Chemometrie, ISBN: 9789491764332
Vierde herziene druk
www.syntaxmedia.nl

Uitwerkingen

In de uitgewerkte voorbeelden worden vanwege de leesbaarheid afgeronde tussenresultaten gepresenteerd. De eindresultaten zijn echter altijd berekend zonder tussentijds afronden.

Hoofdstuk 14

Opmerkingen bij PCR

1. De uitwerkingen voor PCR in dit hoofdstuk zijn opgesplitst in een PCA-deel en een regressiedeel. De PCA is uitgevoerd met Matlab en de regressie in Excel. Deze opsplitsing geeft een beter inzicht in het verloop van de data-analyse. In commerciële software wordt de regressie niet uitgevoerd in Excel maar worden beide stappen geïntegreerd uitgevoerd.
2. Afhankelijk van de gebruikte software kunnen de tekens in de scorematrix ***T*** en de loadingsmatrix ***P*** tegengesteld zijn aan die in deze uitwerkingen. Dat heeft invloed op de score- en loadingplots omdat deze gespiegeld kunnen zijn ten opzichte van de in deze uitwerkingen gepresenteerde plots. Het heeft geen invloed op de conclusies van de data-analyse.

Antwoord 14.1

In de tabel valt op dat Rusland afwijkt van de andere landen met de hoogste consumptie van sterke drank, de laagste consumptie van bier, de laagste levensverwachting en het hoogste aantal hartziekten.

Voordat de data kunnen worden geanalyseerd moeten deze worden voorbewerkt. De aard van de variabelen verschilt sterk en ze worden uitgedrukt in verschillende eenheden. In zo'n geval kan het beste *autoschaling* worden toegepast, zie bladzijde 284.

De ***X***-matrix na autoschaling is gelijk aan de ***Z***-matrix:

Nr.	zx1	zx2	zx3	zx4	zx5
1	0,8213	1,7054	-0,7282	0,6536	-0,8371
2	-0,9308	1,4847	-1,1161	0,6536	-0,5394
3	-0,0548	1,0033	-0,0843	0,6536	-0,4284
4	-0,6023	-0,2122	0,8751	0,6536	0,1726
5	-0,2738	-0,3526	0,8208	0,3424	0,4136
6	0,2738	-0,4850	0,5053	0,0311	0,1997
7	2,2449	-0,7337	-1,3229	-2,1476	1,9828
8	-0,8213	-0,7739	1,8551	-0,9026	1,1716
9	0,3833	-0,8019	-0,3429	0,9649	-1,0754
10	-1,0403	-0,8340	-0,4618	-0,9026	-1,0600
sdev	1,0000	1,0000	1,0000	1,0000	1,0000
var	1,0000	1,0000	1,0000	1,0000	1,0000
som var					5,0000

Merk op dat de standaarddeviaties en de varianties van de variabelen na autoschaling gelijk zijn aan 1. De som van de varianties van de variabelen na autoschaling is gelijk aan het aantal variabelen.

Na autoschaling worden de principale componenten berekend op basis van de correlatiematrix ***R***, zie bladzijde 322 vergelijking 14.6.

De correlatiematrix ***R*** is:

	consumptie sterke drank	consumptie wijn	consumptie bier	levens-verwachting	hartziekten
consumptie sterke drank	1,0000	-0,0440	-0,4792	-0,3730	0,4283
consumptie wijn	-0,0440	1,0000	-0,3898	0,5244	-0,3947
consumptie bier	-0,4792	-0,3898	1,0000	0,0957	0,2497
levensverwachting	-0,3730	0,5244	0,0957	1,0000	-0,7017
hartziekten	0,4283	-0,3947	0,2497	-0,7017	1,0000

Voor een dataset met 5 variabelen kunnen maximaal 5 principale componenten worden berekend. De eigenwaarden van de correlatiematrix, cumulatieve eigenwaarden en percentages van de verklaarde varianties van de principale componenten zijn vermeld in de volgende tabel. Daaruit blijkt dat de eerste twee principale componenten samen 78,1 % van de totale variantie in de dataset verklaren. Hier wordt de regel toegepast die staat op bladzijde 326 dat, na autoschaling van ***X***, de optimale modelcomplexiteit bij PCA gelijk is aan het aantal principale componenten waarvoor geldt $\lambda > 1$.

nummer *j* PC	λ_j	$\sum \lambda_j$	**percentage verklaarde variantie**
1	2,30	2,30	46,0
2	1,61	3,91	78,1
3	0,58	4,49	89,8
4	0,42	4,91	98,3
5	0,09	5,00	100,0

De scorematrix ***T*** voor twee principale componenten is:

nr.	PC1	PC2
1	-1,3953	-1,6192
2	-1,7596	-0,8084
3	-1,1018	-0,3717
4	-0,3317	1,1203
5	0,1619	0,9310
6	0,4452	0,4054
7	3,4085	-2,0556
8	1,4032	2,0760
9	-0,7224	-0,1260
10	-0,1080	0,4480
sdev	1,5170	1,2672
var	2,3014	1,6057
som var		3,9071
eigenwaarde	2,3014	1,6057

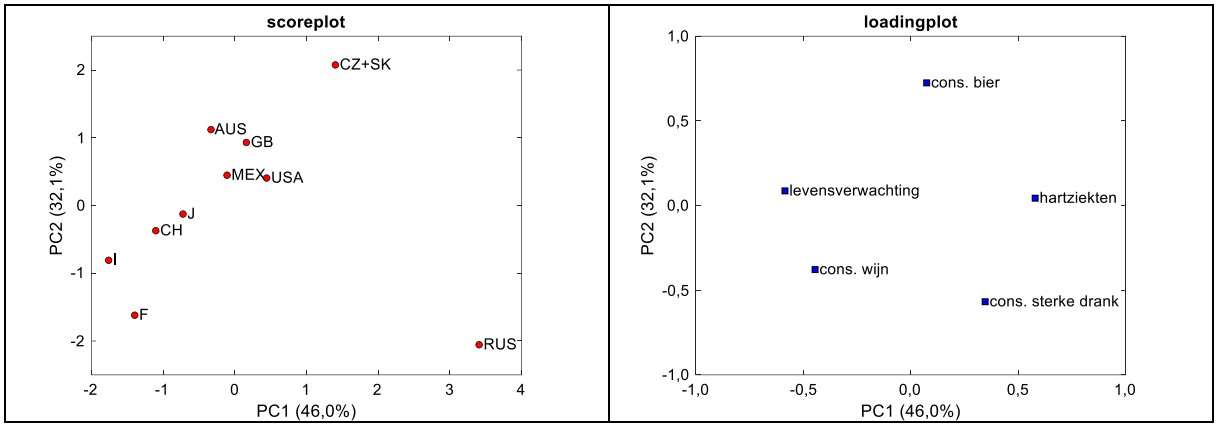
Uit voorgaande tabel met gegevens over de scorematrix ***T*** blijkt dat de varianties van de kolommen in ***T*** gelijk zijn aan de eigenwaarden van de overeenkomstige principale componenten.

De som van de varianties van de eerste twee principale componenten is 3,9071. De som van de varianties van de ***Z***-matrix is 5. De eerste twee principale componenten verklaren daarom $3{,}9071/5 \cdot 100 = 78{,}1$ % van de totale variantie in de dataset.

De loadingsmatrix ***P*** is:

	PC1	PC2
consumptie sterke drank	0,3459	-0,5681
consumptie wijn	-0,4450	-0,3784
consumptie bier	0,0740	0,7244
levensverwachting	-0,5850	0,0864
hartziekten	0,5785	0,0434

Op basis van de scorematrix ***T*** en de loadingsmatrix ***P*** kan respectievelijk een scoreplot en een loadingplot worden gemaakt. Zie de volgende afbeeldingen.



Een *scoreplot* geeft de relaties weer tussen de objecten, in dit geval de landen. In de scoreplot ligt Rusland duidelijk gescheiden van de andere landen. In de andere landen is een trend waarneembaar die begint bij Frankrijk (linksonder) en eindigt bij Tsjechië en Slowakije (rechtsboven).

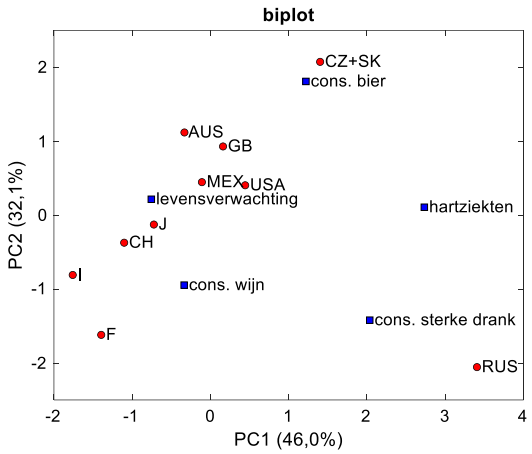
Een *loadingplot* geeft de relaties weer tussen de variabelen en de principale componenten. In de loadingplot zijn de variabelen goed gespreid wat betekent dat ze verschillend zijn. De variabelen ‘levensverwachting’ en ‘hartziekten’ hebben een lage loading op PC2 en respectievelijk een sterk negatieve en sterk positieve loading op PC1. Zij liggen dicht langs de PC1-as, maar in tegengestelde richting. Dat betekent dat ze negatief zijn gecorreleerd. Dat blijkt ook uit de correlatiecoëfficiënt die -0,7017 is, zie de correlatiematrix ***R***. Het is ook logisch dat hoge waarden voor ‘levensverwachting’ samen gaan met lage waarden voor ‘hartziekten’.

