Stel, je hebt van de verkiezingen voor de Tweede Kamer van november 2006 van de tien grootste partijen voor elk van de 459 Nederlandse gemeenten de percentages die ze behaalden. Hoe breng je deze scores overzichtelijk in beeld? Bijvoorbeeld door een principale componentenanalyse te maken. **Klaas-Jan Wieringa** beschrijft hoe dat te doen, hij vertelde er over op NWD 13.

Multivariate statistiek

Inleiding

We beschouwen voor de verkiezingen voor de Tweede Kamer van november 2006 de percentages die de tien grootste partijen in elk van de 459 Nederlandse gemeenten behaalden (ontleend aan www.verkiezingsuitslagen.nl).

Er zijn tien variabelen X_1 tot en met X_{10} (CDA, PVDA, ... tot en met PVV), en we hebben voor elke gemeente in Nederland de scores op deze tien variabelen, ofwel het percentage voor elk van de tien partijen. Hoe breng je deze scores overzichtelijk in beeld? Voor één partij zou je een plaatje kunnen maken van de percentages, voor twee partijen zou je de resultaten tegen elkaar uit kunnen zetten in een assenstelsel. In die gevallen maak je dus een keuze voor één of twee variabelen die je gebruikt. De informatie uit de andere variabelen wordt dan genegeerd. Het zou mooi zijn om de informatie van alle tien variabelen te gebruiken, en toch te komen tot een overzichtelijk plaatje. Daarbij is het belangrijk om er rekening mee te houden dat de variabelen onderling afhankelijk zijn. Multivariate statistiek biedt hiervoor passende methoden. Met één daarvan, principale componentenanalyse, kun je met die tien variabelen de belangrijkste componenten berekenen. Andere methoden zijn onder meer factoranalyse en clusteranalyse.

Tabel 1: percentages voor de tien grootste partijen in alle Nederlandse gemeenten

Nr	Gemeente	CDA	PvdA	VVD	SP	GL	D66	CU	SGP	PVDD	PVV
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	Aa en Hunze	15.23	33.78	21.27	15.32	4.83	1.56	2.19	0.12	1.50	3.49
2	Aalburg	35.88	8.27	9.07	7.55	0.77	0.38	11.51	19.32	1.22	5.33
3	Aalsmeer	34.79	14.38	23.35	9.15	3.09	1.55	4.27	0.37	1.82	6.05
4	Aalten	41.09	17.15	9.61	17.96	2.49	0.91	5.65	0.17	1.15	3.21
5	Abcoude	30.94	14.70	25.52	8.89	6.80	3.51	3.02	0.58	1.86	3.46
6	Achtkarspelen	34.73	25.67	6.09	15.05	1.88	0.30	11.08	0.76	0.92	3.04
7	Alblasserdam	23.08	20.49	9.41	13.40	1.76	1.59	9.65	12.17	1.16	5.86
8	Albrandswaard	27.01	15.43	21.65	12.13	2.62	1.68	3.73	1.11	1.99	9.74
9	Alkemade	42.85	12.89	19.80	11.15	3.24	1.17	1.14	0.11	1.36	5.11
10	Alkmaar	19.36	23.62	13.96	21.96	6.95	2.37	2.37	0.10	2.22	5.66
...	...	...	...	...	...	...	...	...	...	...	...
19	Amsterdam	9.59	30.13	13.94	18.47	12.49	4.48	1.54	0.08	3.46	4.46
...	...	...	...	...	...	...	...	...	...	...	...
459	Zwolle	22.39	24.05	11.65	17.33	5.81	1.97	10.58	0.51	1.52	3.44

Een principale componentenanalyse levert het plaatje onderaan deze bladzijde. De gemeenten zijn weergegeven in een assenstelsel waarbij de assen (*pc1* en *pc2*) overeenkomen met de twee belangrijkste principale componenten.

De eerste en belangrijkste principale component is:

$$pc1 = 0,35 \cdot cda - 0,28 \cdot pvda - 0,26 \cdot vvd - 0,23 \cdot sp - 0,43 \cdot gl$$
$$-0,40 \cdot d66 + 0,28 \cdot cu + 0,31 \cdot sgp - 0,40 \cdot pvdd - 0,02 \cdot pvv$$

De tweede, op één na belangrijkste, principale component is:

$$pc2 = 0,04 \cdot cda - 0,32 \cdot pvda + 0,41 \cdot vvd - 0,56 \cdot sp + 0,19 \cdot gl$$
$$+0,29 \cdot d66 + 0,26 \cdot cu + 0,27 \cdot sgp + 0,23 \cdot pvdd - 0,32 \cdot pvv$$

Berekening van de principale componenten

We hebben n items en we hebben p variabelen, gerangschikt in een matrix x met n rijen en p kolommen:

	Variabele	X_1	X_2	...	X_p
Item					
1		x_{11}	x_{12}	...	x_{1p}
2		x_{21}	x_{22}	...	x_{2p}
. . .		. . .	. . .		. . .
n		x_{n1}	x_{n2}	...	x_{np}

Standaardiseren

Bij een principale componentenanalyse is het gebruikelijk om de gegevens te standaardiseren, al zijn er soms ook argumenten om dat juist niet te doen. Voor deze analyse voer ik een standaardisering als volgt uit.

