K}\ln\ \lambda_i$$

waarbij λ_i, i = 1,…, K de *K* eigenwaarden van de correlatiematrix **R** zijn, en dit omdat $\prod_{i=1}^{K}\lambda_i = |\mathbf{R}|$, het product van de eigenwaarden is gelijk aan de determinant van de correlatiematrix.

De Bartlett toets voor sfeericiteit voor Voorbeeld 1

Bartletts toets voor sfeericiteit kan dus gebruikt worden om na te gaan of de gegevens in voorbeeld 1 kunnen gebruikt worden in een factoranalyse. De correlatiematrix **R** wordt gegeven in Tabel 0-2, hierboven. De determinant van deze matrix is |**R**| = 0.02315. Het aantal variabelen is *K* = 14, en het aantal observaties *N* =6566.

Hypothese

De nulhypothese is dus dat deze 14 variabelen onafhankelijke van elkaar zijn en dat de waargenomen correlaties allen aan toeval te wijten zijn, en dat de populatiecorrelatiematrix dus een identiteitsmatrix is:

$$H_0: \rho = \mathbf{I}$$

$$H_1: \rho \neq \mathbf{I}$$

De alternatieve hypothese stelt dat dit niet het geval is, en dat ten minste één van de waargenomen correlaties niet aan toeval te wijten is.

Steekproevenverdeling

De steekproevenverdeling is een χ^2-verdeling met (1/2) *K*(*K* – 1) vrijheidsgraden. Voor dit voorbeeld betekent dit:

$$df = \frac{1}{2}K\ K-1\ = \frac{1}{2}14\ 14-1\ = 91.$$

Kritieke waarde

De kritieke waarde bij een χ^2-verdeling hangt af van 1) het aantal vrijheidsgraden van de verdeling, en 2) het α-niveau of de Type I fout dat men bereid is te aanvaarden. In het geval van α = 5% en df = 91, dan is de kritieke waarde χ^2_α = 114.268.

Toetsstatistiek

De toetsstatistiek voor dit voorbeeld kan berekend worden met de formule:

$$\begin{aligned}\chi^2 &= -\left(N - 1 - \frac{2K+5}{6}\right)\ln\ |\mathbf{R}| \\ &= -\left(6566 - 1 - \frac{2\times 14+5}{6}\right)\ln\ 0.02315 \\ &= -\ 6559.5\ \times\ -3.766\ = 24701.51\end{aligned}$$

Besluit

Tot een besluit komt men door de berekende toetsstatistiek te vergelijken met de kritieke waarde. Is de toetsstatistiek kleiner dan of gelijk aan de kritieke waarde dan aanvaardt men de nulhypothese dat de geobserveerde correlatiematrix getrokken is uit een populatie waar de variabelen onafhankelijk van elkaar zijn en de ware correlatiematrix dus een identiteitsmatrix. Is de toetsstatistiek groter dan de kritieke waarde dan verwerpen we de nulhypothese en aanvaarden we de alternatieve hypothese en beschouwen we de variabelen als niet-onafhankelijk van elkaar.

In dit voorbeeld was de kritieke waarde χ^2_α = 114.268 en de toetsstatistiek χ^2 = 24701.51. In dit geval is de toetsstatistiek duidelijk groter dan de kritieke waarde ($\chi^2 > \chi^2_\alpha$) en verwerpen we dus de nulhypothese en aanvaarden we de alternatieve hypothese. Men zou tot dezelfde beslissing kunnen komen door het significantieniveau van de toetsstatistiek te vergelijken met de α-waarde. In dit voorbeeld was α = 0.05 en het significantieniveau p < .001. Indien het significantieniveau kleiner is dan de α-waarde verwerpen we de nulhypothese en aanvaarden we de alternatieve hypothese, terwijl wanneer het significantieniveau gelijk is aan of groter is dan de α-waarde we de nulhypothese aanvaarden.

Dat we voor dit voorbeeld de nulhypothese verwerpen betekent dat deze data geschikt zijn voor hoofdcomponenten- of factoranalyse.

Kaiser-Meyer-Olkin toets voor toereikendheid van de steekproef

De Kaiser-Meyer-Olkin toets voor toereikendheid van de steekproef of kortweg KMO-toets gaat na of de variabelen in een set psychometrisch samenhoren (Kaiser, 1970; Dziuban & Shirkey, 1974). Indien de variabelen in een set allen dezelfde onderliggende dimensies vatten, dan kan men verwachten dat de correlatie tussen twee variabelen grotendeels kan verklaard worden door hun relaties met de andere variabelen in de set, of met andere woorden, de partiële correlatie tussen twee variabelen, controlerend voor alle andere variabelen in de set, dienen vrij klein te zijn, d.w.z., dicht tegen 0 aan.

Neem dat men K variabelen heeft, X_1, …, X_K en definieer

$$r_{ij.\bullet} = r_{ij.1,2,\ldots,K,(-i),(-j)}$$

als de partiële correlatie tussen variabelen X_i en X_j, controlerend voor alle andere variabelen in de set. De (*-i*) en (*-j*) tonen aan dat er niet voor de variabelen zelf gecontroleerd wordt. In dit geval wordt de KMO-toetsstatistiek gedefinieerd als:

$$KMO = \frac{\sum_{i=1}^{K}\sum_{\substack{j=1 \\ i \neq j}}^{K} r_{ij}^2}{\sum_{i=1}^{K}\sum_{\substack{j=1 \\ i \neq j}}^{K} r_{ij}^2 + \sum_{i=1}^{K}\sum_{\substack{j=1 \\ i \neq j}}^{K} r_{ij.}^2}$$

De teller hier is niets anders dan de som van de kwadraten van alle correlaties tussen de variabelen in de set, de noemer bevat deze zelfde som, plus de som van de kwadraten van alle partiële correlatiecoëfficiënten controlerend voor de andere variabelen in de set. In het beste geval verklaren de andere variabelen in de set de relatie tussen elke twee variabelen X_i en X_j perfect en is $r_{ij.} = 0$. In dit geval zijn teller en noemer gelijk en is de KMO gelijk aan 1. Indien alle partiële correlaties gelijk zijn aan de zero-orde correlaties dan is de noemer twee maal de teller en zal KMO gelijk zijn aan 0.5. Computerprogramma's berekenen deze partiële correlaties aan de hand van de "anti-image" correlatiematrix, waarvan de elementen gelijk zijn aan minus de partiële correlaties.

De KMO kan ook voor elke variabele apart berekend worden. De formule voor de variabele specifieke KMO voor een variabele X_i is:

$$KMO_i = \frac{\sum_{\substack{j=1 \\ i \neq j}}^{K} r_{ij}^2}{\sum_{\substack{j=1 \\ i \neq j}}^{K} r_{ij}^2 + \sum_{\substack{j=1 \\ i \neq j}}^{K} r_{ij.}^2}$$

Het enige verschil met de bovenstaande statistiek is dat hier de correlaties en partiële correlaties niet over alle paren van variabelen gesommeerd worden, maar alleen deze waarin X_i deelneemt. Het gebruik van deze variabele-specifieke KMO laat toe variabelen te identificeren die weinig gemeen hebben met de andere in de set, en dus weinig zullen bijdragen tot de factoranalyse.

Men streeft dus naar een KMO die zo dicht mogelijk bij 1 ligt. Er zijn geen strikte regels voor welke waarde een KMO minimaal moet hebben, maar Tabel 0-1 geeft toch enige richtlijnen. Scores boven de 0.8 geven aan dat deze set variabelen zeer geschikt is voor factoranalyse, terwijl een waarde onder de 0.5 aangeeft dat men hier beter geen factoranalyse op uitvoert.

Tabel 0-1: Richtlijnen voor de interpretatie van de KMO

Waarde KMO	*Evaluatie*
> 0.90	Schitterend
0.80 – 0.90	Goed
0.70 – 0.80	Middelmatig
0.60 – 0.70	Pover
0.50 – 0.60	Slecht
< 0.50	Onaanvaardbaar

(bron: Dziuban & Shirkey, 1974, p. 359)

De KMO voor Voorbeeld 1

Om de KMO te berekenen voor Voorbeeld 1 dient men eerst de partiële correlaties voor alle variabelenparen, controlerend voor alle andere variabelen te berekenen. Aangezien hier steeds voor 12 variabelen dienen gecontroleerd te worden is dit een omslachtig werkje. Computerprogramma's gebruiken hiervoor de "anti-image" matrix, een transformatie van de inverse van de correlatiematrix[4], waarvan de elementen de gewenste partiële correlatiecoëfficiënten zijn. Tabel 0-2 geeft zowel de zero-orde als partiële correlatiecoëfficiënten voor het voorbeeld weer.

Tabel 0-2: Zero-orde en partiële correlatiecoëfficiënten voor Voorbeeld 1

	V55	V56	V57	V58	V59	V60	V61	V62	V63	V64	V65	V66	V67	V68
V55		*0.391*	*-0.019*	*0.011*	*0.154*	*0.055*	*-0.039*	*0.067*	*0.277*	*-0.037*	*0.133*	*0.004*	*-0.054*	*-0.063*
V56	**0.675**		*0.177*	*0.054*	*0.256*	*-0.038*	*0.026*	*0.004*	*0.180*	*0.038*	*-0.076*	*0.006*	*0.014*	*0.036*
V57	**0.308**	**0.417**		*0.210*	*0.098*	*0.064*	*0.027*	*0.000*	*-0.002*	*0.052*	*-0.003*	*0.038*	*0.035*	*0.077*
V58	**0.338**	**0.390**	**0.399**		*0.199*	*0.041*	*0.025*	*0.007*	*0.019*	*0.106*	*0.083*	*-0.010*	*0.009*	*0.067*
V59	**0.585**	**0.629**	**0.395**	**0.453**		*0.043*	*0.015*	*0.082*	*0.197*	*-0.068*	*0.078*	*-0.012*	*0.020*	*0.041*
V60	**0.187**	**0.179**	**0.212**	**0.195**	**0.215**		*0.177*	*0.038*	*0.039*	*0.081*	*-0.053*	*0.149*	*0.137*	*-0.006*
V61	**0.054**	**0.097**	**0.130**	**0.116**	**0.112**	**0.267**		*0.043*	*-0.020*	*-0.056*	*0.023*	*0.117*	*0.078*	*0.070*
V62	**0.315**	**0.296**	**0.165**	**0.188**	**0.313**	**0.127**	**0.066**		*0.194*	*0.139*	*-0.102*	*-0.009*	*-0.102*	*0.063*
V63	**0.629**	**0.600**	**0.313**	**0.350**	**0.585**	**0.209**	**0.077**	**0.380**		*0.087*	*0.110*	*-0.033*	*0.067*	*-0.008*
V64	**0.185**	**0.208**	**0.198**	**0.241**	**0.180**	**0.192**	**0.050**	**0.207**	**0.260**		*0.184*	*0.124*	*0.009*	*-0.016*
V65	**0.299**	**0.229**	**0.179**	**0.254**	**0.288**	**0.117**	**0.090**	**0.062**	**0.312**	**0.273**		*0.044*	*0.115*	*0.057*
V66	**0.089**	**0.116**	**0.171**	**0.134**	**0.122**	**0.320**	**0.249**	**0.025**	**0.115**	**0.221**	**0.190**		*0.424*	*0.024*
V67	**0.096**	**0.138**	**0.182**	**0.156**	**0.159**	**0.308**	**0.231**	**-0.018**	**0.163**	**0.168**	**0.235**	**0.522**		*0.075*
V68	**0.075**	**0.140**	**0.179**	**0.173**	**0.161**	**0.098**	**0.133**	**0.101**	**0.114**	**0.075**	**0.126**	**0.129**	**0.162**	