PC1 kan worden geïnterpreteerd als een abstracte factor die gerelateerd is aan gezondheid, waarbij de gezondheid toeneemt van rechts naar links. Gezondheid wordt bepaald door verschillende factoren. In deze beperkte dataset wordt deze bepaald door de twee variabelen ‘levensverwachting’ en ‘hartziekten’ en beschreven door PC1. Rechts is het aantal hartziekten hoog en links is de levensverwachting hoog.
De variabele ‘consumptie van bier’ heeft een sterk positieve loading op PC2, en de variabelen ‘consumptie van wijn’ en ‘consumptie van sterke drank’ hebben een negatieve loading op PC2, waarbij de loading van de ‘consumptie van sterke drank’ sterker negatief is dan van de ‘consumptie van wijn’. PC2 kan worden geïnterpreteerd als een abstracte factor die gerelateerd is de sterkte van de alcoholische drank, waarbij de sterkte toeneemt van boven naar beneden.

De variabelen ‘consumptie van sterke drank’ en ‘hartziekten’ hebben beide een positieve loading op PC1. Dat is een indicatie van een positieve correlatie tussen de ‘consumptie van sterke drank en ‘hartziekten’. De correlatiecoëfficiënt tussen beide variabelen is 0,4283, zie de correlatiematrix ***R***. De variabelen ‘consumptie van wijn’ en ‘levensverwachting’ hebben beide een negatieve loading op PC1. Dat is een indicatie van een positieve correlatie tussen de ‘consumptie van wijn’ en ‘levensverwachting’. De correlatiecoëfficiënt tussen beide variabelen is 0,5244, zie de correlatiematrix ***R***.

Uit biplot hierna blijkt welke variabelen de verschillen veroorzaken tussen de landen in de scoreplot. Rusland ligt dicht bij de variabelen ‘consumptie van sterke drank’ en ‘hartziekten’ en ver van de variabelen ‘levensverwachting’ en ‘consumptie van bier’. Dit verklaart de positie van Rusland ten opzichte van de andere landen omdat Rusland de hoogste consumptie heeft van sterke drank en het hoogste aantal hartziekten, én de laagste consumptie van bier en de laagste levensverwachting.

De trend die begint bij Frankrijk (linksonder) en eindigt bij Tsjechië en Slowakije (rechtsboven) correspondeert met een verschuiving in de richting van een hogere bier consumptie en meer hartziekten, terwijl de tegelijkertijd de consumptie van wijn afneemt en de levensverwachting lager wordt. Frankrijk combineert een hoge consumptie van wijn met een hoge levensverwachting. Zie de tabel in de opgave.



Antwoord 14.2

PCA is uitgevoerd na autoschaling van de LSER-factoren in de ***X*** matrix en niet-geschaalde log k_w waarden in de ***y*** vector. Daarna is PCR uitgevoerd met een klassiek kalibratiemodel.

De $\boldsymbol{X}_{\mathbf{kal}}$ - en $\boldsymbol{X}_{\mathbf{test}}$ -matrix na autoschaling:

	Kalibratieset				
nr.	zx1	zx2	zx3	zx4	zx5
1	-0,8232	-1,2942	-0,7872	-0,6082	1,6794
2	-0,7117	-1,5538	-0,7872	-0,3887	3,0341
3	0,8458	1,6395	-0,7872	0,2069	0,0877
4	-1,3372	-0,3336	1,3885	-0,7963	-0,2197
5	0,2575	1,6395	1,6001	-0,1693	-0,3446
6	1,6199	0,5751	1,2677	-0,4514	-0,0147
7	-0,7644	-1,2423	-0,7872	-0,6395	-1,0299
8	-0,4299	-0,9048	-0,7872	-0,8590	-0,6360
9	-1,4053	-0,3596	-0,7872	0,6771	-0,5656
10	-0,1606	-0,2817	1,0259	-0,1379	-0,8410
11	0,5021	-0,3855	-0,7872	-1,0784	0,9076
12	1,0006	0,6529	0,8446	-0,0125	-0,4247
13	1,9915	1,5616	-0,7872	3,1538	1,0421
14	0,4123	0,1856	1,0864	0,6144	0,2191
15	-1,1329	1,3020	-0,7872	1,8998	-0,6969
16	1,4961	-0,2038	-0,7872	-0,4514	0,1518
17	0,1801	0,2116	1,2374	-0,4514	-0,4471
18	-0,7922	-1,2423	-0,7872	-0,6395	-1,0299
19	-0,3556	0,2895	-0,7872	-0,0439	-0,5335
20	-0,3928	-0,2557	0,9957	0,1756	-0,3382
	Testset				
1	0,5052	0,2635	1,7209	-1,0784	-0,0564
2	-0,4609	-0,6452	-0,7872	-0,1693	-0,3894
3	0,4123	1,3020	0,6935	1,0220	-0,2069
4	3,4128	0,8087	-0,7872	-0,5141	1,0934
5	-0,7799	-1,3981	-0,7872	-0,4514	1,6794

Na autoschaling worden de principale componenten berekend op basis van de correlatiematrix $\boldsymbol{R}$, zie bladzijde 322 vergelijking 14.6.

De correlatiematrix $\boldsymbol{R}$ is:

	L1	L2	L3	L4	L5
L1	1,0000	0,5333	0,1480	0,2471	0,1566
L2	0,5333	1,0000	0,2760	0,6042	-0,1963
L3	0,1480	0,2760	1,0000	-0,1466	-0,2445
L4	0,2471	0,6042	-0,1466	1,0000	0,0461
L5	0,1566	-0,1963	-0,2445	0,0461	1,0000

De correlaties tussen de LSER-factoren is relatief gering. Dat is gunstig omdat dan elke variabele specifieke informatie bevat.

Voor een dataset met 5 variabelen kunnen maximaal 5 principale componenten worden berekend. De eigenwaarden van de correlatiematrix, cumulatieve eigenwaarden en percentages van de verklaarde varianties van de principale componenten zijn vermeld in de volgende tabel. Daaruit blijkt dat er vier principale componenten nodig zijn om meer dan 95 % van de totale variantie in de dataset te verklaren. Dus $A = 4$.

nummer *j* PC	λ_j	$\sum \lambda_j$	**percentage verklaarde variantie**
1	1,9775	1,9775	39,55
2	1,3255	3,3030	66,06
3	0,9534	4,2564	85,13
4	0,5409	4,7974	95,95
5	0,2026	5,0000	100,00

N.B. De optimale modelcomplexiteit bij PCA kan op verschillende manieren worden bepaald (zie bladzijde 326):
(*i*) op basis van een voldoende hoog percentage verklaarde variantie;
(*ii*) na autoschaling van ***X*** door alleen de principale componenten te behouden met een eigenwaarde groter dan één, omdat deze een meer dan gemiddelde hoeveelheid variantie verklaren.

In dit geval was de opgave om een aantal principale componenten voor het model te kiezen waarmee tenminste 95% van de aanwezige variantie kan worden verklaard. Dat correspondeert volgens voorgaande tabel met 4 principale componenten. Bij toepassing van de regel $\lambda > 1$ zou de modelcomplexiteit 2 moeten worden en zou slechts 66,02% van de totale variantie worden verklaard. In de praktijk wordt de regel toegepast die het beste bij de probleemstelling en de met de toegepaste regel gevonden oplossing past.

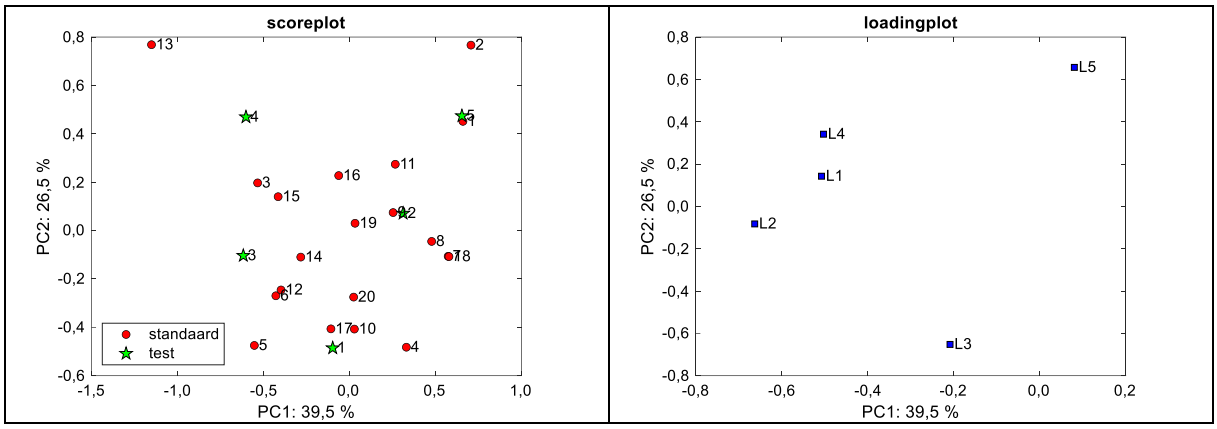
De scorematrix ***T*** voor de kalibratie- en testset voor vier principale componenten is:

nr.	PC1	PC2	PC3	PC4
	Kalibratieset			
1	0,6595	0,4515	0,1847	-0,1633
2	0,7062	0,7666	0,4155	-0,3616
3	-0,5339	0,1970	-0,0590	0,2385
4	0,3307	-0,4822	0,0733	-0,3636
5	-0,5528	-0,4749	0,1707	-0,1882
6	-0,4279	-0,2696	0,4956	0,1306
7	0,5736	-0,1064	-0,2708	0,2419
8	0,4777	-0,0448	-0,1284	0,2710
9	0,2537	0,0744	-0,5332	-0,1232
10	0,0284	-0,4073	0,0096	-0,0705
11	0,2663	0,2741	0,3047	0,2444
12	-0,3977	-0,2455	0,1906	0,0922
13	-1,1506	0,7684	-0,1406	-0,0510
14	-0,2831	-0,1097	0,1644	-0,2268
15	-0,4146	0,1400	-0,7649	-0,2290
16	-0,0626	0,2273	0,2389	0,4600
17	-0,1078	-0,4063	0,1911	-0,0683
18	0,5782	-0,1076	-0,2754	0,2366
19	0,0324	0,0301	-0,2718	0,1515
20	0,0242	-0,2751	0,0050	-0,2211
	Testset			
1	-0,0974	-0,4855	0,4749	-0,0716
2	0,3120	0,0710	-0,2055	0,1396
3	-0,6171	-0,1035	-0,0752	-0,1619
4	-0,6016	0,4694	0,6883	0,6797
5	0,6537	0,4739	0,1729	-0,1733