Voor elke variabele X_j rekenen we uit: het gemiddelde

$$\bar{x}_j = \frac{1}{n} \cdot \sum_{k=1}^{n} x_{kj}$$ en de standaarddeviatie

$$\sigma_j = \sqrt{\frac{\sum_{k=1}^{n} (x_{kj} - \bar{x}_j)^2}{n}}$$

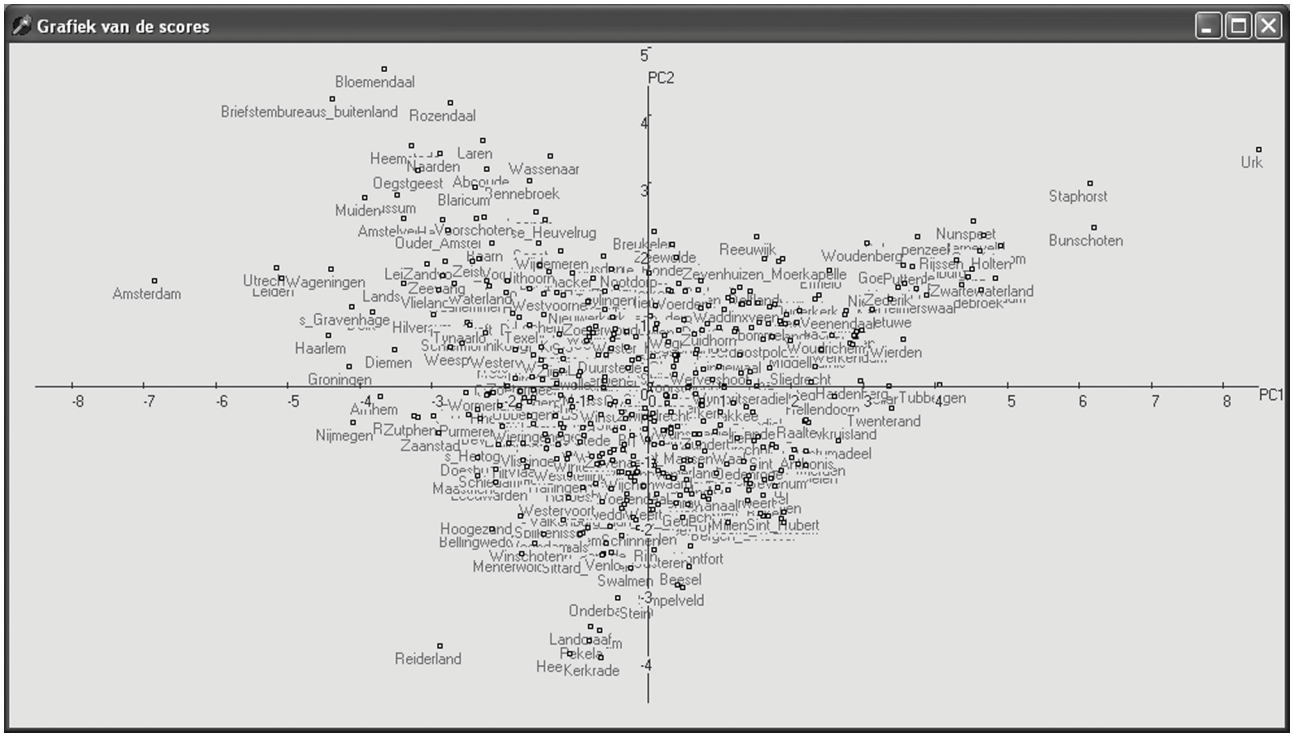
We berekenen een gestandaardiseerde gegevensmatrix z met de formule:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

Voorbeeld: in de gemeente nummer 1 (Aa en Hunze) scoort partij nummer 2 (de PVDA) 33,78 (x_{12}). Landelijk scoort de PVDA gemiddeld 18,99 ($\bar{X}_2$) met een standaarddeviatie van 6,25 (σ_2). Het gemiddelde percentage van de PVDA over alle gemeenten is niet gelijk aan het percentage dat de PVDA landelijk van alle stemmen behaalde, omdat in de berekening van $\bar{X}_j$ alle gemeenten even zwaar meetellen. De gestandaardiseerde score van de PVDA in de gemeente Aa en Hunze is:

$$z_{12} = \frac{x_{12} - \bar{x}_2}{\sigma_2} = \frac{33,78 - 18,99}{6,25} = 2,37$$

We krijgen nu de tabel boven aan de volgende bladzijde.



Tabel 2: gestandaardiseerde scores voor de tien grootste partijen in alle Nederlandse gemeenten