vet: zero-orde correlatie; *cursief*: partiële correlatie

Of men de KMO nu berekent op de volledige correlatiematrix of op de halve, maakt daar correlatiematrices symmetrisch zijn geen verschil uit. De formule voor de KMO is:

$$KMO = \frac{\sum_{i=1}^{K}\sum_{\substack{j=1 \\ i\neq j}}^{K} r_{ij}^2}{\sum_{i=1}^{K}\sum_{\substack{j=1 \\ i\neq j}}^{K} r_{ij}^2 + \sum_{i=1}^{K}\sum_{\substack{j=1 \\ i\neq j}}^{K} r_{ij.}^2}$$

K is in dit voorbeeld gelijk aan 14, en de sommaties van alle zero-orde en partiële correlaties zijn respectievelijk:

$$\sum_{i=1}^{14}\sum_{\substack{j=1 \\ i\neq j}}^{14} r_{ij}^2 = 13.089$$

$$\sum_{i=1}^{14}\sum_{\substack{j=1 \\ i\neq j}}^{14} r_{ij.}^2 = 2.209$$

wat een KMO geeft van:

[4] De anti-image matrix $\mathbf{A} = \mathbf{S} \cdot \mathbf{R}^{-1} \cdot \mathbf{S}$, waarbij $\mathbf{S}^2 = (\text{diag}(\mathbf{R}^{-1}))^{-1}$.

$$KMO = \frac{13.089}{13.089 + 2.209} = 0.856$$

Dit resultaat toont dat deze data inderdaad goed geschikt zijn voor een hoofdcomponenten- of factoranalyse.

Men kan ook de KMO voor elk van de variabelen apart berekenen wat dan aangeeft van hoe goed deze variabele is voor een factoranalyse. De resultaten hiervoor worden getoond in Tabel 0-3.

Tabel 0-3: Variabelenspecifieke KMOs voor Voorbeeld 1

Variabele	$\sum_{\substack{j=1 \\ i \neq j}}^{K} r_{ij}^2$	$\sum_{\substack{j=1 \\ i \neq j}}^{K} r_{ij.\bullet}^2$	KMO_i
V55	1.686	0.289	0.854
V56	1.814	0.296	0.860
V57	0.937	0.101	0.902
V58	1.035	0.112	0.902
V59	1.779	0.200	0.899
V60	0.585	0.097	0.858
V61	0.280	0.066	0.810
V62	0.570	0.096	0.855
V63	1.730	0.213	0.890
V64	0.514	0.108	0.827
V65	0.626	0.114	0.846
V66	0.636	0.237	0.729
V67	0.668	0.243	0.733
V68	0.228	0.036	0.863

Deze resultaten tonen dat geen van de variabelen ongeschikt is om opgenomen te worden in de analyse. Sommige van de variabelen zijn zeer geschikt om opgenomen te worden in een analyse met KMO_i-s van meer dan 0.9, terwijl andere maar middelmatig scoren met waarden in 0.70 range.

Hoofdcomponentenanalyse

In deze sectie gaan we dieper in op de extractie van de latente variabelen, de hoofdcomponenten (of principale componenten). Eerst moet men kiezen of men de analyse baseert op een covariantie- of op een correlatiematrix. Beide zijn mogelijk, maar meestal kiest men er voor de correlatiematrix te gebruiken. In deze sectie gaan we in op de eenvoudigste van de twee technieken, hoofdcomponentenanalyse, om in een volgende sectie de methode uit te breiden tot één vorm van factoranalyse, namelijk 'principal axis factoring'. Er zijn nog andere vormen van factoranalyse die niet aan bod komen in dit hoofdstuk, maar hoewel die andere methoden gebruiken om de factoren te extraheren gebeurt de interpretatie van de resultaten op een gelijkaardige manier. Enkele van deze methoden worden kort overlopen in het volgende hoofdstuk.

Covariantiematrix of correlatiematrix

Hoofdcomponenten- en factoranalyse kunnen vertrekken van ofwel de covariantiematrix of van de correlatiematrix. Indien men kiest voor de covariantiematrix betekent dit dat men werkt met de gecentreerde waarden van de variabelen: $\dot{X} = X - \bar{X}$, waarbij $\bar{\dot{X}} = 0$ en $s_{\dot{X}} = s_X$. De

covariantiematrix is dan niets anders dan $\mathbf{S} = \frac{1}{N}\dot{\mathbf{X}}'\dot{\mathbf{X}}$, waarbij *N* het aantal observaties in de steekproef is, en $\dot{\mathbf{X}}$ de *N*×*K* matrix met de gecentreerde scores voor de *N* observaties op de *K* variabelen.

Het alternatief is gebruik te maken van de correlaties tussen de variabelen. In dit geval gebruikt men de gestandaardiseerde waarden (de z-scores) van de oorspronkelijke variabelen:

$$x = \frac{X - \bar{X}}{s_X}$$

waarbij $\bar{x} = 0$ en s_x = 1. De correlatiematrix voor de variabelen kan dan berekend worden als $\mathbf{R} = \frac{1}{N}\mathbf{x}'\mathbf{x}$, waar **x** de *N*×*K* matrix is met de gestandaardiseerde scores voor de *N* observaties op de *K* variabelen.

Gewoonlijk maakt men gebruik van de correlatiematrix **R** omdat de resultaten die men zo verkrijgt gemakkelijker te interpreteren zijn dan diegene die men verkrijgt door de covariantiematrix **S** te gebruiken. In de rest van dit hoofdstuk zullen we dan ook gebruik maken van de correlaties tussen de variabelen. De ontwikkelde argumenten kunnen eenvoudig uitgebreid worden naar de covarianties tussen de variabelen.

Hoofdcomponentenanalyse

De bedoeling van hoofdcomponentenanalyse is een reeks latente variabelen, de hoofdcomponenten of PC-en, hier aangeduid als **Y**, die maximaal de variantie in een set variabelen **X** verklaren, en die geformuleerd kunnen worden als een lineaire combinatie van de gestandaardiseerde **X** variabelen (Dunteman, 1989; Mardia, Kent, & Bibby, 1979). Indien men *K* waargenomen variabelen heeft in **X** en *P* latente variabelen in **Y**, waarbij $P \leq K$, en *N* observaties, dan kan men de score van de *i*-de observatie op de *p*-de latente variabele kunnen schrijven als:

$$Y_{ip} = \sum_{k=1}^{K} a_{kp} x_{ik} \qquad (06)$$

Hierbij kiest men de PC-en dusdanig dat de eerste PC het grootste deel van de variantie in **X** verklaart, de tweede PC het tweede grootste deel, en zo verder tot de laatste PC die het kleinste deel van de variantie in **X** verklaart:

$$\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \ldots \geq \text{Var}(Y_K) \qquad (07)$$

De schatting van de gewichten a_{kp} is wel onderhevig aan de beperking dat de som van de kwadraten van de gewichten voor een gegeven PC gelijk moeten zijn aan 1:

$$\sum_{k=1}^{K} a_{kp}^2 = 1 \qquad (08)$$

Een verdere beperking is dat de geselecteerde PC-en orthogonaal ten opzichte van elkaar moeten zijn, of m.a.w. dat de covariantie tussen de verschillend PC-en nul moet zijn:

$$\forall i, j, i \neq j : \text{Cov}\ Y_i, Y_j = 0 \qquad (09)$$

In matrixformaat wordt dit:

$$\mathbf{Y} = \mathbf{xA}$$

waar **Y** is een $N \times P$ matrix met de waarden op de P latente variabelen voor de N observaties, **x** een $N \times K$ matrix met de gestandaardiseerde waarden voor de K waargenomen variabelen, en **A** een $K \times P$ gewichtsmatrix. Deze laatste is onderhevig aan de volgende beperking:

$$\mathbf{A'A} = \mathbf{I}$$

wat een samenvatting is van de beperkingen in Vgl. 08 en 09.

Om deze hoofdcomponenten te bepalen gaat men de waargenomen correlatiematrix **R**, een een positief-definite—een matrix waarvan $|\mathbf{R}| > 0$—symmetrische $K \times K$ matrix:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1K} \\ r_{12} & 1 & \dots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1K} & r_{2K} & \dots & 1 \end{pmatrix}$$

decompositeren in een niet-symmetrische $K \times K$ matrix **L**:

$$\mathbf{L} = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1K} \\ l_{21} & l_{22} & \dots & l_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ l_{K1} & l_{K2} & \dots & l_{KK} \end{pmatrix}$$

waarvoor geldt dat:

$$\mathbf{R} = \mathbf{LL'} \qquad (010)$$

Laten we beginnen met een klein voorbeeld: 3 variabelen X_1, X_2, en X_3 die volgende correlatiematrix **R** hebben:

$$\mathbf{R} = \begin{pmatrix} 1 & 0.668 & 0.309 \\ 0.668 & 1 & 0.420 \\ 0.309 & 0.420 & 1 \end{pmatrix}$$

Om de hoofdcomponenten te bepalen berekent men eerst de K eigenwaarden $\lambda_1, \dots, \lambda_K$ voor deze correlatiematrix:

$$\lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_p \\ \vdots \\ \lambda_P \end{pmatrix}$$

waarbij de eigenwaarden gerangschikt worden van groot naar klein:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq \dots \geq \lambda_P > 0$$

