Vervolgdoctoraal Fysische Geografie,
Tentamen Matrix algebra en Multivariate Statistiek (MUVS)
28 nov 1997, Tr I, 102

Maak vraag 1, en kies er twee uit de vragen 2, 3 en 4. Antwoord kort.

VRAAG 1

a bereken het inprodukt van de vectoren $(1, 3, 1, 2)$ en $(4, 5, -2, 8)$

gegeven $A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix}$, $b = [7, 5, 9]$, $c = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$.

b bereken bc, cb en $|cb|$.

c inverteer A, controleer je antwoord.

d los $Ax = b'$ op voor x (hint: gebruik hiervoor het antwoord van vraag c, als je dit hebt).

e is c een eigenvector van A?

VRAAG 2

a Leg de verschillen en overeenkomsten tussen discriminantanalyse en arbitrary origin (k-means) clusteranalyse kort uit.

b Beschouw de dataset met de furanen en dioxines uit de twee havens in Rotterdam en de Rijn, bekeken in het prakticum. Welke indeling is, in termen van Wilk's criterium, de beste: de indeling in herkomst (Chemiehaven, Laurenshaven, Rijn) of die gevonden met behulp van k-means clusteranalyse?

c Had je het antwoord op vraag 2b ook kunnen beantwoorden zonder dat je de dataset had gekend?

d Welke van de twee indelingen bij vraag 2b is het meest waardevol voor de interpretatie van de gegevens, en waarom?

VRAAG 3
Gegeven de dataset in onderstaande kruistabel

	locatie 1	locatie 2	locatie 3
soort a	15	9	0
soort b	6	2	4
soort c	1	1	4
soort d	8	8	2

a bereken de tabel met verwachte frequenties, en de tabel met de chi-kwadraat bijdragen.

b welke soort en welke locatie wijken het sterkste af van de marginale verdeling? Waaraan zie je dit?

c Welke soort en welke locatie lijken het sterkste op de marginale verdeling?

d Stel, je wilt bovenstaande tabel ordenen. Welke techniek zou je hiervoor gebruiken? Op welk criterium worden rijen en kolommen dan gesorteerd?

VRAAG 4
Een onderzoeker (Steven de Jong) heeft een set van 70 satellietbeelden van een enkel gebied, elk voor een smal bandje uit het zichtbare of near infrared spectrum. Om een globale indruk te krijgen van het patroon, aanwezig in al deze beelden, gebruikt hij hoofdcomponentenanalyse.

a Waarom?

b Hoe ziet de data matrix er in dit geval uit?

c De eerste drie hoofdcomponenten lijken de beelden vrij goed samen te vatten, maar de interpreteerbaarheid van deze hoofdcomponenten laat te wensen over: het is niet duidelijk hoe een hoofdcomponent samenhangt met een groep beelden. Is hier verbetering in aan te brengen? Zo ja, hoe?

d wat is, bij hoofdcomponentenanalyse, een loading (of loding)? Wat is een score?

a. $(1,3,1,2)\cdot(4,5,-2,8) = 4 + 15 + (-2) + 16 = 33$

b. $bc = \begin{bmatrix} 7 & 5 & 9 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 14+5+9 \end{bmatrix} = \begin{bmatrix} 28 \end{bmatrix}$

$$cb = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}\begin{bmatrix} 7 & 5 & 9 \end{bmatrix} = \begin{bmatrix} 14 & 10 & 18 \\ 7 & 5 & 9 \\ 7 & 5 & 9 \end{bmatrix}$$

$$|cb| = \begin{vmatrix} 14 & 10 & 18 \\ 7 & 5 & 9 \\ 7 & 5 & 9 \end{vmatrix} = \begin{vmatrix} 14 & 10 & 18 \\ 7 & 5 & 9 \\ 0 & 0 & 0 \end{vmatrix} = \begin{vmatrix} 0 & 0 & 0 \\ 7 & 5 & 9 \\ 0 & 0 & 0 \end{vmatrix} = 0$$

c.