De loadingsmatrix ***P*** is:

	PC1	PC2	PC3	PC4
L1	-0,5076	0,1424	0,5185	0,5843
L2	-0,6629	-0,0828	-0,1116	-0,0360
L3	-0,2087	-0,6522	0,4385	-0,5382
L4	-0,5030	0,3410	-0,4698	-0,3926
L5	0,0803	0,6566	0,5529	-0,4620

Op basis van de scorematrix ***T*** en de loadingsmatrix ***P*** kunnen respectievelijk scoreplots en loadingplots worden gemaakt. In de volgende afbeeldingen zijn deze gegeven voor de eerste twee principale componenten. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



Met behulp van de gereduceerde scorematrix $\boldsymbol{T}_{\mathbf{kal}}$ met vier principale componenten kan voor de kalibratieset een invers PCR-model $\boldsymbol{y}_{\mathbf{kal}} = \boldsymbol{T}_{\mathbf{kal}}\boldsymbol{b}$ worden ontwikkeld met de log k_w -waarden in de $\boldsymbol{y}_{\mathbf{kal}}$ -vector. Het resultaat hiervan is een vector met geschatte regressiecoëfficiënten in $\hat{\boldsymbol{b}}$, zie de volgende regressie-uitvoer van Excel.
Hiermee kunnen geschatte waarden voor $\hat{\boldsymbol{y}}_{\mathbf{kal}}$ en $\hat{\boldsymbol{y}}_{\mathbf{test}}$ worden berekend met respectievelijk $\hat{\boldsymbol{y}}_{\mathbf{kal}} = \boldsymbol{T}_{\mathbf{kal}}\hat{\boldsymbol{b}}$ en $\hat{\boldsymbol{y}}_{\mathbf{test}} = \boldsymbol{T}_{\mathbf{test}}\hat{\boldsymbol{b}}$, zie volgende tabel. De berekeningen kunnen met de bekende scorematrices voor de kalibratieset en test set zonder toepassing van matrixrekening in Excel worden uitgevoerd.

gereduceerde scorematrix $\boldsymbol{T}$								
	Log $k_w = y_i$	PC1	PC2	PC3	PC4	$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	Kalibratieset							
S1	5,4892	0,6595	0,4515	0,1847	-0,1633	4,9567	-0,5325	0,283587
S2	6,0714	0,7062	0,7666	0,4155	-0,3616	6,3229	0,2515	0,063242
S3	1,5692	-0,5339	0,1970	-0,0590	0,2385	1,4293	-0,1399	0,019563
S4	2,5650	0,3307	-0,4822	0,0733	-0,3636	2,2859	-0,2791	0,077895
S5	1,1786	-0,5528	-0,4749	0,1707	-0,1882	0,8406	-0,3380	0,114247
S6	2,3819	-0,4279	-0,2696	0,4956	0,1306	2,3095	-0,0724	0,005245
S7	2,0052	0,5736	-0,1064	-0,2708	0,2419	2,3909	0,3857	0,148784
S8	2,6725	0,4777	-0,0448	-0,1284	0,2710	2,7134	0,0409	0,001670
S9	1,0396	0,2537	0,0744	-0,5332	-0,1232	1,4621	0,4225	0,178486
S10	1,0938	0,0284	-0,4073	0,0096	-0,0705	1,6138	0,5200	0,270416
S11	4,5311	0,2663	0,2741	0,3047	0,2444	4,1198	-0,4113	0,169165
S12	1,6175	-0,3977	-0,2455	0,1906	0,0922	1,5737	-0,0438	0,001921
S13	0,8923	-1,1506	0,7684	-0,1406	-0,0510	1,1585	0,2662	0,070837
S14	1,5758	-0,2831	-0,1097	0,1644	-0,2268	2,0403	0,4645	0,215722
S15	0,3015	-0,4146	0,1400	-0,7649	-0,2290	-0,3189	-0,6204	0,384876
S16	3,0787	-0,0626	0,2273	0,2389	0,4600	3,1753	0,0966	0,009329
S17	1,9694	-0,1078	-0,4063	0,1911	-0,0683	1,8579	-0,1115	0,012435
S18	2,6125	0,5782	-0,1076	-0,2754	0,2366	2,3857	-0,2268	0,051422
S19	1,5680	0,0324	0,0301	-0,2718	0,1515	1,6258	0,0578	0,003345
S20	1,6027	0,0242	-0,2751	0,0050	-0,2211	1,8728	0,2701	0,072961
	Testset							
T1	2,9099	-0,0974	-0,4855	0,4749	-0,0716	2,5164	-0,3935	0,154816
T2	2,0436	0,3120	0,0710	-0,2055	0,1396	2,4263	0,3827	0,146471
T3	0,8308	-0,6171	-0,1035	-0,0752	-0,1619	0,7331	-0,0977	0,009542
T4	4,0185	-0,6016	0,4694	0,6883	0,6797	3,8048	-0,2137	0,045650
T5	4,8854	0,6537	0,4739	0,1729	-0,1733	4,9572	0,0718	0,005151

Gegevens voor de regressie	
Meervoudige correlatiecoëfficiënt R	0,9748
R-kwadraat	0,9502
Aangepaste kleinste kwadraat	0,9370
Standaardfout	0,3790
Waarnemingen	20

Variantieanalyse

	Vrijheidsgraden	Kwadratensom	Gemiddelde kwadraten	F	Significantie F
Regressie	4	41,1544	10,2886	71,61	1,37E-09
Storing	15	2,1551	0,1437		
Totaal	19	43,3096			

	Coëfficiënten	Standaardfout	T-statistische gegevens	P-waarde	Laagste 95%	Hoogste 95%
Snijpunt	2,2908	0,0848	27,0278	3,87E-14	2,1101	2,4715
PC1	1,9190	0,1769	10,8463	1,70E-08	1,5419	2,2962
PC2	1,8951	0,2352	8,0585	7,87E-07	1,3938	2,3963
PC3	2,7767	0,2774	10,0085	4,94E-08	2,1854	3,3681
PC4	-0,1946	0,3669	-0,5303	6,04E-01	-0,9766	0,5874

De regressiecoëfficiënten zijn af te lezen in de Excel-uitvoer onder de kop 'Coëfficiënten' en de significantie ervan onder de kop 'P-waarde'. Uit voorgaande Exceluitvoer blijkt dat $R^2_{\text{kal}} = 0{,}9502$, $RMSEC = s_r = 0{,}3790$. De geschatte log k_w –waarden ($\hat{y}_i$) in de testmonsters zijn 2,5164; 2,4263; 0,7331; 3,8048; 4,9572, zie twee tabellen terug.

De correlatiecoëfficiënten voor de geschatte log k_w en de juiste log k_w kunnen afzonderlijk worden berekend voor de kalibratieset en testset. De kwadraten van deze correlatiecoëfficiënten zijn: $R^2_{\text{kal}} = 0{,}9502$ en $R^2_{\text{test}} = 0{,}9659$.

Voor de kalibratieset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{20} (\hat{y}_i - y_i)^2 = 2{,}1551$$

RMSEC kan worden berekend met (12.14) met p = 5:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}} (\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{2{,}1551}{20-5}} = 0{,}3790$$

Voor de testset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{5} (\hat{y}_i - y_i)^2 = 0{,}3616$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}} (\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{0{,}3616}{5}} = 0{,}2689$$

Voor de RMSEP van het SMLR-model in opgave 13.2 is gevonden 0,1630. Dit is lager dan voor het PCR-model in deze opgave. De predictie van het SMLR-model in opgave 13.2 is beter dan die van het PCR-model.

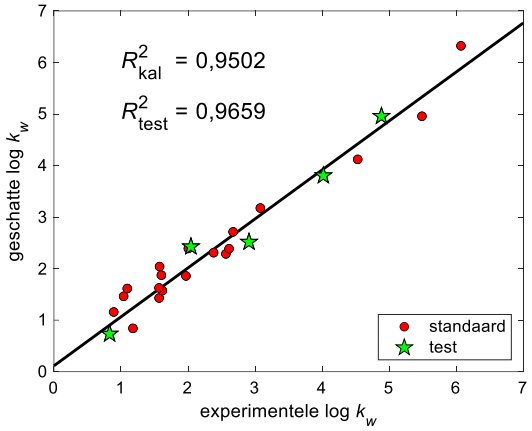
Er kan ook worden getest of de voorspelfout (RMSEP) van het SMLR-model significant beter (lager) is dan van het PCR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten. De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout:

$$F_{PCR/SMLR} = \frac{RMSEP^2_{PCR}}{RMSEP^2_{SMLR}} = \frac{0{,}2689^2}{0{,}1630^2} = 2{,}72$$

$F_{\text{krit}} = F_{(0{,}05;5;5)} = 5{,}05$ (zie tabel 4 van Bijlage 1)

Omdat $F_{\text{PCR/SMLR}} < F_{\text{krit}}$ is er geen significant verschil tussen de varianties van de voorspelfout.

De predictie van het SMLR-model in opgave 13.2 is dus *niet* significant beter dan die van het PCR-model.



Opgave 14.3

PCA is uitgevoerd na centrering van de variabelen in de ***X*** matrix en met niet-geschaalde concentraties in de ***y*** vector. Daarna is PCR uitgevoerd met een invers kalibratiemodel.

Bij PCR wordt alle aanwezige spectrale informatie benut. Na centrering worden de principale componenten berekend op basis van de variantie-covariantiematrix ***S***, zie bladzijde 322 vergelijking 14.5.