In het geval van het voorbeeld leidt dit tot de volgen vector met eigenwaarden:

$$\lambda = \begin{pmatrix} 1.949 \\ 0.732 \\ 0.319 \end{pmatrix}$$

De eerste principale component is nu niets anders dan de genormaliseerde eigenvector $\mathbf{V_1}$ geassocieerd met de grootste eigenwaarde van **R**, λ_1:

$$\mathbf{V}_1 = \begin{pmatrix} v_{11} \\ \vdots \\ v_{k1} \\ \vdots \\ v_{K1} \end{pmatrix}$$

waar v_{11}, …, v_{K1} de elementen van de eigenvector zijn. Daar dit een genormaliseerde eigenvector is geldt dat (zie Vgl. 08):

$$(\mathbf{V_1})'\mathbf{V_1} = 1$$

In het voorbeeld wordt dit:

$$\mathbf{V_1} = \begin{pmatrix} 0.604 \\ 0.637 \\ 0.479 \end{pmatrix}$$

of de genormaliseerde eigenvector geassocieerd met de eerste eigenwaarde λ_1 = 1.949. Dat dit inderdaad de genormaliseerde eigenvector geassocieerd met deze eerste eigenwaarde is wordt bevestigd door:

$$\mathbf{R} \cdot \mathbf{V_1} = \begin{pmatrix} 1 & 0.668 & 0.309 \\ 0.668 & 1 & 0.420 \\ 0.309 & 0.420 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.604 \\ 0.637 \\ 0.479 \end{pmatrix} = \begin{pmatrix} 1.178 \\ 1.242 \\ 0.933 \end{pmatrix} = 1.949 \times \begin{pmatrix} 0.604 \\ 0.637 \\ 0.479 \end{pmatrix} = \lambda_1 \mathbf{V_1}$$

en

$$\mathbf{V_1}\ '\mathbf{V_1} = 0.604^2 + 0.637^2 + 0.479^2 = 1.000$$

De tweede principale component wordt op gelijkaardig wijze verkregen als de genormalizeerde eigenvector $\mathbf{V_2}$ geassocieerd met de 2^e grootste eigenwaarde van de correlatiematrix **R**. In het voorbeeld wordt dit:

$$\mathbf{V_2} = \begin{pmatrix} -0.456 \\ -0.216 \\ 0.863 \end{pmatrix}$$

Dat dit inderdaad een eigenvector is kan nagegaan worden door:

$$\mathbf{R}\cdot\mathbf{V}_2=\begin{pmatrix}1 & 0.668 & 0.309\\ 0.668 & 1 & 0.420\\ 0.309 & 0.420 & 1\end{pmatrix}\cdot\begin{pmatrix}-0.456\\ -0.216\\ 0.863\end{pmatrix}=\begin{pmatrix}-0.334\\ -0.158\\ 0.632\end{pmatrix}=0.732\times\begin{pmatrix}-0.456\\ -0.216\\ 0.863\end{pmatrix}=\lambda_2\mathbf{V}_2$$

Deze tweede PC moet orthogonaal zijn met de eerste. Dit is het geval wanneer de som der kruisproducten der elementen van $\mathbf{V_1}$ en $\mathbf{V_2}$ gelijk is aan 0:

$$\mathbf{V_1}'\mathbf{V_2}=\mathbf{V_2}'\mathbf{V_1}=0.604\times-0.456+0.637\times-0.216+0.479\times0.863=0$$

Het is in feite niet nodig om te testen voor de orthogonaliteit van de eigenvectoren daar bij symmetrische matrices (zoals **R**) de eigenvectoren altijd orthogonaal zijn.

Het totaal aantal hoofdcomponenten dat kan berekend worden is gelijk aan het aantal eigenwaarden, wat op zijn beurt gelijk is aan de rang van de correlatiematrix en deze is gelijk aan het aantal variabelen. Indien men dus *K* variabelen in de hoofdcomponentenanalyse betrekt zal me ook *K* PC-en kunnen berekenen. Dat betekent dus dat ik het voorbeeld, waar *K* = 3, er ook 3 PC-en kunnen geëxtraheerd worden. De derde principale component is de genormaliseerde eigenvector $\mathbf{V_3}$ geassocieerd met de derde en laagste eigenwaarde λ_3:

$$\mathbf{V}_3=\begin{pmatrix}-0.653\\ 0.740\\ -0.160\end{pmatrix}$$

waarvoor ook geldt dat:

$$\mathbf{R}\cdot\mathbf{V}_3=\begin{pmatrix}1 & 0.668 & 0.309\\ 0.668 & 1 & 0.420\\ 0.309 & 0.420 & 1\end{pmatrix}\cdot\begin{pmatrix}-0.653\\ 0.740\\ -0.160\end{pmatrix}=\begin{pmatrix}-0.209\\ 0.236\\ -0.051\end{pmatrix}=0.319\times\begin{pmatrix}-0.653\\ 0.740\\ -0.160\end{pmatrix}=\lambda_3\mathbf{V}_3$$

en

$$\mathbf{V_3}'\cdot\mathbf{V_3}=-0.653^2+0.740^2+-0.160^2=1.000$$

Dat deze laatste ook orthogonaal staat op de andere blijkt uit dat:

$$\mathbf{V_1}'\cdot\mathbf{V_3}=\mathbf{V_3}'\cdot\mathbf{V_1}=0\ \text{ en }\ \mathbf{V}_2'\cdot\mathbf{V_3}=\mathbf{V_3}'\cdot\mathbf{V}_2=0$$

Om dit alles samen te vatten kan men zeggen dat wanneer men een PC-analyse op *K* variabelen doet, en dus een *K*×*K* correlatiematrix heeft, men *K* hoofdcomponenten kan extraheren, en de *j*-de PC is de eigenvector corresponderend met de *j*-de eigenwaarde van de correlatiematrix.

De extractie van de eigenvectoren is echter nog maar de eerste stap van de hoofdcomponentenanalyse. Belangrijker dan deze eigenvectoren is de matrix **L**, de structuurmatrix of de patroonmatrix of de matrix van de (factor)ladingen. Voor elke eigenvector $\mathbf{V_j}$, *j* = 1, …, *K*, kan een corresponderende ladingenvector $\mathbf{L_j}$ berekent worden, als het product van de eigenvector $\mathbf{V_j}$ met de vierkantswortel van de eigenwaarde λ_j geassocieerd met de eigenvector in kwestie:

$$\mathbf{L}_j = \mathbf{V}_j\sqrt{\lambda_j} = \begin{pmatrix} v_{1j} \\ v_{2j} \\ \vdots \\ v_{Kj} \end{pmatrix}\sqrt{\lambda_j} = \begin{pmatrix} v_{1j}\sqrt{\lambda_j} \\ v_{2j}\sqrt{\lambda_j} \\ \vdots \\ v_{Kj}\sqrt{\lambda_j} \end{pmatrix} = \begin{pmatrix} l_{1j} \\ l_{2j} \\ \vdots \\ l_{Kj} \end{pmatrix} \qquad (011)$$

De structuurmatrix **L** is dan gewoon de matrix opgebouwd uit de verschillende kolomvectoren $\mathbf{L_j}$, $j = 1, \ldots, K$:

$$\mathbf{L} = \mathbf{L}_1 \mid \mathbf{L}_2 \mid \ldots \mid \mathbf{L}_K = \begin{pmatrix} l_{11} & l_{12} & \ldots & l_{1K} \\ l_{21} & l_{22} & \ldots & l_{2K} \\ \vdots & \vdots & l_{ij} & \vdots \\ l_{K1} & l_{K2} & \ldots & l_{KK} \end{pmatrix} \qquad (012)$$

waarbij l_{ij} de lading is van variabele X_i op factor Y_j. Deze ladingen zijn niets anders dan de correlaties van de geobserveerde variabelen $X_1 \ldots X_K$ met de latente factoren $Y_1 \ldots Y_K$.

$$\mathbf{L_1} = \mathbf{V_1}\sqrt{\lambda_1} = \begin{pmatrix} 0.604 \\ 0.637 \\ 0.479 \end{pmatrix}\sqrt{1.949} = \begin{pmatrix} 0.844 \\ 0.899 \\ 0.668 \end{pmatrix}$$

wat betekent dat de waargenomen variabele X_1 0.844 laadt op of correleert met de eerste principale component, terwijl de variabele X_2 er 0.899 mee correleert en X_3 0.668. Deze ladingen laten dus toe om de verschillende hoofdcomponenten te interpreteren en te benoemen. Op een analoge manier kunnen $\mathbf{L_2}$ en $\mathbf{L_3}$ berekend worden:

$$\mathbf{L_2} = \mathbf{V}_2\sqrt{\lambda_2} = \begin{pmatrix} -0.456 \\ -0.216 \\ 0.863 \end{pmatrix}\sqrt{0.732} = \begin{pmatrix} -0.390 \\ -0.185 \\ 0.738 \end{pmatrix}$$

en

$$\mathbf{L_3} = \mathbf{V_3}\sqrt{\lambda_3} = \begin{pmatrix} -0.653 \\ 0.740 \\ -0.160 \end{pmatrix}\sqrt{0.319} = \begin{pmatrix} -0.369 \\ 0.418 \\ -0.090 \end{pmatrix}$$

Tabel 0-1: Structuurmatrix voorbeeld

	Y_1	Y_2	Y_3	s^2
X_1	0.844	-0.390	-0.369	1.000
X_2	0.889	-0.185	0.418	1.000
X_3	0.668	0.738	-0.090	1.000
λ	1.949	0.731	0.319	3.000
% var	65.0%	24.4%	10.6%	

Tabel 0-1 geeft de structuurmatrix voor het voorbeeld weer in tabelvorm. In de rijen staan de waargenomen variabelen X_1, X_2 en X_3, en in de kolommen de latente hoofdcomponenten Y_1, Y_2 en Y_3. De ladingen van de waargenomen variabelen staan in de cellen, waarbij l_{ij} is de correlatie van variabele X_i met de principale component Y_j. De structuurmatrix heeft echter enkele interessante eigenschappen.

Bijvoorbeeld, de som der kwadraten van de ladingen over een PC (een kolom) is gelijk aan de eigenwaarde corresponderend met die PC:

$$\sum_{i=1}^{K} l_{ij}^{\ 2} = \lambda_j = \mathbf{L_j}' \mathbf{L_j}$$

dit volgt rechtstreeks uit de constructie van de ladingenvector als $\mathbf{L_j} = (\lambda_j)^{0.5}\mathbf{V_j}$ en uit dat de lengte van $\mathbf{V_j} = 1$. Als men bv. de som van de kwadraten van de ladingen op de eerste PC neemt, krijgt men de eerste eigenwaarde λ_1:

$$\sum_{i=1}^{K} l_{i1}^{\ 2} = 0.844^2 + 0.889^2 + 0.668^2 = 1.949 = \lambda_1$$

De totale variantie in een $K \times K$ correlatiematrix is gelijk aan K. Daar correlaties de covarianties zijn tussen gestandaardiseerde variabelen zijn de diagonale elementen van een correlatiematrix (r_{ii}) de varianties van de gestandaardiseerde variabelen. De totale variatie over alle variabelen is de som van de diagonale elementen van de correlatiematrix (tr(**R**)), en aangezien alle $r_{ii} = 1$, $i = 1, \ldots, K$, is de som van deze elementen ook gelijk aan het aantal variabelen K[5]. In het voorbeeld is de totale variantie dan ook gelijk aan $K = 3$.