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 \\ 3 & 2 & 1 & 0 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 \\ 0 & -4 & -2 & -3 & 0 & 1 \end{array}\right] \rightarrow$$

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & -3 & -2 & -2 & 1 & 0 \\ 0 & -4 & -2 & -3 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} & 0 \\ 0 & -4 & -2 & -3 & 0 & 1 \end{array}\right] \rightarrow$$

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 1 & -1 & -2 & 2 \\ 0 & 1 & \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} & 0 \\ 0 & 0 & -\frac{2}{3} & -\frac{1}{3} & -\frac{4}{3} & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 1 & 1 & -1 \\ 0 & 0 & -\frac{2}{3} & -\frac{1}{3} & -\frac{4}{3} & 1 \end{array}\right]$$

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & -3 & -2 & -2 & 1 & 0 \\ 0 & 0 & \frac{2}{3} & -\frac{1}{3} & -\frac{4}{3} & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 1\frac{1}{2} & 2 & -\frac{3}{2} \\ 0 & -3 & 0 & -3 & -3 & 3 \\ 0 & 0 & \frac{2}{3} & -\frac{1}{3} & -\frac{4}{3} & 1 \end{array}\right] \rightarrow$$

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & -3 & 0 & -3 & -3 & 3 \\ 0 & 0 & \frac{2}{3} & -\frac{1}{3} & -\frac{4}{3} & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 1 & 1 & -1 \\ 0 & 0 & 1 & -\frac{1}{2} & -2 & \frac{3}{2} \end{array}\right]$$

$$A^{-1} = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 1 & -1 \\ -\frac{1}{2} & -2 & \frac{3}{2} \end{bmatrix}$$

controle: $A^{-1}A = I$

$$\begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 1 & -1 \\ -\frac{1}{2} & -2 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$ klopt!

1. $b = [7 \quad 5 \quad 9] \qquad b' = \begin{bmatrix} 7 \\ 5 \\ 9 \end{bmatrix}$

$Ax = b' \rightarrow A^{-1}Ax = A^{-1}b' \rightarrow Ix = A^{-1}b'$

$\rightarrow x = A^{-1}b'$

$$A^{-1}b' = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 1 & -1 \\ -\frac{1}{2} & -2 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} 7 \\ 5 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix} = x$$

$|A - \lambda I| = 0$

$$\begin{vmatrix} 1-\lambda & 2 & 1 \\ 2 & 1-\lambda & 0 \\ 3 & 2 & 1-\lambda \end{vmatrix} = (1-\lambda)\begin{vmatrix} 1-\lambda & 0 \\ 2 & 1-\lambda \end{vmatrix} - 2\begin{vmatrix} 2 & 0 \\ 3 & 1-\lambda \end{vmatrix} + 1\begin{vmatrix} 2 & 1-\lambda \\ 3 & 2 \end{vmatrix}$$

$= (1-\lambda)\left((1-\lambda)^2 - 0\right) - 2\left((2-2\lambda) - 0\right) + \left(4 - (3-3\lambda)\right)$

$= (1-\lambda)^3 + 4\lambda - 4 + 4 - 3 + 3\lambda = (1-\lambda)^3 + 7\lambda - 3 = 0$

als c is eigenvector dan: $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = \lambda \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$ (immers: $Ax = \lambda x$)

$\Rightarrow$ $5 = 2\lambda \rightarrow \lambda = 2\frac{1}{2}$
$5 = \lambda \rightarrow \lambda = 5$
$9 = \lambda \rightarrow \lambda = 9$

Dit kan niet, dus c is geen eigenvector van A

(3) a. Waargenomen:

	locatie 1	locatie 2	locatie 3	totaal (D_r)
soort a	15	9	0	24
soort b	6	2	4	12
soort c	1	1	4	6
soort d	8	8	2	18
totaal (D_k)	30	20	10	60

Verwacht:

	locatie 1	locatie 2	locatie 3	totaal
soort a	12	8	4	24
soort b	6	4	2	12
soort c	3	2	1	6
soord d	9	6	3	18
totaal	30	20	10	60

Waargenomen − Verwacht

	locatie 1	locatie 2	locatie 3	totaal
soort a	3	1	−4	0
soort b	0	−2	2	0
soort c	−2	−1	3	0
soort d	−1	2	−1	0
totaal	0	0	0	0

Chikwadraat bijdrages

	locatie 1	locatie 2	locatie 3	kwadratsom
soort a	0,112	0,046	−0,258	0,081
soort b	0	−0,129	0,183	0,050
soort c	−0,149	−0,091	0,387	0,180
soort d	−0,043	0,105	−0,075	0,018
kwadratsom	0,037	0,038	0,255	

(Chi-kwadraat bijdrage: $D_r^{-.95}$ (waargenomen − verwacht) $D_k^{-.95}$)

b. * soort c wijkt het sterkst af van de vier soorten: de kwadratsom van de X^2-bijdrages is het grootst. Tevens is al in de waargenomen waarden te zien dat de frequentie van soort c toeneemt van links naar rechts, terwijl de kolomtotalen juist afnemen.