Voor een dataset met 13 variabelen kunnen maximaal 13 principale componenten worden berekend. De variantie van elke principale component (= variantie in de overeenkomstige scorevector), de cumulatieve varianties van de principale componenten en de percentages van de verklaarde varianties voor de eerste drie principale componenten zijn vermeld in de volgende tabel. Daaruit blijkt dat er slechts één principale component nodig is om 100 % van de totale variantie in de dataset te verklaren (A = 1). Er is maar één principale component nodig omdat er ook maar één chemische component aanwezig is in de zuivere kobaltoplossingen.

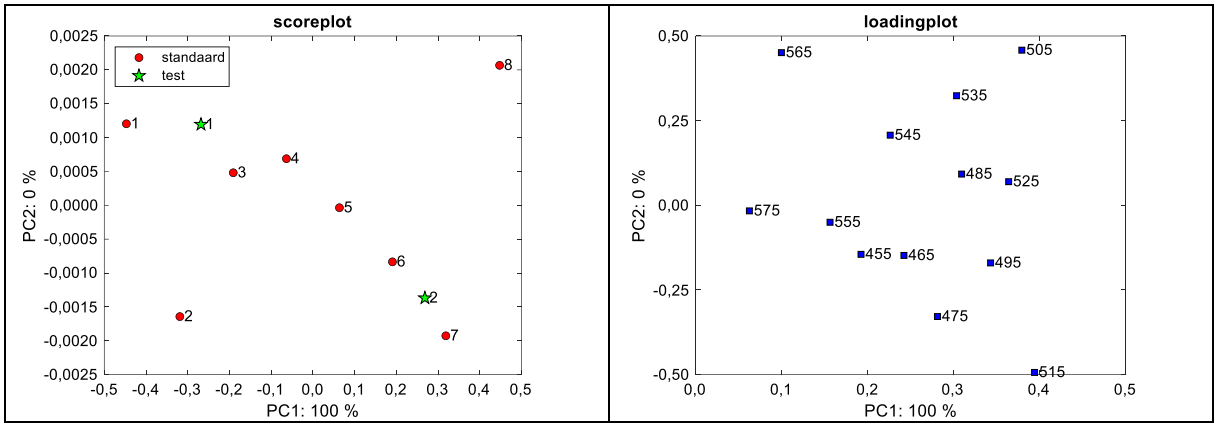
nummer *j* PC	λ_j	$\sum \lambda_j$	percentage verklaarde variantie
1	0,0977	0,0977	100,00
2	0,0000	0,0977	100,00
3	0,0000	0,0977	100,00

De (eendimensionale) scorematrix ***T*** voor de kalibratie- en de testset is gegeven in de volgende tabel.

De (eendimensionale) getransponeerde loadingsmatrix $\boldsymbol{P}$ voor PC1 is:

[0,1921 0,2420 0,2810 0,3091 0,3428 0,3791 0,3939 0,3638 0,3032 0,2261
0,1561 0,0996 0,0624]

Op basis van de score- en loadingsmatrix voor de eerste twee principale componenten kunnen respectievelijk de volgende score- en loadingplot worden gemaakt. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



Met behulp van de gereduceerde scorematrix $\boldsymbol{T}_{\mathbf{kal}}$ met één principale component kan voor de kalibratieset een invers PCR-model $\boldsymbol{y}_{\mathbf{kal}} = \boldsymbol{T}_{\mathbf{kal}}\boldsymbol{b}$ worden ontwikkeld met de kobaltconcentraties in de $\boldsymbol{y}_{\mathbf{kal}}$ -vector. Het resultaat hiervan is een vector met geschatte regressiecoëfficiënten in $\hat{\boldsymbol{b}}$, zie de volgende regressie-uitvoer van Excel.
Hiermee kunnen geschatte waarden voor $\hat{\boldsymbol{y}}_{\mathbf{kal}}$ en $\hat{\boldsymbol{y}}_{\mathbf{test}}$ worden berekend met respectievelijk $\hat{\boldsymbol{y}}_{\mathbf{kal}} = \boldsymbol{T}_{\mathbf{kal}}\hat{\boldsymbol{b}}$ en $\hat{\boldsymbol{y}}_{\mathbf{test}} = \boldsymbol{T}_{\mathbf{test}}\hat{\boldsymbol{b}}$, zie volgende tabel. De berekeningen kunnen met de bekende scorematrices voor de kalibratieset en test set met Excel worden uitgevoerd.

	Co-concentratie	***T*** voor PC1		$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset					
1	0,010	-0,447059		0,0100	-0,000027	7,55E-10
2	0,020	-0,318913		0,0200	0,000013	1,65E-10
3	0,030	-0,190875		0,0300	0,000045	2,00E-09
4	0,040	-0,063480		0,0400	0,000026	6,93E-10
5	0,050	0,063595		0,0500	-0,000017	2,98E-10
6	0,060	0,190663		0,0599	-0,000061	3,76E-09
7	0,070	0,318592		0,0700	-0,000038	1,45E-09
8	0,080	0,447477		0,0801	0,000060	3,63E-09
	testset					
1	0,024	-0,268212		0,0240	-0,000015	2,15E-10
2	0,066	0,268445		0,0660	0,000033	1,08E-09

Gegevens voor de regressie	
Meervoudige correlatiecoëfficiënt R	1,0000
R-kwadraat	1,0000
Aangepaste kleinste kwadraat	1,0000
Standaardfout	4,61E-05
Waarnemingen	8

Variantieanalyse

	Vrijheidsgraden	Kwadratensom	Gemiddelde kwadraten	F	Significantie F
Regressie	1	0,0042	0,0042	1976460	8,74E-18
Storing	6	1,28E-08	2,125E-09		
Totaal	7	0,0042			

	Coëfficiënten	Standaardfout	T-statistische gegevens	P-waarde	Laagste 95%	Hoogste 95%
Snijpunt	0,0450	1,63E-05	2761,07	1,52E-19	0,0450	0,0450
PC1	0,0784	5,57E-05	1405,87	8,74E-18	0,0782	0,0785

De regressiecoëfficiënten zijn af te lezen in de Excel-uitvoer onder de kop 'Coëfficiënten' en de significantie ervan onder de kop 'P-waarde'. Beide regressiecoëfficiënten zijn significant omdat $p < 0{,}05$. Uit de Exceluitvoer blijkt dat $R^2_{\text{kal}} = 1{,}0000$, $RMSEC = s_r = 4{,}61 \cdot 10^{-5}$. De geschatte kobaltconcentraties in de testmonsters zijn 0,0240 en 0,0660, zie twee tabellen terug.

De correlatiecoëfficiënt voor de geschatte en juiste kobaltconcentraties voor de kalibratieset is $R^2_{\text{kal}} = 1{,}0000$. Voor de testset wordt voor de geschatte $\hat{y}$ en de juiste y geen correlatiecoëfficiënt berekend omdat er slechts twee testmonsters zijn.

Voor de kalibratieset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{8} (\hat{y}_i - y_i)^2 = 1{,}275 \cdot 10^{-8}$$

RMSEC kan worden berekend met (12.14) met $p = 2$:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}} (\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}275 \cdot 10^{-8}}{8-2}} = 4{,}61 \cdot 10^{-5}$$

Voor de testset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{2} (\hat{y}_i - y_i)^2 = 1{,}299 \cdot 10^{-9}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}} (\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{1{,}299 \cdot 10^{-9}}{2}} = 2{,}55 \cdot 10^{-5}$$

RMSEP van het SMLR-model met b_0-term in opgave 13.3 is $1{,}85 \cdot 10^{-4}$. Dit is hoger dan voor het PCR-model in deze opgave. De predictie van het PCR-model in deze opgave is beter dan die van het SMLR-model in opgave 13.3. Dit kan worden verklaard door het feit dat bij PCR de aanwezige spectrale informatie in alle golflengten kan worden benut. Deze informatie is in het PCR-model geconcentreerd in één principale component. Bij SMLR kan alleen de spectrale informatie van de twee geselecteerde golflengten worden benut.

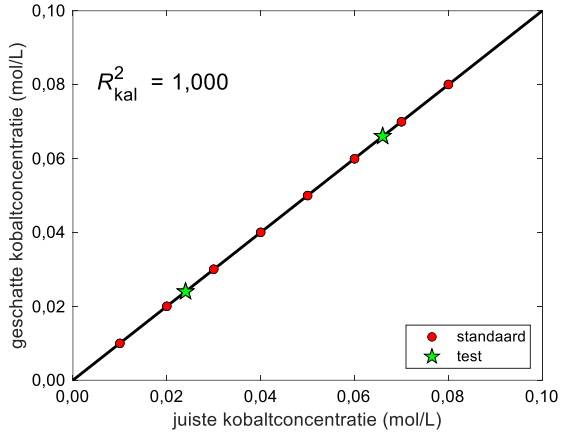
Er kan ook worden getest of de voorspelfout (RMSEP) van het PCR-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten. De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout:

$$F_{SMLR/PCR} = \frac{RMSEP^2_{SMLR}}{RMSEP^2_{PCR}} = \frac{(1{,}85 \cdot 10^{-4})^2}{(2{,}55 \cdot 10^{-5})^2} = 52{,}78$$

$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Omdat $F_{SMLR/PCR} > F_{krit}$ is er een significant verschil tussen de varianties van de voorspelfout.

De predictie van het PCR-model is dus *significant beter* dan die van het SMLR-model in opgave 13.3.



Antwoord 14.4

PCA is uitgevoerd na centrering van de variabelen in de ***X*** matrix en met niet-geschaalde vetpercentages in de ***y*** vector. Daarna is PCR uitgevoerd met een invers kalibratiemodel.

Na centrering worden de principale componenten berekend op basis van de variantie-covariantiematrix ***S***, zie bladzijde 322 vergelijking 14.5. De eigenwaarden van de covariantiematrix, cumulatieve eigenwaarden en percentages van de verklaarde varianties van de principale componenten zijn vermeld in de volgende tabel.