Anderzijds is ook de som van K eigenwaarden van **R** gelijk aan de totale variantie K. In het voorbeeld heeft men:

$$\sum_{i=1}^{3} \lambda_i = 1.949 + 0.731 + 0.319 = 3.000 = K$$

Dit is ook maar logisch daar hoofdcomponentenanalyse de variantie van de waargenomen variabelen herverdeelt over de latente PC-en. De som van de varianties van de PC-en moet dan ook gelijk zijn aan die van de waargenomen variabelen. De variantie van een PC is gelijk aan de eigenwaarde voor die PC: $\text{Var}(Y_j) = \lambda_j$. De proportie van de totale variantie verklaard door een gegeven PC Y_j is de ratio van de eigenwaarde λ_j over de totale variantie of het aantal waargenomen variabelen K:

$$\%\ \text{verklaarde variantie door } Y_j = \frac{\lambda_j}{K}.$$

De term “verklaarde variantie” dient hier met een korreltje zout genomen worden. In het Engels maakt men een onderscheid tussen de variantie “accounted for” en de variantie “explained”. Bij hoofdcomponenten (en factoren) wordt ‘verklaard’ in de eerste betekenis gebruikt. Wat deze technieken doen is de oorspronkelijke variantie van de set variabelen herverdelen over een beperkter aantal PCs of factoren. Wanneer men dus stelt dat een PC X% van de variantie verklaard, wil men in feit zeggen dat deze X% van de totale variantie van de oorspronkelijke variabelenset vertegenwoordigd.

Waar de som van de kwadraten van de ladingen over een PC gelijk is aan de eigenwaarde of de variantie van die PC, is de som van de kwadraten van de ladingen over een waargenomen variabele gelijk aan de variantie van deze variabele. In het geval van gestandaardiseerde variabelen moet de som van de kwadraten van de ladingen over een waargenomen variabele dus

[5] Wanneer men de covariantiematrix gebruikt i.p.v. de correlatiematrix, dan is de totale variantie gelijk aan tr(**S**), of de som van de variaties van alle waargenomen variabelen.

gelijk zijn aan 1. Dit is een gevolg van de orthogonaliteit van de weerhouden PCs, waardoor elke PC een uniek deel van de variantie van een variabele dekt.

$$\sum_{j=1}^{K} l_{ij}^{\ 2} = 1$$

We kunnen de score van observatie *i* op de waargenomen variabele X_j schrijven als een lineaire combinatie van de scores van *i* op de *K* hoofdcomponenten:

$$X_{ij} = v_{j1}Y_{i1} + v_{j2}Y_{i2} + \ldots + v_{jK}Y_{iK},$$

waar Y_{ik} de score van observatie *i* op de *k*-de PC is, en v_{jk} het element van de eigenvector voor de *k*-de PC is, corresponderend met de *j*-de waargenomen variabele. Als we deze variabelen allen centreren:

$$\bar{X}_j = \bar{Y}_1 = \bar{Y}_2 = \ldots = \bar{Y}_K = 0,$$

kunnen we de variantie van X_j schrijven als:

$$\begin{aligned}
\text{Var}\ X_j &= \frac{1}{N}\sum_{i=1}^{N} X_{ij}^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} \left(v_{j1}Y_{i1} + v_{j2}Y_{i2} + \ldots + v_{jK}Y_{iK}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} \left(v_{j1}^2Y_{i1}^2 + v_{j2}^2Y_{i2}^2 + \ldots + v_{jK}^2Y_{iK}^2 + 2v_{j1}v_{j2}Y_{i1}Y_{i2} + \ldots + 2v_{jK-1}v_{jK}Y_{iK-1}Y_{iK}\right) \\
&= v_{j1}^2\frac{1}{N}\sum_{i=1}^{N}Y_{i1}^2 + v_{j2}^2\frac{1}{N}\sum_{i=1}^{N}Y_{i2}^2 + \ldots + v_{jK}^2\frac{1}{N}\sum_{i=1}^{N}Y_{iK}^2 \\
&\quad + 2v_{j1}v_{j2}\frac{1}{N}\sum_{i=1}^{N}Y_{i1}Y_{i2} + \ldots + 2v_{jK-1}v_{jK}\frac{1}{N}\sum_{i=1}^{N}Y_{iK-1}Y_{iK} \\
&= v_{j1}^2\,\text{Var}\left(Y_1\right) + v_{j2}^2\,\text{Var}\left(Y_2\right) \ldots + v_{jK}^2\,\text{Var}\left(Y_K\right) \\
&\quad + 2v_{j1}v_{j2}\,\text{Cov}\left(Y_1,Y_2\right) \ldots + 2v_{jK-1}v_{jK}\,\text{Cov}\left(Y_{K-1},Y_K\right)
\end{aligned}$$

daar de PCs Y_1, …, Y_K orthogonaal zijn, zijn alle Cov(Y_i,Y_j) = 0, *i*, *j* = 1,…, *K*, $i \neq j$, en kan dit vereenvoudigd worden tot:

$$\begin{aligned}
\text{Var}\ X_j &= v_{j1}^2\,\text{Var}\ Y_1 + v_{j2}^2\,\text{Var}\ Y_2 + \ldots + v_{jK}^2\,\text{Var}\ Y_K \\
&= \sum_{k=1}^{K} v_{jk}^2\,\text{Var}\ Y_k \\
&= \sum_{k=1}^{K} v_{jk}^2\lambda_k \\
&= \sum_{k=1}^{K} l_{jk}^2
\end{aligned}$$

Wanneer men dit toepast op de 1^{e} variabele in het voorbeeld, krijgt men:

$$\sum_{k=1}^{K} l_{1k}^{\ 2} = 0.844^2 + -0.390^{\ 2} + -0.369^{\ 2} = 1,$$

en de variantie van een gestandaardiseerde variabele is altijd 1[6].

Zoals gesteld in Vgl. 010 is de geobserveerde correlatiematrix **R** gelijk aan het product van de structuurmatrix **L** met zijn getransponeerde **L'**:

$$\mathbf{LL'} = \hat{\mathbf{R}},$$

indien **L** een $K \times K$ matrix is en dus alle K PC-en weerhouden werden, dan zal **L·L' R** perfect reproduceren: $\hat{\mathbf{R}} = \mathbf{R}$**.** Dit wordt bv. hieronder gedemonstreerd voor het 3-variabelen voorbeeld en waar—op enkele afrondingsfouten na—de correlatiematrix **R** geproduceerd wordt.

$$\mathbf{L} \cdot \mathbf{L'} = \begin{pmatrix} 0.844 & -0.390 & -0.369 \\ 0.889 & -0.185 & 0.418 \\ 0.668 & 0.738 & -0.090 \end{pmatrix} \cdot \begin{pmatrix} 0.844 & 0.889 & 0.668 \\ -0.390 & -0.185 & 0.738 \\ -0.369 & 0.418 & -0.090 \end{pmatrix}$$
$$= \begin{pmatrix} 1.001 & 0.668 & 0.309 \\ 0.668 & 0.999 & 0.420 \\ 0.309 & 0.420 & 0.999 \end{pmatrix}$$

Indien men echter niet alle hoofdcomponenten weerhoudt en alleen de belangrijkste weerhoudt en $\mathbf{L_r}$ een $K \times P$ matrix is, waarbij $P < K$, dan zal $(\mathbf{L_r})(\mathbf{L_r})'$ de correlatiematrix **R** niet perfect reproduceren. Als men bv. alleen de eerste principale component weerhoudt in het voorbeeld, dan is $\mathbf{L_r}$:

$$\mathbf{L_r} = \begin{pmatrix} 0.844 \\ 0.889 \\ 0.668 \end{pmatrix}$$

en is $\hat{\mathbf{R}} = \mathbf{L_r} \cdot \mathbf{L_r'}$:

$$\mathbf{L_r} \cdot \mathbf{L_r'} = \begin{pmatrix} 0.844 \\ 0.889 \\ 0.668 \end{pmatrix} \cdot \begin{matrix} 0.844 & 0.889 & 0.668 \end{matrix}$$
$$= \begin{pmatrix} 0.712 & 0.750 & 0.564 \\ 0.750 & 0.790 & 0.594 \\ 0.564 & 0.594 & 0.446 \end{pmatrix}$$

wat duidelijk verschilt van de correlatiematrix **R**. Het verschil tussen $\hat{\mathbf{R}}$ en **R** is:

[6] Indien men met de covariantiematrix werkt is de som der kwadraten van de ladingen over een waargenomen variabele gelijk aan de variantie van die variabele s^2.

$$\hat{\mathbf{R}} - \mathbf{R} = \begin{pmatrix} 0.712 & 0.750 & 0.564 \\ 0.750 & 0.790 & 0.594 \\ 0.564 & 0.594 & 0.446 \end{pmatrix} - \begin{pmatrix} 1 & 0.668 & 0.309 \\ 0.668 & 1 & 0.420 \\ 0.309 & 0.420 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} -0.288 & 0.082 & 0.255 \\ 0.082 & -0.210 & 0.174 \\ 0.255 & 0.174 & -0.554 \end{pmatrix}$$

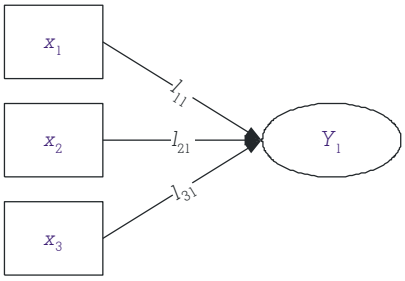
Dit verschil tussen de gereproduceerde correlatiematrix en de waargenomen correlatiematrix kan gebruikt worden om de weerhouden hoofdcomponenten oplossing te diagnoseren. Hier zien we bv. dat de eerste PC wel vrij goed de variantie in X_1 en X_2 vat en ook de correlatie tussen X_1 en X_2, maar relatief minder goed de variantie in X_3 en de correlatie van X_3 met X_1 en X_2.

De waarden op de diagonalen van $\mathbf{L_rL_r'}$ zijn de communaliteiten van de geobserveerde variabelen, d.w.z. de variantie in de geobserveerde variabelen verklaard door de weerhouden hoofdcomponenten.