Nr	Gemeente	CDA	PvdA	VVD	SP	GL	D66	CU	SGP	PVDD	PVV
		Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
1	Aa en Hunze	-1,83	2,37	1,27	0,00	0,76	0,15	-0,52	-0,43	-0,19	-0,89
2	Aalburg	0,59	-1,72	-1,19	-1,69	-1,62	-1,31	1,81	3,67	-0,68	-0,20
3	Aalsmeer	0,47	-0,74	1,69	-1,34	-0,26	0,14	0,00	-0,38	0,36	0,08
4	Aalten	1,21	-0,29	-1,08	0,57	-0,61	-0,65	0,35	-0,42	-0,80	-1,00
5	Abcoude	0,01	-0,69	2,12	-1,40	1,91	2,58	-0,31	-0,33	0,43	-0,90
6	Achtkarspelen	0,46	1,07	-1,79	-0,06	-0,97	-1,41	1,70	-0,30	-1,20	-1,06
7	Alblasserdam	-0,91	0,24	-1,12	-0,42	-1,04	0,19	1,34	2,14	-0,78	0,00
8	Albrandswaard	-0,45	-0,57	1,34	-0,70	-0,54	0,30	-0,13	-0,22	0,66	1,47
9	Alkemade	1,41	-0,98	0,97	-0,91	-0,18	-0,33	-0,78	-0,44	-0,44	-0,28
10	Alkmaar	-1,35	0,74	-0,20	1,44	2,00	1,16	-0,47	-0,44	1,06	-0,07
...											
19	Amsterdam	-2,49	1,78	-0,21	0,69	5,25	3,78	-0,68	-0,44	3,21	-0,52
...											
459	Zwolle	-0,99	0,81	-0,67	0,44	1,33	0,66	1,58	-0,35	-0,16	-0,91

Lineaire combinaties met maximale variantie

Beschouw nogmaals de eerste principale component:

$$pc1 = 0,35 \cdot cda - 0,28 \cdot pvda - 0,26 \cdot vvd - 0,23 \cdot sp - 0,43 \cdot gl$$
$$-0,40 \cdot d66 + 0,28 \cdot cu + 0,31 \cdot sgp - 0,40 \cdot pvdd - 0,02 \cdot pvv$$

En uitgedrukt in de gestandaardiseerde waarden:

$$pc1 = 0,35 \cdot z_1 - 0,28 \cdot z_2 - 0,26 \cdot z_3 - 0,23 \cdot z_4 - 0,43 \cdot z_5$$
$$-0,40 \cdot z_6 + 0,28 \cdot z_7 + 0,31 \cdot z_8 - 0,40 \cdot z_9 - 0,02 \cdot z_{10}$$

De waarden van *pc1* en *pc2* voor de eerdergenoemde gemeenten:

Nr	Gemeente	*pc1*	*pc2*
1	Aa en Hunze	-2,20	-0,13
2	Aalburg	4,53	1,71
3	Aalsmeer	0,03	1,65
4	Aalten	1,49	-0,82
5	Abcoude	-2,23	3,20
6	Achtkarspelen	2,21	-1,19
7	Alblasserdam	1,73	0,27
8	Albrandswaard	-0,48	0,68
9	Alkemade	0,76	0,82
10	Alkmaar	-2,97	-0,44
...			
19	Amsterdam	-6,85	1,57
...			
459	Zwolle	-0,92	0,20

De varianties (kwadraat van de standaarddeviatie) van de principale componenten *pc1* en *pc2* zijn

$$\sigma_{pc1}{}^2 \approx \frac{(-2,20)^2 + 4,53^2 + 0,03^2 + \ldots + (-0,92)^2}{459} \approx 3,7$$

en

$$\sigma_{pc2}{}^2 \approx \frac{(-0,13)^2 + (1,71)^2 + \ldots + (0,20)^2}{459} \approx 2,1 .$$

Deze zijn opvallend hoog in vergelijking met de varianties van de gestandaardiseerde variabelen z_1 tot en met z_{10}, want die hebben een variantie van 1. Voor de gestandaardiseerde variabelen Z_j geldt: het gemiddelde $\overline{Z_j} = 0$ en de standaarddeviatie $\sigma_z = 1$, en dus de variantie $\sigma_z{}^2 = 1$

Als principale componenten komen in aanmerking de lineaire combinaties van de gestandaardiseerde variabelen $a_1 \cdot z_1 + a_2 \cdot z_2 + \ldots + a_{10} \cdot z_{10}$

waarvoor geldt dat $a_1{}^2 + a_2{}^2 + \ldots + a_{10}{}^2 = 1$.

Je kunt deze lineaire combinaties beschouwen als vectoren met lengte 1 in een tiendimensionale vectorruimte.

– De eerste principale component *pc1* is de lineaire combinatie van z_1 tot en met z_{10} met de grootste variantie.

– De tweede principale component *pc2* is van de lineaire combinaties van z_1 tot en met z_{10} die loodrecht staan op *pc1* degene met de grootste variantie. Dat *pc1* en *pc2* inderdaad loodrecht staan op elkaar is te controleren door het inproduct te berekenen van de vectoren *pc1* en *pc2*.
– De derde principale component *pc3* is van de lineaire combinaties van van z_1 tot en met z_{10} die loodrecht staan op *pc1* en *pc2* degene met de grootste variantie.
– Op deze wijze kun je doorgaan voor de principale componenten *pc4* tot en met *pc10*.

De principale componenten zijn zo nog niet helemaal vastgelegd. Er zijn voor elke principale component twee mogelijkheden die elkaars tegengestelde zijn. We kunnen van deze twee mogelijkheden er willekeurig één enkele kiezen.

Correlatiematrix

Het uitgangspunt voor de berekening van de principale componenten is de 'correlatiematrix'. Bereken tussen elk tweetal variabelen Z_i en Z_j de correlatie r_{ij}:

$$r_{ij} = \frac{\sum_{k=1}^{n} (z_{ki} - \bar{z}_i) \cdot (z_{kj} - \bar{z}_j)}{\sigma_{z_i} \cdot \sigma_{z_j}} = \frac{\sum_{k=1}^{n} (z_{ki} - 0) \cdot (z_{kj} - 0)}{1 \cdot 1} = \sum_{k=1}^{n} z_{ki} \cdot z_{kj}$$

In de correlatiematrix is element *i,j* gelijk aan de correlatie r_{ij} tussen de variabelen *i* en *j*.