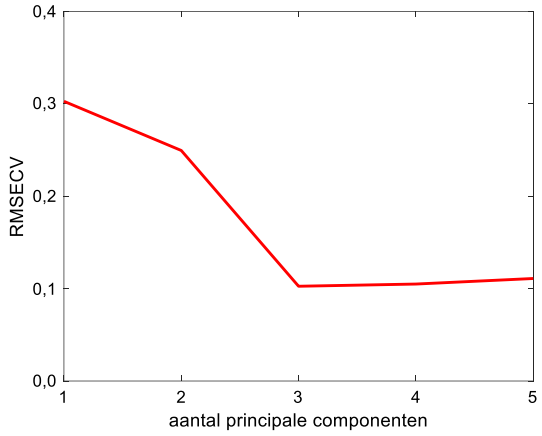
nummer *j* PC	λ_j	$\sum \lambda_j$	percentage verklaarde variantie
1	0,2903	0,2903	99,69
2	0,0005	0,2909	99,87
3	0,0004	0,2912	100,00

N.B. Op bladzijde 326 in het boek staat dat de optimale modelcomplexiteit bij PCA op de volgende manieren kan worden bepaald:
(*i*) op basis van een voldoende hoog percentage verklaarde variantie;
(*ii*) na autoschaling van ***X***, door alleen de principale componenten te behouden met $\lambda > 1$.

Als er een afhankelijke variabele *y* bestaat is er nog een derde optie:
(*iii*) door het aantal componenten te kiezen dat correspondeert met het minimum van de RMSECV-curve bij kruisvalidatie of met het punt waarna de RMSECV nog maar weinig daalt, bijvoorbeeld minder dan 2 %.
Dat geldt zowel voor PCR als voor PLS. Dit is bij PCR toegepast in paragraaf 14.5, afbeelding 14.9 en bij PLS in paragraaf 15.4. Deze optie wordt ook toegepast bij de uitwerking van deze opgave.

De optimale modelcomplexiteit *A* is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal principale componenten in het model. LOO-kruisvalidatie wordt bij de voorbeelden en alle opgaven in het boek toegepast omdat het aantal kalibratiemonsters gering is.

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal principale componenten in het model en hiervan is de volgende grafiek getekend. Het minimum in deze curve ligt bij *A* = 3. Dit is de optimale modelcomplexiteit voor het PCR-model. Uit voorgaande tabel blijkt dat daarbij 100% van de aanwezige variantie wordt verklaard.



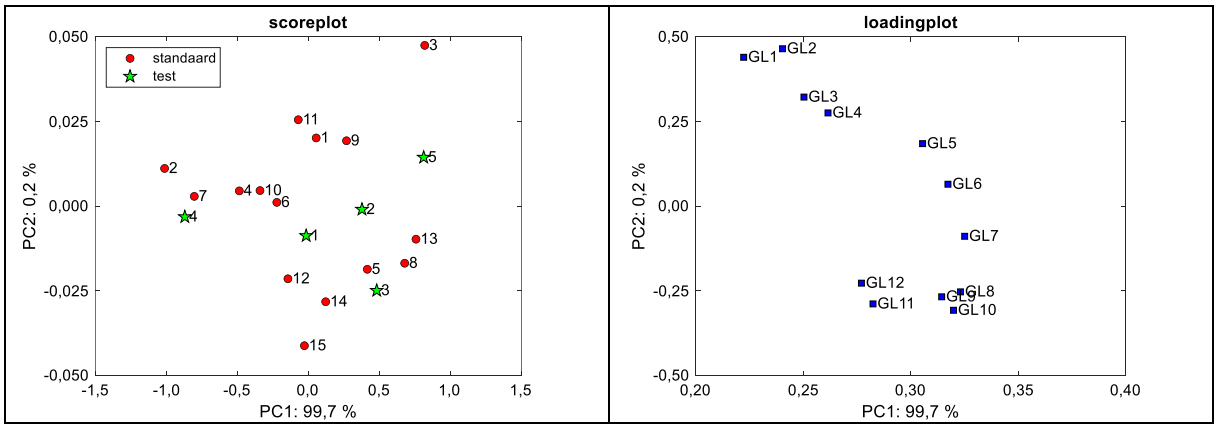
De scorematrix ***T*** voor de kalibratie- en testset voor drie principale componenten is:

Nr.	PC1	PC2	PC3
	kalibratieset		
1	0,05448873	0,02010150	0,00056970
2	-1,01276742	0,01109767	-0,01762596
3	0,81861701	0,04743256	-0,00891411
4	-0,48675145	0,00447361	0,03501660
5	0,41449609	-0,01868196	0,00499806
6	-0,22270578	0,00107759	0,01575449
7	-0,80358686	0,00285303	0,00620355
8	0,67749011	-0,01689145	-0,02773680
9	0,26777639	0,01927317	0,00987294
10	-0,34104292	0,00456460	-0,03062466
11	-0,07152738	0,02547949	0,00586132
12	-0,14447722	-0,02148080	-0,02793523
13	0,75706862	-0,00979414	0,01956013
14	0,12132893	-0,02826466	0,00391312
15	-0,02840686	-0,04124021	0,01108685
	testset		
1	-0,01600809	-0,00878427	-0,01976086
2	0,37697754	-0,00100527	-0,00504046
3	0,48080912	-0,02496406	-0,01142655
4	-0,87027206	-0,00320008	0,00463059
5	0,81139698	0,01435973	-0,00567355

De loadingsmatrix $\boldsymbol{P}$ is:

	PC1	PC2	PC3
GL1	0,2223	0,4392	-0,3012
GL2	0,2404	0,4644	-0,2906
GL3	0,2503	0,3217	-0,1396
GL4	0,2615	0,2749	0,0678
GL5	0,3055	0,1843	0,4303
GL6	0,3173	0,0643	0,4465
GL7	0,3251	-0,0894	0,3516
GL8	0,3231	-0,2534	0,0715
GL9	0,3144	-0,2681	-0,0024
GL10	0,3199	-0,3079	-0,2012
GL11	0,2824	-0,2890	-0,2655
GL12	0,2771	-0,2278	-0,4202

Op basis van de score- en loadingsmatrix voor de eerste twee principale componenten kunnen respectievelijk de volgende score- en loadingplot worden gemaakt. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



Met behulp van de gereduceerde scorematrix $\boldsymbol{T}_{\mathbf{kal}}$ met drie principale componenten kan voor de kalibratieset een invers PCR-model $\boldsymbol{y}_{\mathbf{kal}} = \boldsymbol{T}_{\mathbf{kal}}\boldsymbol{b}$ worden ontwikkeld met de vetpercentages in de $\boldsymbol{y}_{\mathbf{kal}}$ -vector. Het resultaat hiervan is een vector met geschatte regressiecoëfficiënten in $\widehat{\boldsymbol{b}}$, zie de volgende regressie-uitvoer van Excel.
Hiermee kunnen geschatte waarden voor $\widehat{\boldsymbol{y}}_{\mathbf{kal}}$ en $\widehat{\boldsymbol{y}}_{\mathbf{test}}$ worden berekend met respectievelijk $\widehat{\boldsymbol{y}}_{\mathbf{kal}} = \boldsymbol{T}_{\mathbf{kal}}\widehat{\boldsymbol{b}}$ en $\widehat{\boldsymbol{y}}_{\mathbf{test}} = \boldsymbol{T}_{\mathbf{test}}\widehat{\boldsymbol{b}}$, zie volgende tabel. De berekeningen kunnen met de bekende scorematrices voor de kalibratieset en testset met Excel worden uitgevoerd.

Nr.	Vetpercentage y_i	$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset			
1	41,7	41,7169	0,016871	2,85E-04
2	41,5	41,4705	-0,029545	8,73E-04
3	41,7	41,6211	-0,078894	6,22E-03
4	41,7	41,7874	0,087361	7,63E-03
5	42,2	42,2695	0,069472	4,83E-03
6	41,9	41,8734	-0,026557	7,05E-04
7	41,7	41,6645	-0,035491	1,26E-03
8	42,2	42,2749	0,074911	5,61E-03
9	41,7	41,8049	0,104942	1,10E-02
10	41,8	41,7238	-0,076184	5,80E-03
11	41,6	41,6271	0,027059	7,32E-04
12	42,0	42,0810	0,080964	6,56E-03
13	42,4	42,2949	-0,105072	1,10E-02
14	42,3	42,2884	-0,011616	1,35E-04
15	42,5	42,4018	-0,098221	9,65E-03
	testset			
1	41,9	41,9892	0,089160	7,95E-03
2	42,0	42,0424	0,042447	1,80E-03
3	42,3	42,3337	0,033681	1,13E-03
4	41,6	41,7104	0,110407	1,22E-02
5	42,1	41,9976	-0,102355	1,05E-02

Gegevens voor de regressie	
Meervoudige correlatiecoëfficiënt R	0,9740
R-kwadraat	0,9487
Aangepaste kleinste kwadraat	0,9347
Standaardfout	0,0811
Waarnemingen	15

Variantieanalyse

	Vrijheidsgraden	Kwadratensom	Gemiddelde kwadraten	F	Significantie F
Regressie	3	1,3370	0,4457	67,76	2,23E-07
Storing	11	0,0723	0,0066		
Totaal	14	1,4093			

	Coëfficiënten	Standaardfout	T-statistische gegevens	P-waarde	Laagste 95%	Hoogste 95%
Snijpunt	41,9267	0,0209	2002,32	6,06E-32	41,8806	41,9728
PC1	0,2986	0,0402	7,42	1,32E-05	0,2101	0,3871
PC2	-11,2921	0,9340	-12,09	1,08E-07	-13,3477	-9,2364
PC3	1,6153	1,1391	1,42	1,84E-01	-0,8919	4,1224

De regressiecoëfficiënten zijn af te lezen in de Excel-uitvoer onder de kop ‘Coëfficiënten’ en de significantie ervan onder de kop ‘P-waarde’. Alleen de regressiecoëfficiënt voor PC3 is niet significant omdat $p > 0{,}05$. Uit de Exceluitvoer blijkt dat $R^2_{\text{kal}} = 0{,}9487$, $RMSEC = s_r = 0{,}0811$. Voor de geschatte vetpercentages in de testmonsters, zie twee tabellen terug.