De geschatte correlatie tussen twee variabelen X_i en X_j, $\hat{r}_{ij}$ is element *ij* van matrix **LL'**, in het geval dat het aantal weerhouden vectoren *P* gelijk is aan het aantal variabelen *K* zal $(\mathbf{LL'})_{ij}$ of $\hat{r}_{ij}$ exact gelijk zijn aan r_{ij}. Indien, echter $P < K$ zal $(\mathbf{LL'})_{ij}$ r_{ij} niet exact reproduceren ($\hat{r}_{ij} \neq r_{ij}$). De geschatte waarde $\hat{r}_{ij}$ is dan de correlatie tussen X_i en X_j die veroorzaakt wordt door de weerhouden PCs of factoren. In het voorbeeld is de waargenomen correlatie tussen X_1 en X_3, $r_{13} = 0.309$. De geschatte correlatie $\hat{r}_{13}$ op basis van de eerste PC alleen is:

$$\hat{r}_{13} = l_{11}l_{31} = 0.844 \times 0.668 = 0.564,$$

een duidelijke overschatting van de geobserveerde correlatie. Hoe we tot dit resultaat komen wordt duidelijk in de onderstaande figuur. De enige band tussen X_1 en X_3 of x_1 en x_3 wordt gevormd door Y_1. Aangezien we werken met gestandaardiseerde variabelen zijn l_{11} en l_{31} gelijk aan respectievelijk $r_{Y_1X_1}$ en $r_{Y_1X_3}$. Het indirecte verband tussen x_1 en x_3 is dan niets anders dan het product van hun relaties met de latente variabele Y_1: $\hat{r}_{13} = r_{Y_1X_1}r_{Y_1X_1} = l_{11}l_{31}$.

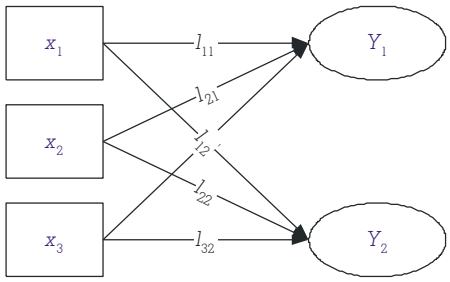


Figuur 0-1: Diagram van een hoofdcomponentenanalyse met 1 latente variabele

Hetzelfde principe geldt wanneer we $\hat{r}_{13}$ berekenen op basis van de eerste twee PCs. In dit geval is de geschatte correlatie:

$$\hat{r}_{13} = l_{11}l_{31} + l_{12}l_{32} = 0.844 \times 0.668 + \ -0.390 \ \times 0.738 = 0.276,$$

wat al een stuk dichter bij de waargenomen waarde voor r_{13} ligt. In dit geval is de geschatte correlatie tussen X_1 en X_3 de som van de indirect verbanden via Y_1 en Y_2, zoals duidelijk wordt uit onderstaande figuur.



Figuur 0-2: Diagram van een hoofdcomponentenanalyse met 2 latente variabelen

In het algemeen kan men dus stellen dat de correlatie tussen twee waargenomen variabelen X_i en X_j kan geschat worden als de som over alle factoren van het product van hun ladingen op de factoren:

$$\hat{r}_{ij} = \sum_{p=1}^{P} l_{ip} l_{jp} \ .$$

Indien men dit doet over alle K PC-en zal men zo de waargenomen r_{ij} exact reproduceren.

Statistische kenmerken van de latente hoofdcomponenten en factoren

Alhoewel de factoren niet gemeten worden kan men hun statistische eigenschappen toch afleiden uit de wijze waarop ze aangemaakt worden (Mardia et al., 1979).

Indien men vertrekt van een reeks van K variabelen **X** met gemiddelden μ en variantie-covariantie Σ en de latente factoren **Y** dan kan men stellen dat:

1. $\mathrm{E}\ Y_i = 0$,
2. $\mathrm{Var}\ Y_i = \lambda_i$,
3. $\mathrm{Cov}\ Y_i, Y_j = 0, i \neq j$,
4. $\mathrm{Var}\ Y_1 \geq \mathrm{Var}\ Y_2 \geq \ldots \geq \mathrm{Var}\ Y_K$,
5. $\sum_{i=1}^{K} \mathrm{Var}\ Y_i = \mathrm{tr}\,\Sigma = \sum_{i=1}^{K} \sigma_{ii}$, en
6. $\prod_{i=1}^{K} \mathrm{Var}\ Y_i = |\Sigma|$.

Indien men vertrekt van gestandaardiseerde variabelen is **X** = **x**, μ = **0** en $\Sigma = \mathbf{R}$. Punten 1 t.e.m. 4 volgen rechtstreeks van de manier waarop de hoofdcomponenten werden geëxtraheerd. Punt 5 werd ook reeds aangehaald daar de totale variantie van de PC-en gelijk moet zijn aan die van de waargenomen variabelen, maar volgt ook uit de eigenschap van eigenwaarden dat de som der eigenwaarden gelijk is aan de som van de diagonale elementen van de matrix. Punt 6 is ook

gewoon een eigenschap van eigenwaarden, het product van de eigenwaarden van een matrix is gelijk aan de determinant van die matrix.

Voorbeeld 1: Hoofdcomponentenanalyse

In Voorbeeld 1 analyseert men de redenen die respondenten geven voor hun sociale participatie, en vraagt men zich af of de 14 redenen die vermeld worden kunnen herleid worden tot een kleiner aantal redenen. Een hoofdcomponentenanalyse op deze 14 variabelen begint met het bereken van de eigenwaarden voor de correlatiematrix **R**. De resultaten worden getoond in de vector λ, waar de eigenwaarden gerangschikt zijn van hoogste naar laagste. Het aantal eigenwaarden is gelijk aan aantal variabelen (K = 14).

$$\lambda = \begin{pmatrix} 4.221 \\ 1.812 \\ 1.032 \\ 0.984 \\ 0.932 \\ 0.844 \\ 0.781 \\ 0.669 \\ 0.610 \\ 0.579 \\ 0.473 \\ 0.393 \\ 0.375 \\ 0.294 \end{pmatrix}$$

Er is een duidelijk verschil in de grootte van de verschillende eigenwaarden. De grootste eigenwaarde is 4.2 terwijl de kleinste minder dan 0.3 bedraagt. De principale component geassocieerd met de grootste eigenwaarde zal dan ook evenveel variantie verklaren als 4.22 waargenomen variabelen of 4.221/14 of 30% van de totale variantie, terwijl de PC geassocieerd met de kleinste eigenwaarde minder dan 30% van de variantie van een enkele waargenomen variabele verklaard of 0.294/14 = 2% van de totale variantie.

De 1ste principale component is de eigenvector geassocieerd met grootste eigenwaarde λ_1 = 4.221. Deze kan worden berekend door de vergelijking op te lossen:

$$\mathbf{R} \cdot \mathbf{V_1} = \lambda_1 \mathbf{V_1}$$

De ladingen van de geobserveerde variabelen op deze eerste PC, of de correlatie van de geobserveerde variabelen *V55* tot *V68* met deze eerste PC, zijn:

$$\mathbf{L_1} = \sqrt{\lambda_1}\mathbf{V_1}$$

Als we dit uitrekenen krijgen we voor $\mathbf{V_1}$ en $\mathbf{L_1}$:

$$\mathbf{V_1} = \begin{pmatrix} 0.359 \\ 0.375 \\ 0.284 \\ 0.298 \\ 0.376 \\ 0.206 \\ 0.125 \\ 0.215 \\ 0.369 \\ 0.208 \\ 0.230 \\ 0.173 \\ 0.185 \\ 0.137 \end{pmatrix} \text{ en } \mathbf{L_1} = \sqrt{4.221} \times \mathbf{V_1} = \begin{pmatrix} 0.738 \\ 0.771 \\ 0.584 \\ 0.612 \\ 0.772 \\ 0.423 \\ 0.256 \\ 0.441 \\ 0.759 \\ 0.427 \\ 0.473 \\ 0.356 \\ 0.379 \\ 0.281 \end{pmatrix}$$

Daarna kan men aanvangen met de extractie van de 2e principale component, deze moet orthogonaal zijn op de eerste en correspondeert met de tweede hoogste eigenwaarde λ_2 = 1.812. Deze wordt berekent door volgende vergelijkingen op te lossen:

$$\mathbf{R} \cdot \mathbf{V_2} = \lambda_2 \mathbf{V_2}$$

$$(\mathbf{V_1})' \mathbf{V_2} = (\mathbf{V_2})' \mathbf{V_1} = 0$$

De laatste vergelijking, met de orthogonaliteitsconditie, kan men in de praktijk negeren daar eigenvectoren uit symmetrische matrices altijd orthogonaal zijn. De vector met ladingen voor deze PC is:

$$\mathbf{L}_2 = \sqrt{\lambda_2} \mathbf{V}_2$$

wat de volgende resultaten geeft:

$$\mathbf{V_2} = \begin{pmatrix} -0.253 \\ -0.216 \\ 0.019 \\ -0.026 \\ -0.177 \\ 0.326 \\ 0.344 \\ -0.201 \\ -0.198 \\ 0.117 \\ 0.098 \\ 0.511 \\ 0.495 \\ 0.152 \end{pmatrix} \text{ en } \mathbf{L_2} = \sqrt{1.812} \times \mathbf{V_2} = \begin{pmatrix} -0.341 \\ -0.291 \\ 0.025 \\ -0.036 \\ -0.239 \\ 0.438 \\ 0.464 \\ -0.270 \\ -0.266 \\ 0.157 \\ 0.132 \\ 0.688 \\ 0.666 \\ 0.204 \end{pmatrix}$$

Deze procedure herhaalt men dan voor alle eigenwaarden tot en met de kleinste eigenwaarde (> 0), en er steeds voor zorgend dat de gekozen eigenvectoren orthogonaal staan op alle vorige. In dit voorbeeld betekent men dat men 14 hoofdcomponenten kan extraheren.

In praktijk extraheert men niet alle factoren maar alleen de belangrijkste, d.w.z. deze die het meeste van de totale variantie verklaren. Gewoonlijk extraheert men alleen de factoren corresponderend met eigenwaarden groter dan 1. Het heeft weinig zin om de resterende factoren met eigenwaarden kleiner dan 1 te bepalen daar die minder van de totale variantie verklaren dan een individuele geobserveerde variabele. Daar het doel van hoofdcomponentenanalyse het reduceren van het aantal variabelen is heeft het geen zin om factoren aan te maken die minder variantie verklaren dan een individuele variabele.