De correlatiematrix voor de dataset van de Tweede Kamerverkiezingen 2006 is onderaan deze bladzijde te vinden.

Een correlatie van 1 betekent dat twee partijen volledig positief met elkaar correleren en een correlatie van -1 betekent dat twee partijen volledig negatief met elkaar correleren. Een correlatie van 0 betekent dat er geen correlatie is tussen twee partijen. Uiteraard correleert elke politieke partij perfect met zichzelf, dus de waarden op de hoofddiagonaal zijn 1. De correlatie van partij A met partij B is gelijk aan de correlatie van partij B met partij A, dus de correlatiematrix is symmetrisch ten opzichte van de hoofddiagonaal. In deze correlatiematrix zijn een aantal correlaties tussen partijen opvallend. Bijvoorbeeld CDA en PVDA hebben een onderlinge correlatie van -0,727, dus die zijn behoorlijk sterk negatief gecorreleerd, niet verrassend gezien de politieke standpunten. GroenLinks en D66 hebben een onderlinge correlatie van 0,817, dus zijn sterk met elkaar gecorreleerd. Dat betekent dat een gemeente waarin GroenLinks relatief hoog scoort doorgaans ook een relatief hoge score voor D66 kent, en een relatief lage score van GroenLinks gaat meestal samen met een relatief lage score van D66.

Berekening met eigenwaarden en eigenvectoren

Van de correlatiematrix *C* berekenen we de eigenwaarden en eigenvectoren. Op de volgende bladzijde staat *C* als matrix genoteerd.

Om het geheugen op te frissen, volgt hierna een korte beschrijving van wat eigenwaarden en eigenvectoren zijn. Wie hiermee onbekend is, verwijs ik naar literatuur op het gebied van de lineaire algebra, bijvoorbeeld *Linear algebra and its applications* (Lay, 2003). Gegeven is een $n \times n$ matrix A. Een eigenvector van A is een vector x, waarvan tenminste één element ongelijk aan 0 is, zodanig dat $A \cdot x = \lambda \cdot x$ voor een zeker getal λ. Het getal λ heet een eigenwaarde van A als er een vector x bestaat (met tenminste één element ongelijk aan 0) die voldoet aan de vergelijking $A \cdot x = \lambda \cdot x$. Deze vector x heet een eigenvector bij eigenwaarde λ.

Tabel 3: correlatiematrix Tweede Kamerverkiezingen 2006

	CDA	**PvdA**	**VVD**	**SP**	**GL**	**D66**	**CU**	**SGP**	**PVDD**	**PVV**
CDA	1,000	-0,727	-0,133	-0,357	-0,508	-0,442	0,003	0,020	-0,560	-0,039
PvdA	-0,727	1,000	-0,216	0,555	0,349	0,171	-0,116	-0,339	0,198	-0,114
VVD	-0,133	-0,216	1,000	-0,298	0,363	0,519	-0,345	-0,241	0,636	-0,064
SP	-0,357	0,555	-0,298	1,000	0,164	0,037	-0,529	-0,531	0,058	0,269
GL	-0,508	0,349	0,363	0,164	1,000	0,817	-0,229	-0,361	0,599	-0,206
D66	-0,442	0,171	0,519	0,037	0,817	1,000	-0,234	-0,281	0,595	-0,172
CU	0,003	-0,116	-0,345	-0,529	-0,229	-0,234	1,000	0,581	-0,295	-0,390
SGP	0,020	-0,339	-0,241	-0,531	-0,361	-0,281	0,581	1,000	-0,264	-0,149
PVDD	-0,560	0,198	0,636	0,058	0,599	0,595	-0,295	-0,264	1,000	0,093
PVV	-0,039	-0,114	-0,064	0,269	-0,206	-0,172	-0,390	-0,149	0,093	1,000

$$C = \begin{bmatrix} 1 & -0,727 & -0,133 & -0,357 & -0,508 & -0,442 & -0,003 & 0,020 & -0,560 & -0,039 \\ -0,727 & 1 & -0,216 & -0,555 & 0,349 & 0,171 & -0,116 & -0,339 & 0,198 & -0,114 \\ -0,133 & -0,216 & 1 & -0,298 & 0,363 & 0,519 & 0,345 & -0,241 & 0,636 & -0,064 \\ -0,357 & 0,555 & -0,298 & 1 & 0,164 & 0,037 & -0,529 & -0,531 & 0,058 & 0,269 \\ -0,508 & 0,349 & 0,363 & 0,164 & 1 & 0,817 & -0,229 & -0,361 & 0,599 & -0,206 \\ -0,422 & 0,171 & 0,519 & 0,037 & 0,817 & 1 & -0,234 & -0,281 & 0,595 & -0,172 \\ 0,003 & -0,116 & -0,345 & -0,529 & -0,229 & -0,234 & 1 & 0,581 & -0,295 & -0,390 \\ 0,020 & -0,339 & -0,241 & -0,531 & -0,361 & -0,281 & 0,581 & 1 & -0,264 & -0,149 \\ -0,560 & 0,198 & 0,636 & 0,058 & 0,599 & 0,595 & -0,295 & -0,264 & 1 & 0,093 \\ -0,039 & -0,114 & -0,064 & 0,269 & -0,206 & -0,172 & -0,390 & -0,149 & 0,093 & 1 \end{bmatrix}$$

Matrix van C

Voorbeeld: Gegeven is de matrix $A = \begin{pmatrix} 1 & 6 \\ 5 & 2 \end{pmatrix}$.
De vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is een eigenvector van A,

want $\begin{pmatrix} 1 & 6 \\ 5 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 7 \\ 7 \end{pmatrix} = 7 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

De bijbehorende eigenwaarde λ is 7.