De correlatiecoëfficiënten voor het geschatte vetgehalte en het juiste vetgehalte kunnen afzonderlijk worden berekend voor de kalibratieset en testset. De kwadraten van deze correlatiecoëfficiënten zijn: $R^2_{\text{kal}} = 0{,}9487$ en $R^2_{\text{test}} = 0{,}9066$.

Voor de kalibratieset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{15} (\hat{y}_i - y_i)^2 = 7{,}234 \cdot 10^{-2}$$

RMSEC kan worden berekend met (12.14) met p = 4:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}}(\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{7{,}234 \cdot 10^{-2}}{15-4}} = 8{,}11 \cdot 10^{-2}$$

Voor de testset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{2}(\hat{y}_i - y_i)^2 = 3{,}355 \cdot 10^{-2}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}}(\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{3{,}355 \cdot 10^{-2}}{5}} = 8{,}19 \cdot 10^{-2}$$

RMSEP van het SMLR-model in opgave 13.4 is $1{,}20 \cdot 10^{-1}$. Dit is hoger dan voor het PCR-model in deze opgave. De predictie van het PCR-model in deze opgave is beter dan die van het SMLR-model in opgave 13.4. Dit kan worden verklaard door het feit dat bij PCR de aanwezige spectrale informatie in alle golflengten kan worden benut.

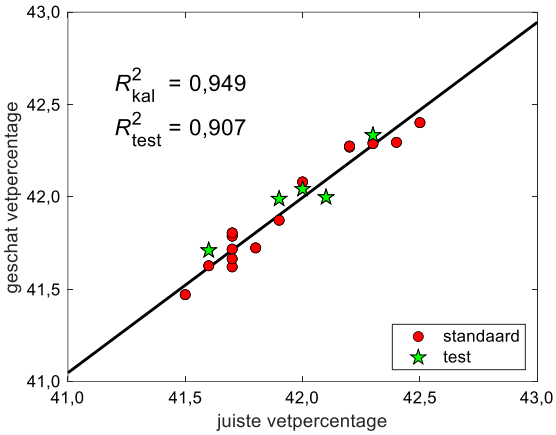
Er kan ook worden getest of de voorspelfout (RMSEP) van het PCR-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten. De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout:

$$F_{SMLR/PCR} = \frac{RMSEP_{SMLR}^2}{RMSEP_{PCR}^2} = \frac{(1{,}20 \cdot 10^{-1})^2}{(8{,}19 \cdot 10^{-2})^2} = 2{,}15$$

$F_{\text{krit}} = F_{(0{,}05;5;5)} = 5{,}05$ (zie tabel 4 van Bijlage 1)

Omdat $F_{\text{SMLR/PCR}} < F_{\text{krit}}$ is er geen significant verschil tussen de varianties van de voorspelfout.

De predictie van het PCR-model is dus *niet* significant beter dan die van het SMLR-model in opgave 13.4.



Antwoord 14.5

PCA is uitgevoerd na centrering van de variabelen in de $\boldsymbol{X}$ matrix en met niet-geschaalde concentraties in de $\boldsymbol{y}$ vector. Daarna is PCR uitgevoerd met een invers kalibratiemodel.

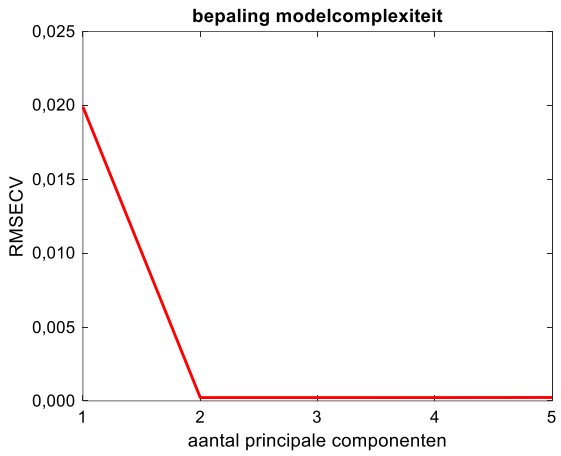
Kobalt

Na centrering worden de principale componenten berekend op basis van de variantie-covariantiematrix $\boldsymbol{S}$, zie bladzijde 322 vergelijking 14.5.

De optimale modelcomplexiteit A is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal principale componenten in het model.
De eigenwaarden van de covariantiematrix, cumulatieve eigenwaarden en percentages van de verklaarde varianties van de principale componenten zijn vermeld in de volgende tabel.

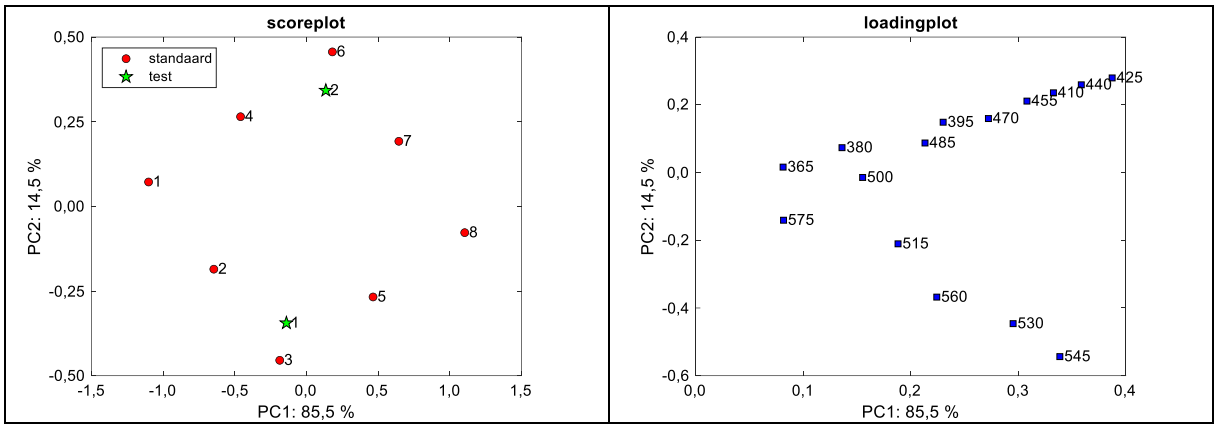
nummer j PC	λ_j	$\sum \lambda_j$	percentage verklaarde variantie
1	0,5369	0,5369	85,46
2	0,0913	0,6282	99,99
3	0,0000	0,6282	99,99
4	0,0000	0,6282	100,00
5	0,0000	0,6282	100,00

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal principale componenten in het model en hiervan is de volgende grafiek getekend. Het minimum in deze curve ligt bij $A = 2$. Dit is de optimale modelcomplexiteit voor het PCR-model. Zie ook de informatie bij opgave 14.4 over de bepaling van de modelcomplexiteit.
Er zijn twee principale componenten nodig omdat het oplossingen betreft van twee zuivere componenten (kobalt en nikkel).



De gereduceerde scorematrix $\boldsymbol{T}$ voor de kalibratie- en testset met twee principale componenten staat in de volgende tabel.

Op basis van de score- en loadingsmatrix voor de eerste twee principale componenten kunnen respectievelijk de volgende score- en loadingplot worden gemaakt. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De scoreplot geeft het experimenteel design van de mengsels in de ruimte van de eerste twee principale componenten. Merk op dat dit design er anders uitziet dan het experimenteel design in de concentratieruimte in afbeelding 13.1c. In standaard 6 is de kobaltconcentratie hoog (0,180) ten opzichte van de nikkelconcentratie (0,100) en in standaard 3 is dit omgekeerd. In de loadingplot corresponderen de golflengten die liggen op de lijn van het midden naar rechtsboven voornamelijk met de nikkelpiek en de golflengten die liggen op de lijn van het midden naar rechtsonder met de kobaltpiek, zie afbeelding 13.1a.

Met behulp van de gereduceerde scorematrix $\boldsymbol{T}_{\mathbf{kal}}$ met twee principale componenten kan voor de kalibratieset een invers PCR-model worden ontwikkeld met de kobaltconcentraties in de $\boldsymbol{y}_{\mathbf{kal}}$ -vector. Deze regressie kan worden uitgevoerd met Excel, zie de volgende regressie-uitvoer van Excel. Met de regressiecoëfficiënten van dit model kunnen geschatte kobaltconcentraties, de bijbehorende residuen en de kwadraten van de residuen voor de kalibratieset en de testset worden berekend, zie volgende tabel.

Nr.	*Kobalt concentratie y_i*	***T*** (PC1)	***T*** (PC2)		$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset						
1	0,060	-1,10122266	0,07199916		0,0600	1,3319E-05	1,77393E-10
2	0,080	-0,64621738	-0,18546330		0,0801	6,1982E-05	3,84177E-09
3	0,100	-0,18641644	-0,45453224		0,0998	-1,8575E-04	3,45026E-08
4	0,120	-0,45953563	0,26474106		0,1200	2,0909E-05	4,37198E-10
5	0,160	0,46451975	-0,26749980		0,1602	2,0221E-04	4,08908E-08
6	0,180	0,18076107	0,45618552		0,1798	-1,5175E-04	2,30274E-08
7	0,200	0,64429796	0,19216317		0,2002	1,7394E-04	3,02549E-08
8	0,220	1,10381332	-0,07759357		0,2199	-1,3487E-04	1,81893E-08
	testset						
1	0,110	-0,13985398	-0,34439227		0,1097	-3,4192E-04	1,16907E-07
2	0,170	0,13506407	0,34221936		0,1699	-1,4793E-04	2,18829E-08

Gegevens voor de regressie	
Meervoudige correlatiecoëfficiënt R	1,0000
R-kwadraat	1,0000
Aangepaste kleinste kwadraat	1,0000
Standaardfout	1,74E-04
Waarnemingen	8

Variantieanalyse

	Vrijheidsgraden	Kwadratensom	Gemiddelde kwadraten	F	Significantie F
Regressie	2	0,024000	0,012000	396505	9,98E-14
Storing	5	1,513E-07	3,026E-08		
Totaal	7	0,024			

	Coëfficiënten	Standaardfout	T-statistische gegevens	P-waarde	Laagste 95%	Hoogste 95%
Snijpunt	0,1400	6,151E-05	2276	3,11E-16	0,1398	0,1402
PC1	0,0764	8,974E-05	851	4,25E-14	0,0761	0,0766
PC2	0,0571	0,0002176	262	1,53E-11	0,0565	0,0577

De regressiecoëfficiënten zijn af te lezen in de Excel-uitvoer onder de kop 'Coëfficiënten' en de significantie ervan onder de kop 'P-waarde'. Alle regressiecoëfficiënten zijn significant omdat $p < 0{,}05$. Uit voorgaande Exceluitvoer blijkt dat $R^2_{\text{kal}} = 1{,}0000$, $RMSEC = s_r = 1{,}74 \cdot 10^{-4}$. Voor de geschatte kobaltconcentraties in de testmonsters, zie twee tabellen terug.