In de onderstaande patroonmatrix zijn de ladingen voor alle 14 hoofdcomponenten weergegeven. Maar alleen de eerste 3 hebben een eigenwaarde groter dan 1. De eerste PC verklaart het grootste deel van de totale variantie, in dit geval meer dan 30%, de tweede PC verklaart nog een extra 13% en de derde nog zo'n 7%. Samen zijn der eerste drie PCs goed voor 50.5% van de totale variantie in de variabelen met betrekking tot redenen voor sociale participatie.

Tabel 0-2: Patroonmatrix voor de hoofdcomponentenanalyse voor Voorbeeld 1

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12	PC 13	PC 14
V55	0.738	-0.341	-0.037	-0.140	-0.209	0.165	-0.024	-0.061	-0.129	-0.059	0.109	0.241	0.180	0.347
V56	0.771	-0.291	0.083	-0.038	-0.191	0.032	-0.089	0.057	-0.217	-0.106	0.038	0.040	0.243	-0.379
V57	0.584	0.025	0.193	0.269	-0.041	-0.497	-0.083	0.180	-0.194	0.464	-0.021	0.036	-0.056	0.056
V58	0.612	-0.035	0.048	0.305	0.029	-0.407	0.147	0.005	0.455	-0.332	0.031	0.149	-0.002	-0.010
V59	0.772	-0.239	0.095	0.020	-0.162	0.029	-0.010	-0.014	0.069	-0.085	-0.012	-0.534	-0.022	0.110
V60	0.424	0.438	0.166	-0.369	0.053	-0.199	-0.026	-0.639	0.028	0.100	0.017	-0.002	0.022	-0.037
V61	0.256	0.464	0.463	-0.201	0.008	0.101	0.612	0.230	-0.137	-0.077	-0.038	0.013	-0.011	0.018
V62	0.441	-0.270	0.143	-0.346	0.587	0.162	-0.100	0.203	0.322	0.234	-0.003	0.009	0.108	-0.018
V63	0.759	-0.266	-0.073	-0.150	-0.046	0.190	-0.038	-0.024	-0.041	-0.032	-0.141	0.136	-0.489	-0.073
V64	0.427	0.157	-0.520	-0.034	0.541	-0.220	0.076	0.033	-0.339	-0.213	-0.081	-0.063	0.031	0.042
V65	0.473	0.132	-0.515	0.282	-0.086	0.277	0.415	-0.120	0.151	0.326	0.078	-0.010	0.055	-0.069
V66	0.356	0.688	-0.125	-0.122	-0.075	0.056	-0.275	0.236	0.040	-0.052	0.464	-0.018	-0.099	-0.012
V67	0.379	0.667	-0.118	-0.004	-0.207	0.132	-0.273	0.146	0.131	-0.011	-0.456	0.033	0.121	0.034
V68	0.280	0.204	0.378	0.619	0.354	0.380	-0.170	-0.186	-0.122	-0.068	0.032	0.024	-0.009	0.020
□	4.221	1.812	1.032	0.984	0.932	0.844	0.781	0.669	0.610	0.579	0.473	0.393	0.375	0.294
% var.	30.2%	12.9%	7.4%	7.0%	6.7%	6.0%	5.6%	4.8%	4.4%	4.1%	3.4%	2.8%	2.7%	2.1%

Dat ook hier de som der kwadraten van de ladingen over een gegeven PC gelijk is aan de eigenwaarde voor die PC blijkt wanneer men bv. de som berekent voor de 1ste en 4de PC:

$$1^{ste}\ PC: \sum_{i=1}^{K} l_{i1}^2 = (0.738)^2 + (0.771)^2 + (0.584)^2 + (0.612)^2 + (0.772)^2 + (0.424)^2 + (0.256)^2 + (0.441)^2 + (0.759)^2 + (0.427)^2 + (0.473)^2 + (0.356)^2 + (0.379)^2 + (0.280)^2 = 4.221 = \lambda_1$$

$$4^{de}\ PC: \sum_{i=1}^{K} l_{i4}^2 = (-0.140)^2 + (-0.038)^2 + (0.269)^2 + (0.305)^2 + (0.020)^2 + (-0.369)^2 + (-0.201)^2 + (-0.346)^2 + (-0.150)^2 + (-0.034)^2 + (0.282)^2 + (-0.122)^2 + (-0.004)^2 + (0.619)^2 = 0.984 = \lambda_4$$

En de som der kwadraten van de ladingen over een geobserveerde variabele is hier ook steeds gelijk aan 1, zoals we hier demonstreren voor *V55* en *V67*:

$$V55: \sum_{j=1}^{K} l_{V55,j}^2 = (0.738)^2 + (-0.341)^2 + (-0.037)^2 + (-0.140)^2 + (-0.209)^2 + (0.165)^2$$
$$+ (-0.024)^2 + (-0.061)^2 + (-0.129)^2 + (-0.059)^2 + (0.109)^2 + (0.241)^2 + (0.180)^2 + (0.347)^2 = 1$$

$$V67: \sum_{j=1}^{K} l_{V67,j}^2 = (0.379)^2 + (0.667)^2 + (-0.118)^2 + (-0.004)^2 + (-0.207)^2 + (0.132)^2 + (-0.273)^2 +$$
$$(0.146)^2 + (0.131)^2 + (-0.011)^2 + (-0.456)^2 + (0.033)^2 + (0.121)^2 + (0.034)^2 = 1$$

Ook hier kan de correlatiematrix **R** gereconstrueerd worden vanuit **L**, ook hier is **R** = **L**·**L'** en als we de ladingen voor alle 14 PC-en gebruiken zal **L**·**L'** perfect de correlatiematrix in Tabel 0-2 reproduceren. Gebruiken we echter niet alle 14 PC-en, dan zal de geschatte correlatiematrix $\hat{\mathbf{R}}$ de geobserveerde correlatiematrix **R** alleen maar benaderen. Hoe minder PC-en men weerhoudt in $\mathbf{L}_r$ hoe slechter de schatting $\hat{\mathbf{R}} = \mathbf{L}_r \cdot \mathbf{L}_r'$ de matrix **R** zal reproduceren. Het verschil tussen **R** en $\hat{\mathbf{R}}$ kan gebruikt worden om de fit van de PC analyse te bepalen. De proportie van de variantie verklaard door de weerhouden PC-en doet dit echter ook en is eenvoudiger te berekenen. Indien *P* hoofdcomponenten weerhouden worden dan is de proportie van de totale variantie verklaard door deze PC-en gelijk aan:

$$\frac{\sum_{i=1}^{P} \lambda_i}{K}$$

Interpretatie van de latente variabelen

De ladingen in de structuurmatrix zijn ook belangrijk voor het identificeren van de verschillende hoofdcomponenten. Door te kijken naar de correlaties van de waargenomen variabelen met de latente variabelen kan men deze laatsten identificeren. Een eerste stap hierbij is de grens te bepalen waarboven we de ladingen sterk genoeg vinden om ze te gebruiken voor de identificatie van de hoofdcomponenten. Er is geen strikt statistisch gegeven om deze grens te bepalen, hoewel men het significantieniveau van de ladingen zou kunnen berekenen voor dit doel. Veel gebruikte grenzen zijn 0.30 en 0.40. Indien de absolute waarde van de lading van een waargenomen variabele groter is dan deze grens dan telt deze variabele mee voor de identificatie van de PC. Bij het bepalen van deze grens moet je oppassen dat je niet te veel maar ook niet te weinig relevante variabelen krijgt. Een groot aantal relevante variabelen maakt het vaak moeilijker om een gemeenschappelijke dimensie te onderscheiden, terwijl bij een klein aantal variabelen men niet zeker kan zijn dat men de juiste interpretatie geeft aan een PC.

Tabel 0-3: Identificatie van hoofdcomponenten, een voorbeeld

	l_{ij}
1ste PC	
V 59 E) Identifying with people who were suffering	0.772
V 56 B) Compassion for those in need	0.771
V 63 I) To help give disadvantaged people hope and dignity	0.759
V 55 A) A sense of solidarity with the poor and disadvantaged	0.738
V 58 D) A sense of duty, moral obligation	0.612

V 57 C) An opportunity to repay something, give something back	0.584
V 65 K) To bring about social or political change	0.473
V 62 H) Religious beliefs	0.441
V 64 J) To make a contribution to my local community	0.427
V 60 F) Time on my hands, wanted something worthwhile to do	0.424
2de PC	
V 66 L) For social reasons, to meet people	0.688
V 67 M) To gain new skills and useful experience	0.667
V 61 G) Purely for personal satisfaction	0.464
V 60 F) Time on my hands, wanted something worthwhile to do	0.438
3de PC	
V 64 J) To make a contribution to my local community	-0.520
V 65 K) To bring about social or political change	-0.515
V 61 G) Purely for personal satisfaction	0.463

In dit voorbeeld trokken we de grens bij 0.40 en besloten we alleen de eerste drie hoofdcomponenten te weerhouden. Dit leverde 10 relevante variabelen op voor de 1ste PC, 4 voor de 2de en 3 voor de 3de. In Tabel 0-3 zijn de relevante items voor de eerste drie PC-en en hun ladingen weergegeven. De items zijn gesorteerd volgens de sterkte van hun lading op elk van de PC-en.

Eerst kijken we naar de items die het hoogst laden op de eerste PC, dit zijn variabelen *V59*, *V56*, *V63*, en *V55*. Deze variabelen meten het belang van volgende redenen: identificatie met anderen die lijden, medelijden met mensen in nood, geven van hoop en waardigheid aan minderbedeelden, en solidariteit met de armen en minderbedeelden. Elk van deze items laadt positief op de PC, wat betekent dat een hoge score op deze variabelen—en dus een groter belang van deze redenen—gepaard gaat met een hogere score op de latente variabele. Op basis van deze 4 items zou je deze PC "solidariteit" of "sociaal bewustzijn" kunnen dopen. De resterende items voor deze PC lijken deze identificatie te steunen.

De items die hoog en positief laden op de 2de PC laden verwijzen niet langer naar derden die mogelijk profiteren van de inzet van de respondent of naar andere sociale redenen voor participatie, maar wel naar het profijt dat de respondent persoonlijk haalt uit zijn of haar sociale participatie. Men zou deze 2de PC dan ook kunnen samenvatten als "persoonlijke redenen" voor sociale participatie.

De twee items die het sterkst laden op de 3de PC, laden negatief op deze PC wat betekent dat een hoge score op de variabelen gepaard gaat met een lage score op de PC. Voor respondenten die hoog scoren op de 3de PC zijn "bijdragen aan de gemeenschap" en "streven naar sociale en politieke verandering" dus onbelangrijke redenen voor sociale participatie. Wat wel een belangrijke reden is, is "persoonlijk genot". Om deze redenen zou men deze PC "egoïsme" kunnen labellen.