Er zijn numerieke methoden voor de berekening van eigenwaarden en eigenvectoren bij een gegeven matrix. Voor de berekening van de eigenwaarden en eigenvectoren van een symmetrische matrix, zoals een correlatiematrix, zijn er speciale numerieke methoden met aantrekkelijke eigenschappen.

Met behulp van de theorie van kwadratische vormen (uit de lineaire algebra) kan bewezen worden dat een correlatiematrix alleen reële (dus geen complexe) niet-negatieve eigenwaarden heeft. Een $n \times n$ correlatiematrix heeft dus n eigenwaarden $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. Deze eigenwaarden zijn de varianties van de principale componenten. De principale componenten zijn de bijbehorende eigenvectoren.

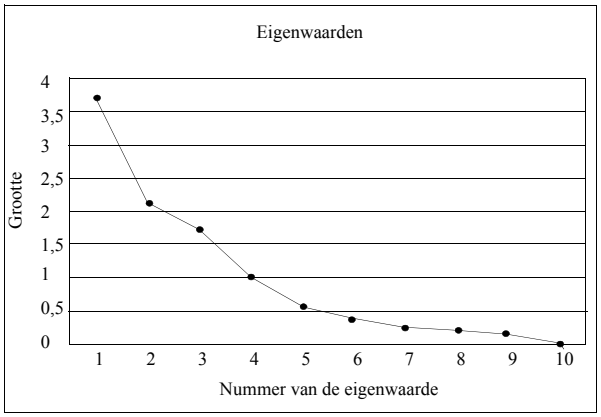
Voor de correlatiematrix C zijn de eigenwaarden $\lambda_1 \approx 3,7$ en $\lambda_2 \approx 2,1$.

De eigenvector bij λ_1 is $\begin{pmatrix} 0,35 \\ -0,28 \\ -0,26 \\ -0,23 \\ -0,43 \\ -0,40 \\ 0,28 \\ 0,31 \\ -0,40 \\ -0,02 \end{pmatrix}$ en bij λ_2 $\begin{pmatrix} 0,04 \\ -0,32 \\ 0,41 \\ -0,56 \\ 0,19 \\ 0,29 \\ 0,26 \\ 0,27 \\ 0,23 \\ -0,32 \end{pmatrix}$

Voor de eigenwaarden geldt dat de som van de eigenwaarden gelijk is aan het 'spoor' van C, dat is de som van de elementen op de hoofddiagonaal van C.
Dus $\lambda_1 + \lambda_2 + \ldots + \lambda_n = c_{11} + c_{22} + \ldots + c_{nn}$. Bij een correlatiematrix zijn de elementen op de hoofddiagonaal 1, en geldt dus $\lambda_1 + \lambda_2 + \ldots + \lambda_n = n$. De som van de varianties van de principale componenten is dus gelijk aan n; dat is ook de som van de varianties van de oorspronkelijke gestandaardiseerde variabelen, want die hebben alle variantie 1. Hiermee kunnen we bepalen hoeveel van de totale variantie de eerste twee principale componenten verklaren. We weten al dat $\lambda_1 \approx 3,7$, $\lambda_2 \approx 2,1$ en $\lambda_1 + \lambda_2 + \ldots + \lambda_{10} = 10$.
De eerste twee principale componenten verklaren dus

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_{10}} \approx 58\%$$ van de totale variantie.

Hieronder een grafiek van alle eigenwaarden van C in aflopende grootte:



Er zijn vuistregels om te bepalen welke principale componenten je in beschouwing moet nemen. Bijvoorbeeld: de derde eigenwaarde is maar weinig kleiner dan de tweede, en deze zou eigenlijk beschouwd moeten worden. Een probleem is dan wel dat een grafiek met drie principale componenten met drie assen niet eenvoudig te visualiseren is, al zijn er wel mogelijkheden voor. Een andere vuistregel is dat je principale componenten behorende bij eigenwaarden groter dan 1 zou moeten beschouwen. Ook daarom zou je de derde principale component eigenlijk niet mogen negeren, maar de vierde en volgende zijn relatief onbelangrijk.