De correlatiecoëfficiënt voor de geschatte en juiste kobaltconcentraties voor de kalibratieset is $R^2_{\text{kal}} = 1{,}0000$. Voor de testset wordt voor de geschatte $\hat{y}$ en de juiste y geen correlatiecoëfficiënt berekend omdat er slechts twee testmonsters zijn.

Voor de kalibratieset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{8} (\hat{y}_i - y_i)^2 = 1{,}513 \cdot 10^{-7}$$

RMSEC kan worden berekend met (12.14) met $p = 3$:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}} (\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}513 \cdot 10^{-7}}{8-3}} = 1{,}74 \cdot 10^{-4}$$

Voor de testset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{2} (\hat{y}_i - y_i)^2 = 1{,}388 \cdot 10^{-7}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}} (\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{1{,}388 \cdot 10^{-7}}{2}} = 2{,}63 \cdot 10^{-4}$$

RMSEP van het SMLR-model in opgave 13.5 is $4{,}20 \cdot 10^{-4}$. Dit is hoger dan voor het PCR-model in deze opgave. De predictie van het PCR-model in deze opgave is beter dan die van het SMLR-model in opgave 13.5. Dit kan worden verklaard door het feit dat bij PCR de aanwezige spectrale informatie in alle golflengten kan worden benut.

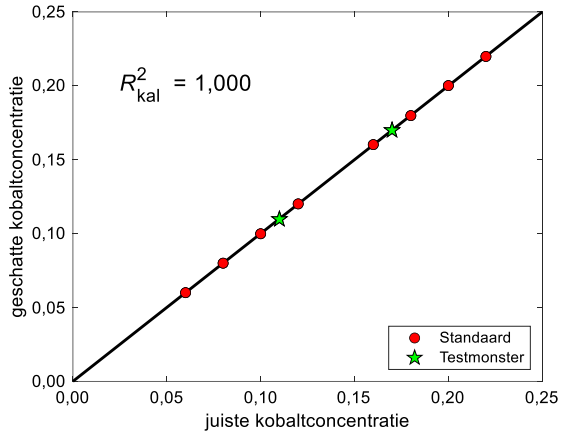
Er kan ook worden getest of de voorspelfout (RMSEP) van het PCR-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten. De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout:

$$F_{SMLR/PCR} = \frac{RMSEP^2_{SMLR}}{RMSEP^2_{PCR}} = \frac{(4{,}20 \cdot 10^{-4})^2}{(2{,}63 \cdot 10^{-4})^2} = 2{,}54$$

$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Omdat $F_{\text{SMLR/PCR}} < F_{\text{krit}}$ is er geen significant verschil tussen de varianties van de voorspelfout.

De predictie van het PCR-model voor kobalt is dus *niet* significant beter dan die van het SMLR-model in opgave 13.5.

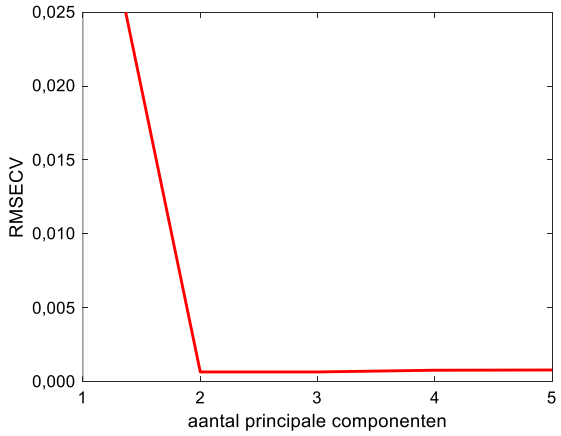


<u>Nikkel</u>

Correctie antwoorden voor nikkel:
RMSEC = $1{,}74 \cdot 10^{-4}$ moet zijn $5{,}12 \cdot 10^{-4}$.

Het PCR model voor nikkel wordt ontwikkeld na de PCA die hiervoor is beschreven bij kobalt. De optimale modelcomplexiteit A voor het PCR-model met nikkel kan anders zijn dan die voor kobalt en moet apart voor nikkel worden bepaald. Echter, omdat het oplossingen betreft van twee zuivere componenten (kobalt en nikkel) mag worden verwacht dat de modelcomplexiteit opnieuw zal liggen bij $A = 2$.

Er is een nieuwe leave-one-out (LOO) kruisvalidatie met de nikkelconcentraties herhaald uitgevoerd met een oplopend aantal principale componenten in het model. Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal principale componenten in het model en hiervan is de volgende grafiek getekend. Het minimum in deze curve ligt bij $A = 2$. Dit is de optimale modelcomplexiteit voor het PCR-model. Zie ook de informatie bij opgave 14.4 over de bepaling van de modelcomplexiteit.



De gereduceerde scorematrix $\boldsymbol{T}$ voor de kalibratie- en testset met twee principale componenten staat in de volgende tabel. Deze zijn identiek aan die voor kobalt. Dat geldt ook voor de score- en loadingplot.

Met behulp van de gereduceerde scorematrix $\boldsymbol{T}_{\mathbf{kal}}$ met twee principale componenten kan voor de kalibratieset een invers PCR-model worden ontwikkeld met de nikkelconcentraties in de $\boldsymbol{y}_{\mathbf{kal}}$-vector. Deze regressie kan worden uitgevoerd met Excel. Met de regressiecoëfficiënten van dit model kunnen geschatte nikkelconcentraties, de bijbehorende residuen en kwadraten van de residuen voor de kalibratieset en de testset worden berekend, zie volgende tabel.

Nr.	*Nikkel concentratie y_i*	***T*** (PC1)	***T*** (PC2)		$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset						
1	0,060	-1,10122266	0,07199916		0,0606	6,0340E-04	3,64091E-07
2	0,120	-0,64621738	-0,18546330		0,1193	-6,5121E-04	4,24072E-07
3	0,180	-0,18641644	-0,45453224		0,1797	-2,7361E-04	7,48647E-08
4	0,080	-0,45953563	0,26474106		0,0801	1,3257E-04	1,75752E-08
5	0,200	0,46451975	-0,26749980		0,2005	5,0392E-04	2,53934E-07
6	0,100	0,18076107	0,45618552		0,0997	-2,8034E-04	7,85925E-08
7	0,160	0,64429796	0,19216317		0,1598	-2,3610E-04	5,57447E-08
8	0,220	1,10381332	-0,07759357		0,2202	2,0138E-04	4,05537E-08
	testset						
1	0,170	-0,13985398	-0,34439227		0,1702	1,9001E-04	3,61055E-08
2	0,110	0,13506407	0,34221936		0,1097	-2,5216E-04	6,35829E-08

Gegevens voor de regressie	
Meervoudige correlatiecoëfficiënt R	1,0000
R-kwadraat	0,9999
Aangepaste kleinste kwadraat	0,9999
Standaardfout	5,12E-04
Waarnemingen	8

Variantieanalyse

	Vrijheidsgraden	Kwadratensom	Gemiddelde kwadraten	F	Significantie F
Regressie	2	0,023999	0,011999	45819	2,2E-11
Storing	5	1,309E-06	2,619E-07		
Totaal	7	0,024			

	Coëfficiënten	Standaardfout	T-statistische gegevens	P-waarde	Laagste 95%	Hoogste 95%
Snijpunt	0,1400	1,809E-04	773,78	6,842E-14	0,1395	0,1405
PC1	0,0647	2,640E-04	244,91	2,154E-11	0,0640	0,0653
PC2	-0,1139	6,402E-04	-177,92	1,064E-10	-0,1156	-0,1123

De regressiecoëfficiënten zijn af te lezen in de Excel-uitvoer onder de kop ‘Coëfficiënten’ en de significantie ervan onder de kop ‘P-waarde’. Alle regressiecoëfficiënten zijn significant omdat $p < 0{,}05$. Uit voorgaande Exceluitvoer blijkt dat $R^2_{\text{kal}} = 0{,}9999$, $RMSEC = s_r = 5{,}12 \cdot 10^{-4}$. Voor de geschatte nikkelconcentraties in de testmonsters, zie twee tabellen terug.

De correlatiecoëfficiënt voor de geschatte en juiste nikkelconcentraties voor de kalibratieset is $R^2_{\text{kal}} = 0{,}9999$. Voor de testset wordt voor de geschatte $\hat{y}$ en de juiste y geen correlatiecoëfficiënt berekend omdat er slechts twee testmonsters zijn.