* locatie 3 wijkt het sterkst af van de drie locaties: de kwadratsom van de X^2-bijdrages is het grootst. In de waargenomen waarden is te zien dat de frequentie voor locatie 3 toe- en afneemt, waar de rijtotalen juist af- en toenemen.

c. * locatie 1 lijkt het sterkst op de marginaal: laagste kwadraatsom en het verloop van de waargenomen frequenties is ~~gelijk~~ gelijk aan dat van de rijtotalen

* soort d lijkt het sterkst op de marginaal: laagste kwadraatsom en het verloop van de waargenomen frequenties lijkt op dat van de kolomtotalen.

d. Omdat er sprake is van kwalitatieve gegevens ~~en~~ (de frequentie in voorkomen van een soort op een locatie is weergegeven) is het het beste om een correspondentie-analyse (CA) uit te voeren.

~~De rijen en kolommen worden gesorteerd op de kwadraatsom van de~~

De rijen en kolommen worden gesorteerd op de loadingen voor een bepaalde dimensie. Deze loadingen worden berekend op basis van de X^2-bijdrages.

7) a. Het doel van een hoofdcomponenten-analyse (PCA) is:
- * data reductie
- * data-exploratie (in dit geval: is er een patroon zichtbaar?)
- * het verkrijgen van data-redundancy
- * het samenvatten van de data

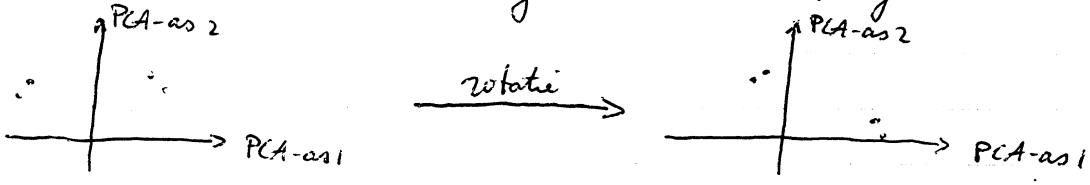
Omdat er per pixel van het gebied 70 ~~en~~ variabelen zijn (~~het~~ de dataset is 70-dimensionaal) lijkt het handig om een data-reductie te bewerkstelligen. Het zou handig zijn als de dataset samengevat kan worden in minder dimensies. Het is dan makkelijker een patroon te ontdekken.

b. De data-matrix ziet er als volgt uit:

~~pixelnr.~~	spectrale band (genummerd)			
pixelnr.	1	2	...	70
1	$DN_{1,1}$	$DN_{1,2}$	...	$DN_{1,70}$
2	$DN_{2,1}$	$DN_{2,2}$	...	$DN_{2,70}$
⋮	⋮	⋮		⋮
n	$DN_{n,1}$	$DN_{n,2}$	...	$DN_{n,70}$

Hierin geldt:
n = aantal pixels per beeld
$DN_{i,j}$ = Digital Number voor pixel i en spectrale band j

c. Er zou een verbetering aangebracht kunnen worden door rotatie van de drie PCA-assen.
Zonder rotatie is er weinig verschil in de loadingen van de ~~en~~ variabelen op de PCA-assen. Het is dan niet duidelijk welke as door welke ~~en~~ variabelen gevormd wordt. Door rotatie komen bepaalde variabelen dichter bij een bepaalde as te liggen (er ontstaat dan meer verschil in loadingen tussen de variabelen).
In een 2-dimensionaal geval kan het volgende aan de hand zijn:



Na rotatie hebben de variabelen een hoge loading voor de ene PCA-as en een lage loading voor de andere (voor een 2-dimensionaal voorbeeld is gekozen omdat dit makkelijker te tekenen is dan een meer-dimensionale situatie).

Een loading van een variabele op een PCA-as geeft de samenhang weer van de variabele met die PCA-as. Een loading die sterk positief of negatief is, geeft een resp. sterk positief of sterk negatief verband tussen de variabele en de PCA-as.

Een PCA-as kan beschreven als: $Y = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$

waarbij geldt: α_i is loading van variabele X_i op de PCA-as

'n score voor een PCA-as is een nieuwe waarde van een object die berekend wordt op basis van de loadingen. De berekening is dan: $Y = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$

waarbij geldt: X_i = waarde voor variabele i

n = totaal aantal variabelen

α_i = loading van variabele i op de PCA-as

Y = score voor de PCA-as

Dit kan samengevat worden als: $Y = X C$

waarbij geldt: X = matrix met de waarden voor de variabelen

C = transponeerde matrix met

op die manier kan voor elke PCA-as een score uitgerekend worden.