De formule voor de derde principale component is:

$$pc3 = 0,43 \cdot cda - 0,46 \cdot pvda + 0,36 \cdot vvd - 0,04 \cdot sp - 0,13 \cdot gl$$
$$-0,02 \cdot d66 - 0,48 \cdot cu - 0,28 \cdot sgp + 0,08 \cdot pvdd + 0,37 \cdot pvv$$

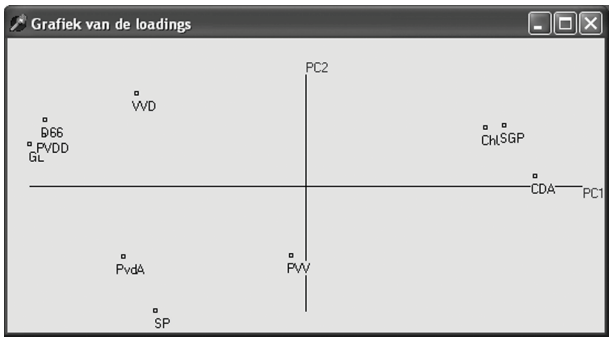
Interpretatie van de principale componenten met een loadingplot

Je krijgt meer inzicht als je probeert de principale componenten te interpreteren. Een hulpmiddel hierbij is de 'loadingplot'. We gebruiken hiervoor dezelfde assen *pc1* en *pc2* als voor de scoreplot. We bepalen de coördinaten van de oorspronkelijke variabelen (de partijen) uit de overeenkomstige coëfficiënten in de principale componenten:

$$pc1 = 0,35 \cdot cda - 0,28 \cdot pvda - 0,26 \cdot vvd - 0,23 \cdot sp - 0,43 \cdot gl$$
$$-0,40 \cdot d66 + 0,28 \cdot cu + 0,31 \cdot sgp - 0,40 \cdot pvdd - 0,02 \cdot pvv$$

$$pc2 = 0,04 \cdot cda - 0,32 \cdot pvda + 0,41 \cdot vvd - 0,56 \cdot sp + 0,19 \cdot gl$$
$$+ 0,29 \cdot d66 + 0,26 \cdot cu + 0,27 \cdot sgp + 0,23 \cdot pvdd - 0,32 \cdot pvv$$

Het CDA heeft dus coördinaten (0,35; 0,04), de PVDA heeft coördinaten (-0,28; -0,32), enzovoort.



Eerste principale component

De partijen CDA, ChristenUnie en SGP hebben een positieve waarde en de partijen PVDA, SP, VVD, D66, GroenLinks en Partij voor de Dieren een negatieve. Je kan deze principale component wellicht interpreteren als christelijk versus seculier. Overigens zijn de aanduiding positief en negatief willekeurig. Alle getallen in de principale component met -1 vermenigvuldigen, levert ook een passende eigenvector op. Ik heb die eigenvector genomen waarvoor de eerste variabele (in dit geval het CDA) een positieve waarde heeft. Vergelijk de loadinggrafiek met de eerdere scoregrafiek. Opvallend zijn de gemeenten op de tegenovergestelde uiteinden van de as van *pc1*: Amsterdam (-6,85) en Urk (8,53).

Tweede principale component

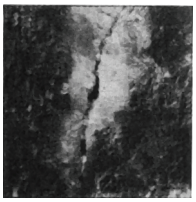
Bij *pc2* scoren VVD, D66, Groen Links, Partij voor de Dieren, ChristenUnie en SGP positief, en PVDA, SP en PVV negatief. Je zou dit kunnen interpreteren als 'rechts' versus 'links'. Gemeenten die op deze as aan de tegenovergestelde uiteinden liggen zijn Bloemendaal (4,68) aan de ene kant; Kerkrade (-4,02) en Reiderland (-3,84) aan de andere kant.

Toepassing met beeldverwerking

In de beeldverwerking worden multivariate methoden veel gebruikt. Een voorbeeld is het verwerken van satellietbeelden. Satellieten kunnen een beeld van een deel van het aardoppervlak maken, waarbij de sensoren energie op verschillende banden van golflengten opnemen, bijvoorbeeld verschillende banden zichtbaar licht, maar ook banden in het infrarood licht. Beschouw de volgende drie foto's, genomen van Railroad Valley, Nevada, USA (overgenomen uit *Linear algebra and its applications* (Lay, 2003)).

Spectraalband 1
visueel blauw

Spectraalband 2
nabij infrarood



Spectraalband 3
midden-infrarood

Sommige informatie is zichtbaar in alledrie de plaatjes, omdat het betreffende object op de aarde in meerdere banden te zien is. Andere objecten zijn door hun afwijkende kleur of temperatuur slechts in één of twee van de banden te zien. Het zou mooi zijn als in één plaatje, eventueel twee, zoveel mogelijk oppervlaktekarakteristieken zichtbaar zijn. Dat kan bereikt worden door een principale componentenanalyse uit te voeren.

Het beeld voor een zekere band van golflengten wordt gedigitaliseerd als een rechthoek van beeldpunten. We nummeren de beeldpunten en leggen voor elk beeldpunt de intensiteit vast als een getal.

Nummering van de beeldpunten

1	2	⋯	1000
1001	1002	⋯	2000
⋮	⋮	⋱	⋮
999001	999002	⋯	1000000

Intensiteit in een bepaalde band

11	23	⋯	35
15	27	⋯	37
⋮	⋮	⋱	⋮
16	19	⋯	40

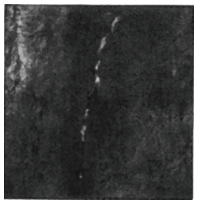
Er kunnen vele rijen en kolommen zijn. Hierboven is uitgegaan van een schema met 1000 bij 1000 beeldpunten.