Factoranalyse

Hoewel filosofisch er duidelijk verschillen zijn tussen hoofdcomponentenanalyse en factoranalyse, is de basismethode niet zo erg verschillend. Waar hoofdcomponentenanalyse vertrekt van **R**, de correlatiematrix voor de variabelen, vertrekt factoranalyse van een aangepaste correlatiematrix $\mathbf{R_c}$ waarin de waarden op de diagonaal (1-en) vervangen zijn door de communaliteiten van de variabelen c_i^2. De communaliteit van een variabele is de proportie van de variantie van deze variabele die verklaard wordt door de weerhouden factoren. Men gaat er vanuit dat elke variabele

een deel gemeenschappelijke variantie (de communaliteit) bezit maar ook een deel unieke variantie. Dit betekent dat;

$$\mathbf{R} = \mathbf{R_c} + \mathbf{U}$$

waarbij **U** een diagonale matrix is met op de diagonale cellen het unieke deel van de variantie van de geobserveerde variabelen.

$$\begin{pmatrix} 1 & r_{12} & \dots & r_{1K} \\ r_{12} & 1 & \dots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1K} & r_{2K} & \dots & 1 \end{pmatrix} = \begin{pmatrix} c_1^2 & r_{12} & \dots & r_{1K} \\ r_{12} & c_2^2 & \dots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1K} & r_{2K} & \dots & c_K^2 \end{pmatrix} + \begin{pmatrix} u_1^2 & 0 & \dots & 0 \\ 0 & u_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & u_K^2 \end{pmatrix}$$

De $\mathbf{R_c}$ matrix wordt dan gebruikt om de eigenwaarden λ, eigenvectoren $\mathbf{V}$ en de ladingen **L** te schatten. In de meest gebruikte methode, principal axis factoring (PAF) of principale assen factoranalyse (SPSS, 1991), wordt deze procedure iteratief toegepast en wordt in elke stap de communaliteiten herberekend afhankelijk van het aantal weerhouden factoren *m*:

$$c_i^2 = \sum_{j=1}^{m} l_{ij}^2$$

waar c_i^2 de communaliteit voor variabele X_i is. Deze communaliteiten worden dan in $\mathbf{R_c}$ ingevoerd en de eigenwaarden, eigenvectoren en ladingen opnieuw berekend. Dit gaat door tot men een stabiele oplossing gevonden heeft. Voor de eerste iteratie wordt $\mathbf{R_c}$ gelijkgesteld aan **R**, of m.a.w. $\forall i, i = 1 \ldots K : c_i^2 = 1$. De eerste stap voor een PAF analyse is dus een PC analyse.

Waar in hoofdcomponentenanalyse alleen als er *K* PC-en weerhouden worden **LL'** = **R**, kan in factoranalyse ook reeds met minder dan *K* factoren **LL'** gelijk zijn aan $\mathbf{R_c}$. Trouwens in praktijk daar men ook een uniek deel van de variantie moet voorzien is het maximum aantal te extraheren latente variabelen in factoranalyse gelijk aan *K* – 1. Ook hier is echter zo dat hoe groter het aantal weerhouden factoren hoe beter **LL'** de aangepaste correlatiematrix $\mathbf{R_c}$ voorspelt.

Als men dit toepast op het voorbeeld met de drie variabelen beginnen we met een PC analyse op de correlatiematrix $\mathbf{R_{(0)}}$, waarbij het subscript (0) staat voor de nulde iteratie. In deze nulde iteratie is $\mathbf{R_{(0)}} = \mathbf{R}$. De elementen van deze matrix duiden we aan als $r_{ij(0)}$ de correlatie tussen variabelen X_i en X_j in iteratie 0.

$$\mathbf{R}_0 = \begin{pmatrix} 1 & 0.688 & 0.309 \\ 0.668 & 1 & 0.420 \\ 0.309 & 0.420 & 1 \end{pmatrix}$$

De eigenwaarden voor $\mathbf{R_{(0)}}$ worden berekend, en in $\lambda_{(0)}$ is er maar 1 eigenwaarde groter dan 1. Slechts één enkele factor zal dus weerhouden worden voor de verdere analyse, *P* = 1. De selectie van het aantal factoren te weerhouden gebeurt in deze nulde stap en dit aantal wordt aangehouden tijdens de rest van factoranalyse.

$$\lambda_0 = \begin{pmatrix} 1.949 \\ 0.732 \\ 0.319 \end{pmatrix}$$

De eigenvector $\mathbf{V_{1(0)}}$ en ladingenvector $\mathbf{L_{1(0)}}$ geassocieerd met $\lambda_{1(0)} = 1.949$ zijn respectievelijk:

$$\mathbf{V}_{1\,0} = \begin{pmatrix} 0.604 \\ 0.637 \\ 0.479 \end{pmatrix} \textbf{ en } \mathbf{L}_{1\,0} = \sqrt{1.949}\mathbf{V}_{1\,0} = \begin{pmatrix} 0.844 \\ 0.889 \\ 0.668 \end{pmatrix}$$

Dit resultaat is identiek aan de hoofdcomponentenanalyse beschreven in sectie XXXX

Na dit voorbereidend werk begint de iteratieve procedure. In elke stap *t* worden de waarden op de diagonaal van de correlatiematrix, de varianties van de gestandaardiseerde variabelen, vervangen door de communaliteiten geschat op de weerhouden factoren:

$$\begin{cases} r_{ii\,t} = \sum_{p=1}^{P} l_{ip}^2 \\ r_{ij\,t} = r_{ij\,t-1}, i \neq j \end{cases} \qquad (013)$$

waarop dan opnieuw de factoren geschat worden. Deze procedure blijft doorgaan tot een vooraf bepaald criterium bereikt is. Dit brengt dan ook mee dat de resultaten zullen verschillen naargelang het aantal factoren dat weerhouden worden, daar dit een effect gaat hebben op de communaliteit van de variabelen. Hoe meer factoren weerhouden worden, hoe groter de communaliteiten. Indien het aantal weerhouden factoren gelijk is aan het aantal variabelen is de communaliteit voor alle variabelen gelijk aan 1, en zullen de resultaten van de factoranalyse identiek zijn aan die van de hoofdcomponentenmethode.

In het voorbeeld worden de communaliteiten voor de waargenomen variabelen dus berekend op alleen de eerste factor:

$$c_{1\,1}^2 = r_{11\,1} = l_{11\,0}^2 = 0.844^2 = 0.712$$
$$c_{2\,1}^2 = r_{22\,1} = l_{21\,0}^2 = 0.889^2 = 0.791$$
$$c_{3\,1}^2 = r_{33\,1} = l_{31\,0}^2 = 0.668^2 = 0.447$$

De aangepaste correlatiematrix $\mathbf{R_{(1)}}$ wordt dan:

$$\mathbf{R}_1 = \begin{pmatrix} 0.712 & 0.688 & 0.309 \\ 0.668 & 0.791 & 0.420 \\ 0.309 & 0.420 & 0.447 \end{pmatrix}$$

De eigenvectoren voor $\mathbf{R_{(1)}}$ zijn:

$$\lambda_1 = \begin{pmatrix} 1.645 \\ 0.244 \\ 0.061 \end{pmatrix}$$

en de eigenvector $\mathbf{V_{(1)}}$ en ladingenvector $\mathbf{L_{(1)}}$ geassocieerd met de grootste eigenwaarde $\lambda_{1(1)}$ zijn respectievelijk:

$$\mathbf{V}_{1\,1} = \begin{pmatrix} 0.618 \\ 0.679 \\ 0.397 \end{pmatrix} \text{ en } \mathbf{L}_{1\,1} = \sqrt{1.645}\,\mathbf{V}_{1\,1} = \begin{pmatrix} 0.792 \\ 0.871 \\ 0.509 \end{pmatrix}$$

De ladingen in L1(1) worden dan op hun beurt gebruikt om de communaliteiten voor de tweede iteratie ($c^2_{i\,2}$) te berekenen. Dit iteratief proces gaat dan door tot een stabiele oplossing bereikt werd. In dit voorbeeld werd een stabiele oplossing bereikt in de 31ste iteratie. De communaliteiten in deze iteratie zijn:

$$c^2_{1\,31} = r_{11\,31} = l^2_{11\,30} = 0.495$$
$$c^2_{2\,31} = r_{22\,31} = l^2_{21\,30} = 0.900$$
$$c^2_{3\,31} = r_{33\,31} = l^2_{31\,30} = 0.195$$

De enige weerhouden factor verklaart dus 49.5% van de variantie in X_1, 90.0% van de variantie in X_2 en 19.5% van de variantie in X_3. Dat betekent dat variabele X_1 50.5% unieke variantie bezit, X_2 10.0% en X_3 80.5%. Deze factor duidt er dus op dat vooral X_3 maar weinig variantie gemeenschappelijk heeft met de twee andere variabelen. De aangepaste correlatie matrix voor deze iteratie $\mathbf{R}_{(31)}$ is:

$$\mathbf{R}_{31} = \begin{pmatrix} 0.495 & 0.668 & 0.309 \\ 0.668 & 0.900 & 0.420 \\ 0.309 & 0.420 & 0.195 \end{pmatrix}$$

De eigenwaarden voor deze correlatiematrix zijn:

$$\lambda_{31} = \begin{pmatrix} 1.590 \\ 0.002 \\ -0.002 \end{pmatrix}$$

De uiteindelijke factor en ladingenvector geassocieerd met de grootste eigenwaarde $\lambda_{1(31)}$ = 1.590 zijn dan respectievelijk:

$$\mathbf{V}_1 = \mathbf{V}_{1\,31} = \begin{pmatrix} 0.558 \\ 0.753 \\ 0.350 \end{pmatrix} \text{ en } \mathbf{L}_1 = \mathbf{L}_{1\,31} = \sqrt{1.590}\,\mathbf{V}_{1\,31} = \begin{pmatrix} 0.703 \\ 0.949 \\ 0.441 \end{pmatrix}$$

Het is dan ook de tweede variabele X_2 die heel sterk laadt op de factor met een lading van 0.949, terwijl X_3 maar matig tot zwak laadt.