Zo'n digitalisering is er voor verschillende banden van golflengten, waarbij telkens hetzelfde stuk van het aardoppervlak is gefotografeerd, steeds in een andere band. Voor elk beeldpunt (een klein stukje van het oppervlak) zijn dus scores beschikbaar voor alle gebruikte banden. Rangschik de data nu zo dat elke rij een beeldpunt is en elke kolom een band van golflengten. Dan krijg je zoiets als:

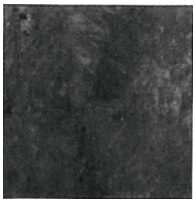
Beeldpuntnummer	**Band 1**	**Band 2**	**Band 3**
1	11	..	..
2	23	..	..
...	..	..	..

Met deze datastructuur kan je een principale componentenanalyse uitvoeren. Omdat er drie variabelen zijn, heeft de correlatiematrix drie rijen en drie kolommen. Je krijgt drie principale componenten. Voor elke principale component kun je een plaatje maken:

Principale component 1: 93.5%



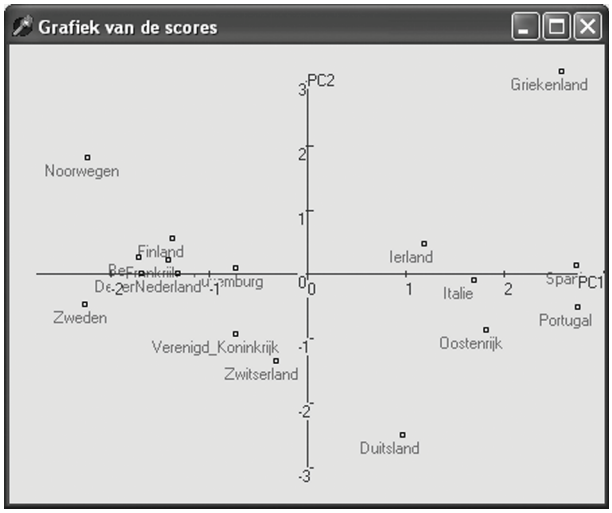
Principale component 2: 5,3%



Principale component 3: 1,2%

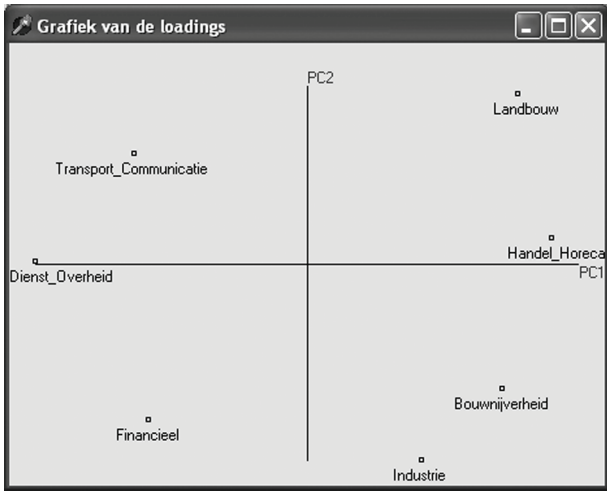
De oorspronkelijke drie plaatjes zijn nog wel in één oogopslag waar te nemen, dus de datareductie is niet zo heel groot. Echter, satellieten kunnen ook in vele honderden smalle banden van golflengten meten, waardoor je niet drie, maar honderden variabelen krijgt. Dan is datareductie natuurlijk extra belangrijk. Er zijn veel van dergelijke toepassingen van multivariate statistische methoden in de beeldverwerking.

Toepassing op economische sectoren in Europese landen



De grafiek met de scores per land. In het oog springt dat sommige landen sterk op elkaar lijken voor wat betreft de structuur van hun economieën.

Uitgangspunt zijn de gegevens van de werkgelegenheid (percentages economisch actieven per sector) in een zevental economische sectoren (Landbouw & Visserij, Industrie, Bouwnijverheid, Transport & Communicatie, Handel & Horeca, Financieel, Dienstverlening & Overheid) in 17 Europese landen (bron: *Britannica Encyclopedie jaarboek 1998, Employment and labour*). We beschouwen de scores en de loadings. De eigenwaarden bij *pc1* en *pc2* zijn 3,3 respectievelijk 1,5. Samen verklaren ze 69% van de variantie. De eigenwaarde bij de derde principale component is 0,95. Deze kan dus buiten beschouwing blijven.



De grafiek met de loadings (coëfficiënten van de variabelen in de principale componenten) is:

Op de eerste principale component zie je een contrast tussen enerzijds landbouw, handel en horeca, bouwnijverheid en industrie, en anderzijds transport en communicatie, dienstverlening en overheid, financieel.

Bedrijfswiskunde aan de Noordelijke Hogeschool Leeuwarden

De vijf bedrijfswiskunde-opleidingen leiden studenten op voor het ondersteunen van beslissingen met wiskundige methoden en met gebruikmaking van ICT. Belangrijke toepassingsgebieden zijn statistiek, actuariële wiskunde, logistiek en wiskundige softwareontwikkeling.

De opleiding Bedrijfswiskunde van de Noordelijke Hogeschool Leeuwarden is één van deze vijf. Studenten krijgen onder meer een aantal statistiekmodulen waarin zaken als beschrijvende statistiek, kansrekening, kansverdelingen, toetsen en schatten, prognosetechnieken en industriële statistiek aan de orde komen. Ook zijn er modulen op het gebied van analyse, operations research, matrixrekening en programmeren.