Voorbeeld 1: Factoranalyse

Een dergelijke factoranalyse kan ook verricht worden op de correlatiematrix voor de redenen voor sociale participatie. In de nulde iteratie wordt de hoofdcomponentenanalyse uit sectie 0herhaald. De 14×1 vector λ is de vector met de eigenwaarden $\lambda_{(0)}$ voor de correlatiematrix $\mathbf{R}_{(0)}$. Maar 3 van de 14 eigenwaarden zijn groter dan 1 wat betekent dat alleen de eerste 3 factoren weerhouden worden. De ladingen van de waargenomen variabelen op deze drie factoren worden weergeven in de 14×3 patroonmatrix $\mathbf{L} = \mathbf{L}_{(0)}$. De vector **C** is de vector met de communaliteiten van de

waargenomen variabelen geschat op basis van de drie weerhouden factoren en die in de volgende iteratie op de diagonaal van $\mathbf{R}_{(1)}$ zullen geplaatst worden. De elementen van deze vector **C**, $c^2_{k(1)}$ worden berekent als:

$$c^2_{k\,t} = r_{kk\,t} = \sum_{p=1}^{3} l^2_{kp\,t-1}$$

$$\lambda: \begin{bmatrix} 0.610 \\ 0.579 \\ 0.669 \\ 0.473 \\ 0.393 \\ 0.375 \\ 0.781 \\ 0.844 \\ 0.294 \\ 0.932 \\ 0.984 \\ 1.032 \\ 1.812 \\ 4.221 \end{bmatrix} \quad L: \begin{bmatrix} 0.738 & -0.341 & -0.036 \\ 0.771 & -0.291 & 0.084 \\ 0.584 & 0.025 & 0.189 \\ 0.612 & -0.036 & 0.045 \\ 0.772 & -0.239 & 0.095 \\ 0.423 & 0.438 & 0.171 \\ 0.256 & 0.464 & 0.465 \\ 0.441 & -0.270 & 0.148 \\ 0.759 & -0.266 & -0.071 \\ 0.427 & 0.157 & -0.519 \\ 0.473 & 0.132 & -0.518 \\ 0.356 & 0.688 & -0.125 \\ 0.379 & 0.666 & -0.117 \\ 0.281 & 0.204 & 0.371 \end{bmatrix} \quad C: \begin{bmatrix} 0.662 \\ 0.686 \\ 0.377 \\ 0.378 \\ 0.663 \\ 0.401 \\ 0.497 \\ 0.289 \\ 0.652 \\ 0.476 \\ 0.510 \\ 0.616 \\ 0.602 \\ 0.258 \end{bmatrix}$$

De eerste iteratie maakt gebruik van correlatiematrix $\mathbf{R}_{(1)}$ waarbij de diagonale elementen vervangen zijn door de communaliteiten uit **C**. In de resultaten hieronder staat λ voor $\lambda_{(1)}$, de vector met eigenwaarden voor de eerste iteratie, **L** voor $\mathbf{L}_{(1)}$, de patroonmatrix voor iteratie 1 en **C** voor $\mathbf{C}_{(2)}$, de vector met de communaliteiten voor gebruik in de tweede iteratie.

$$\lambda := \begin{bmatrix} 0.124 \\ 0.075 \\ 0.068 \\ 0.018 \\ -0.024 \\ -0.038 \\ -0.068 \\ 0.182 \\ 0.297 \\ 0.361 \\ 0.389 \\ 0.516 \\ 1.380 \\ 3.785 \end{bmatrix} \quad L := \begin{bmatrix} 0.723 & -0.301 & 0.014 \\ 0.758 & -0.260 & 0.112 \\ 0.524 & 0.029 & 0.083 \\ 0.550 & -0.015 & -9.310 \cdot 10^{-3} \\ 0.753 & -0.206 & 0.100 \\ 0.380 & 0.343 & 0.150 \\ 0.235 & 0.381 & 0.377 \\ 0.388 & -0.178 & 0.044 \\ 0.739 & -0.226 & -0.044 \\ 0.391 & 0.146 & -0.389 \\ 0.441 & 0.123 & -0.398 \\ 0.338 & 0.637 & -0.041 \\ 0.359 & 0.609 & -0.040 \\ 0.241 & 0.141 & 0.074 \end{bmatrix} \quad C := \begin{bmatrix} 0.613 \\ 0.654 \\ 0.282 \\ 0.302 \\ 0.620 \\ 0.284 \\ 0.342 \\ 0.184 \\ 0.599 \\ 0.325 \\ 0.368 \\ 0.521 \\ 0.502 \\ 0.083 \end{bmatrix}$$

Dit gaat door tot de oplossing stabiliseert. In dit geval gebeurde dit in iteratie 14: De uiteindelijke uitkomst van de factoranalyse wordt hieronder getoond, waar λ de eigenwaarden voor de finale oplossing zijn, **L** de patroonmatrix, en **C** de vector met communaliteiten voor de waargenomen variabelen.

$$\lambda := \begin{bmatrix} -0.062 \\ -0.068 \\ -0.053 \\ 0.001 \\ 0.015 \\ 0.045 \\ 0.079 \\ -0.127 \\ 0.190 \\ 0.220 \\ -0.241 \\ 0.356 \\ 1.262 \\ 3.718 \end{bmatrix} \quad L := \begin{bmatrix} 0.736 & -0.297 & -0.226 \\ 0.757 & -0.236 & -0.034 \\ 0.528 & 0.050 & 0.305 \\ 0.561 & 0.003 & 0.333 \\ 0.750 & -0.180 & 0.060 \\ 0.364 & 0.324 & -0.011 \\ 0.211 & 0.303 & 0.043 \\ 0.381 & -0.162 & -0.026 \\ 0.744 & -0.210 & -0.188 \\ 0.358 & 0.131 & 0.010 \\ 0.405 & 0.111 & -0.033 \\ 0.327 & 0.638 & -0.138 \\ 0.346 & 0.601 & -0.110 \\ 0.231 & 0.138 & 0.161 \end{bmatrix} \quad C := \begin{bmatrix} 0.681 \\ 0.630 \\ 0.374 \\ 0.426 \\ 0.598 \\ 0.237 \\ 0.138 \\ 0.172 \\ 0.633 \\ 0.145 \\ 0.177 \\ 0.533 \\ 0.493 \\ 0.098 \end{bmatrix}$$

De eerste factor vertoont grote gelijkenissen met de eerste principale component uit sectie 0De variabelen die hoog laden op de eerste factor ($|l_{k1}| > 0.40$) laadden ook allen hoog op de eerste PC. De eerste factor meet dan ook “solidariteits” redenen voor sociale participatie. Op de tweede factor laden maar 2 variabelen hoog: *V66* en *V67*, juist de twee variabelen die ook het hoogste laadden op de 2de PC in sectie 0, deze tweede factor kan dan ook geïnterpreteerd worden als “persoonlijke redenen” voor sociale participatie. Op de derde factor laadt echter geen enkele variabele hoog. Dit maak het moeilijk om deze factor te identificeren. De maximum lading hier is

0.333 (voor variabele *V58*) wat te laag is om een factor mee te identificeren. Dus waar de eerste twee factoren zijn plusminus gelijk aan de eerste twee PC-en uit sectie 0 kan de derde factor hier niet duidelijk geïnterpreteerd worden.

De vector met communaliteiten geeft aan dat deze drie factoren een goed deel van de variantie in variabelen *V55*, *V56*, *V59*, *V63* en *V66* verklaart, terwijl andere variabelen (*V61*, *V62*, *V65* en *V68*) nauwelijks iets gemeenschappelijk hebben met deze drie factoren.

Doordat men hier rekening houdt met de unieke variantie van elk van de variabelen, verklaren de uiteindelijke factoren ook minder van de totale variantie dan de oorspronkelijke factoren of de hoofdcomponenten. De drie weerhouden factoren verklaren respectievelijk 26.6%, 9.0% en 2.6% van de totale variantie in de waargenomen variabelen, of samen 38.1%. Dit is substantieel minder dan de 50.5% die de eerste drie PC-en in sectie 0verklaarden.

Samenvatting

- Hoofdcomponentenanalyse en hoofdassen-factoranalyse zijn de twee meest gebruikte vormen van factoranalyse in de ruime zin.
- Vooraleer tot een factoranalyse over te gaan dient men na te gaan of de gebruikte variabelen er wel geschikt voor zijn. Hiervoor bestaan twee toetsen: de Bartlett toets voor sfeericiteit, en de Kaiser-Meyer-Olkin of KMO toets voor toereikendheid van de steekproef. Deze toetsen zijn nu minder belangrijk geworden gezien de huidige statistische programma's factoranalyse heel snel kunnen uitvoeren. Indien de variabelen niet geschikt zijn voor factoranalyse, dan zullen de resultaten ook weinig of niet-bruikbaar zijn.
- Hoofdcomponenten analyse is gebaseerd op de extractie van eigenvectoren uit de correlatiematrix, en elke hoofdcomponent is niets anders dan een genormaliseerde eigenvector vermenigvuldigd met de vierkantswortel van de bijhorende eigenwaarde.
- De eigenwaarde van een hoofdcomponent is proportioneel aan de proportie van de totale variantie van de variabelen in de analyse die door deze hoofdcomponent gevat wordt. De proportie variantie gevat door een hoofdcomponent is gelijk aan de bijhorende eigenwaarde gedeeld door het aantal variabelen in de analyse.
- Hoofdassen-factoranalyse bouwt voort op hoofdcomponentenanalyse. De eerste stap is gelijk aan een hoofdcomponentenanalyse, maar in volgende stappen worden de gestandaardiseerde varianties van de variabelen in de correlatiematrix vervangen door hun communaliteiten, d.w.z., door de variantie van de variabele die gevat wordt door de weerhouden factoren. Deze procedure wordt iteratief uitgevoerd tot de resultaten convergeren.

Referenties

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3*, 77-85.

Dunteman, G. H. (1989). *Principal components analysis* (Sage university paper series on quantitative applications in the social sciences No. 07-069). Newbury park: Sage.

Dziuban, C. D., & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin, 81*(6), 358-361.

Jobson, J. D. (1992). *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods*. New York: Springer Verlag.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*, 401-416.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

MathSoft. (1997). Mathcad (Version 7 Professional) [CD-R]. Cambridge, MA: MathSoft.

SPSS. (1991). *SPSS statistical algorithms* (2nd ed.). Chicago, IL : SPSS.

SPSS. (2006). *SPSS 15.0 Algorithms*. Chicago, IL: SPSS Inc.

World values survey, 1981-1984 and 1990-1993 [Data file]. (1994). World Values Study Group (producer). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

CENTER FOR SOCIAL THEORY – DEPARTMENT OF SOCIOLOGY – GHENT UNIVERSITY
KORTE MEER 3-5, 9000 GENT, BELGIUM
SOCIO@UGENT.BE

There is nothing more practical than a good theory

(K. Lewin, 1952)

CENTRE FOR SOCIAL THEORY - DEPARTMENT OF SOCIOLOGY - UNIVERSITEIT GENT - KORTE MEER 3-5, 9000 GENT, BELGIUM