Daarnaast hebben we een module Case study (in het tweede jaar) waarin studenten een project uitvoeren. Afgelopen jaren was het onderwerp multivariate statistiek en in 2007/2008 zal dat ook zo zijn. De heer Gerben Mooiweer van Sara Lee/Coffee Tea R&D (voorheen Douwe Egberts) is opdrachtgever, verstrekt data, brengt veel deskundigheid in op het gebied van multivariate statistiek en de implementatie in software. Dit is een zeer vruchtbare samenwerking. Op deze manier komen actuele onderwerpen aan de orde die veel toegepast worden in de praktijk.

Ook leren de studenten hierbij in een projectvorm te werken. De studenten zijn in groepen verdeeld. Elke groep heeft een programma geschreven om de principale com-

ponentenanalyse uit te voeren, nadat ze na een globale instructie hadden uitgezocht hoe de methode werkt en hoe de berekening kan plaatsvinden. Toen de programma's klaar waren hebben ze verschillende datasets doorgerekend en geïnterpreteerd.

Sara Lee/Coffee Tea R&D had ons een dataset geleverd van een aantal soorten koffie die op een dertigtal criteria qua smaak en geur zijn beoordeeld, bijvoorbeeld kruidige smaak, lichte smaak, sterke geur, enzovoort. Met principale componentenanalyse kan je een plaatje maken waarbij je op twee assen de verschillende soorten koffie in beeld brengt, of koffie van verschillende productiemomenten. Je kunt zo in één oogopslag zien of een bepaalde soort koffie (te veel) van andere verschilt. Een constante smaak bij producten is belangrijk, want de consument heeft een verwachtingspatroon voor de smaak, en waardeert het niet als een product anders smaakt dan de vorige keer.

Klaas-Jan Wieringa
Noordelijke Hogeschool Leeuwarden
Afdeling Exacte Vakken, Bedrijfswiskunde

Literatuur

Johnson, R.A. & D.W. Wichern (2002). *Applied multivariate statistical analysis.* Prentice Hall.

Lay, D.C. (2003). *Linear algebra and its applications*. Pearson Education.

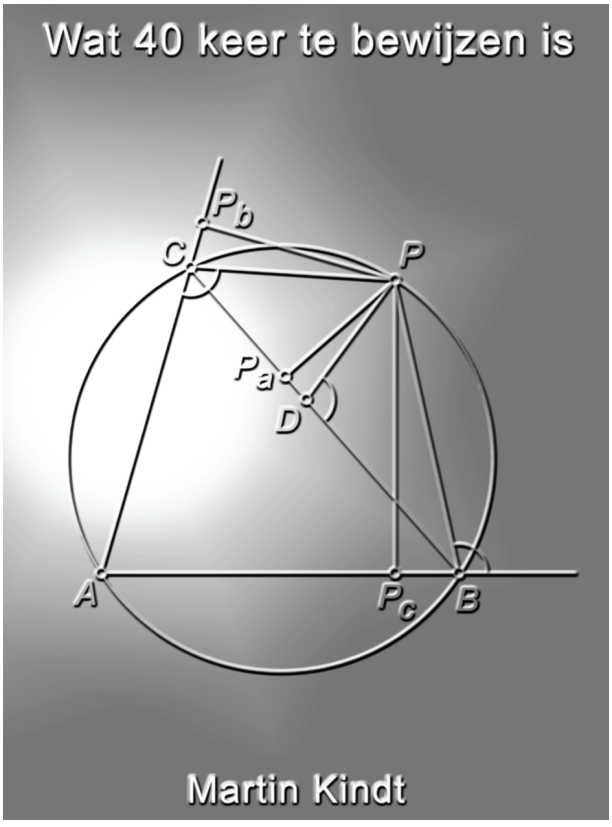
Manly, B.F.J. (1994). *Multivariate statistical methods. A primer*. Chapman & Hall.

Het boek van Manly geeft een inleiding op veel concepten uit de multivariate statistiek, zonder veel wiskundige onderbouwing. Voor een eerste inleiding zou ik dit aanbevelen.

Het boek van Johnson en Wichern gaat er dieper op in en geeft veel wiskundige onderbouwing en geeft bewijzen van relevante stellingen.

Het boek van Lay legt de achtergrond van de gebruikte lineaire algebra helder uit, en heeft een hoofdstuk (pagina 447–492) over symmetrische matrices en kwadratische vormen, en de toepassing daarvan op principale componentenanalyse. Overigens gebruiken we bij de opleiding Bedrijfswiskunde in jaar 1 een eigen dictaat over matrixrekening, dat beknopter en eenvoudiger is dan het boek van Lay.

Software
Er is professionele software voor multivariate data-analyse, bijvoorbeeld het pakket Unscrambler® van Camo; zie www.camo.com. De grafieken in deze tekst heb ik gemaakt met software die ik zelf geschreven heb.



Tien jaar ‘Wat te bewijzen is’ !

Om dat te vieren hebben we alle veertig edities gebundeld. Op de NWD is uitgebreid stilgestaan bij dit heuglijke feit; alle deelnemers hebben een exemplaar van de bundel ontvangen.

Uiteraard willen we de abonnees van de *Nieuwe Wiskrant* die niet op de NWD waren ook in de gelegenheid stellen, tegen verzend- en behandelingskosten (€ 7,50), een exemplaar van deze bundel te bestellen.

Stuur een mailtje met uw adresgegevens naar wiskrant@fi.uu.nl als u ‘Wat 40 keer te bewijzen is’ wilt ontvangen.