Voor de kalibratieset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{8} (\hat{y}_i - y_i)^2 = 1{,}309 \cdot 10^{-6}$$

RMSEC kan worden berekend met (12.14) met p = 3:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}}(\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}309 \cdot 10^{-6}}{8-3}} = 5{,}12 \cdot 10^{-4}$$

Voor de testset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{2}(\hat{y}_i - y_i)^2 = 9{,}969 \cdot 10^{-8}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}}(\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{9{,}969 \cdot 10^{-8}}{2}} = 2{,}23 \cdot 10^{-4}$$

RMSEP van het SMLR-model in opgave 13.5 is $4{,}79 \cdot 10^{-3}$. Dit is hoger dan voor het PCR-model in deze opgave. De predictie van het PCR-model in deze opgave is beter dan die van het SMLR-model in opgave 13.5. Dit kan worden verklaard door het feit dat bij PCR de aanwezige spectrale informatie in alle golflengten kan worden benut.

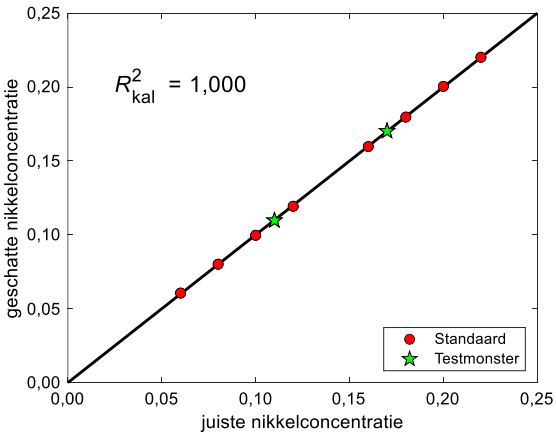
Er kan ook worden getest of de voorspelfout (RMSEP) van het PCR-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten. De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout:

$$F_{SMLR/PCR} = \frac{RMSEP_{SMLR}^2}{RMSEP_{PCR}^2} = \frac{(4{,}79 \cdot 10^{-3})^2}{(2{,}23 \cdot 10^{-4})^2} = 460$$

$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Omdat $F_{\text{SMLR/PCR}} > F_{\text{krit}}$ is er een significant verschil tussen de varianties van de voorspelfout.

De predictie van het PCR-model voor nikkel is dus *significant beter* dan die van het SMLR-model in opgave 13.5.



Antwoord 14.6

PCA is uitgevoerd na centrering van de variabelen in de $\boldsymbol{X}$ matrix en met niet-geschaalde concentraties in de $\boldsymbol{y}$ vector. Daarna is PCR uitgevoerd met een invers kalibratiemodel.

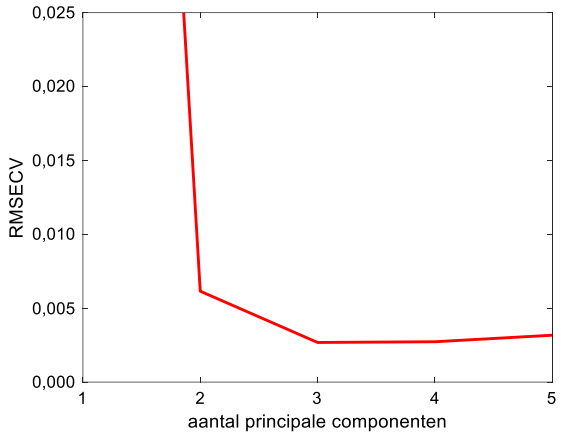
Antraceen

Na centrering worden de principale componenten berekend op basis van de variantie-covariantiematrix $\boldsymbol{S}$, zie bladzijde 322 vergelijking 14.5.

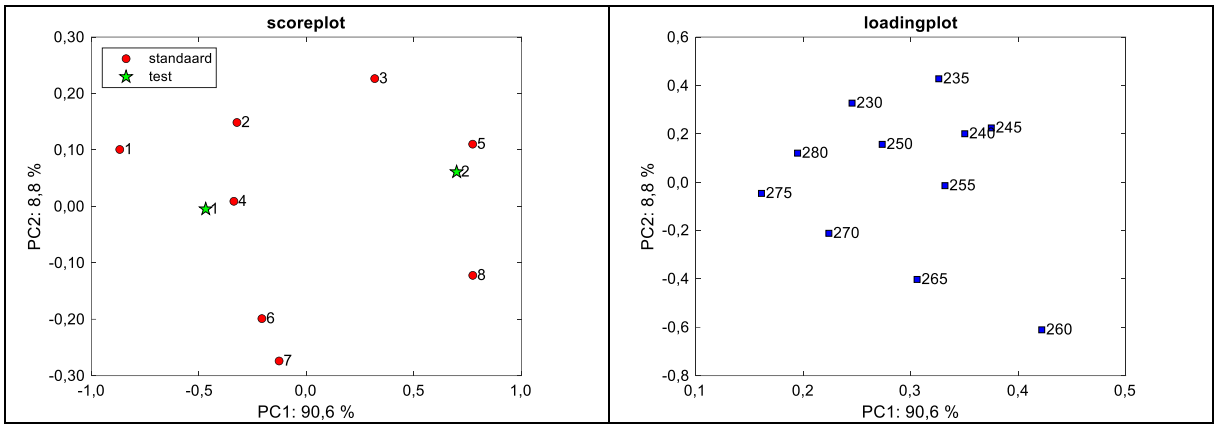
De optimale modelcomplexiteit A is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal principale componenten in het model.
De eigenwaarden van de covariantiematrix, cumulatieve eigenwaarden en percentages van de verklaarde varianties van de principale componenten zijn vermeld in de volgende tabel.

nummer j PC	λ_j	$\sum \lambda_j$	percentage verklaarde variantie
1	0,3325	0,3325	90,56
2	0,0322	0,3647	99,32
3	0,0025	0,3671	99,99
4	0,0000	0,3672	99,99
5	0,0000	0,3672	100,00

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal principale componenten in het model en hiervan is de volgende grafiek getekend. Het minimum in deze curve ligt bij $A = 3$. Dit is de optimale modelcomplexiteit voor het PCR-model. Er zijn drie principale componenten nodig omdat de oplossingen drie componenten bevatten: antraceen, benzo(a)antraceen en een interferent.



De gereduceerde scorematrix $\boldsymbol{T}$ voor de kalibratie- en testset met drie principale componenten staat in de volgende tabel.
Op basis van de score- en loadingsmatrix voor de eerste twee principale componenten kunnen respectievelijk de volgende score- en loadingplot worden gemaakt. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De scoreplot geeft het experimenteel design van de mengsels in de ruimte van de eerste twee principale componenten. Merk op dat dit design er anders uitziet dan het experimenteel design in de concentratieruimte, zie aan het begin van de uitwerking van opgave 13.6.

In de loadingplot corresponderen de golflengten die rechtsboven liggen voornamelijk met de antraceenpiek en de golflengten die liggen op de lijn van het midden naar rechtsonder met benzo(a)antraceen, zie de spectra aan het begin van de uitwerking van opgave 13.6.

Met behulp van de gereduceerde scorematrix $\boldsymbol{T}_{\mathbf{kal}}$ met drie principale componenten kan voor de kalibratieset een invers PCR-model worden ontwikkeld met de antraceenconcentraties in de $\boldsymbol{y}_{\mathbf{kal}}$ -vector. Deze regressie kan worden uitgevoerd met Excel, zie de volgende regressie-uitvoer van Excel. Met de regressiecoëfficiënten van dit model kunnen geschatte antraceenconcentraties, de bijbehorende residuen en kwadraten van de residuen voor de kalibratieset en de testset worden berekend, zie volgende tabel.

Nr.	*Antraceen concentratie* y_i	***T*** (PC1)	***T*** (PC2)	***T*** (PC3)		$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset							
1	0,150	-0,86820317	0,10078417	-0,01015241		0,1494	-5,792E-04	3,355E-07
2	0,200	-0,32334223	0,14862749	-0,03283074		0,2009	9,035E-04	8,162E-07
3	0,250	0,31741464	0,22637978	-0,03523001		0,2498	-2,075E-04	4,304E-08
4	0,300	-0,33753920	0,00900144	0,06892035		0,2982	-1,823E-03	3,322E-06
5	0,400	0,77287672	0,11022333	0,03385461		0,4017	1,718E-03	2,951E-06
6	0,450	-0,20729925	-0,19885976	0,05875725		0,4517	1,658E-03	2,748E-06
7	0,500	-0,12726035	-0,27392110	-0,07269030		0,5005	5,029E-04	2,530E-07
8	0,550	0,77335284	-0,12223535	-0,01062875		0,5478	-2,173E-03	4,720E-06
	testset							
1	0,275	-0,46839081	-0,00472525	-0,03697741		0,2771	2,067E-03	4,273E-06
2	0,425	0,69831468	0,06103279	0,08663271		0,4268	1,755E-03	3,081E-06

Gegevens voor de regressie	
Meervoudige correlatiecoëfficiënt R	0,9999
R-kwadraat	0,9999
Aangepaste kleinste kwadraat	0,9998
Standaardfout	1,95E-03
Waarnemingen	8

Variantieanalyse

	Vrijheidsgraden	Kwadratensom	Gemiddelde kwadraten	F	Significantie F
Regressie	3	0,1500	0,0500	13166	1,92E-08
Storing	4	1,519E-05	3,797E-06		
Totaal	7	0,15			

	Coëfficiënten	Standaardfout	T-statistische gegevens	P-waarde	Laagste 95%	Hoogste 95%
Snijpunt	0,3500	6,890E-04	508,0191	9,008E-11	0,3481	0,3519
PC1	0,1550	1,277E-03	121,3521	2,765E-08	0,1515	0,1585
PC2	-0,6457	4,106E-03	-157,2716	9,805E-09	-0,6571	-0,6343
PC3	0,0915	1,487E-02	6,1540	3,538E-03	0,0502	0,1328

De regressiecoëfficiënten zijn af te lezen in de Excel-uitvoer onder de kop 'Coëfficiënten' en de significantie ervan onder de kop 'P-waarde'. Alle regressiecoëfficiënten zijn significant omdat $p < 0{,}05$. Uit voorgaande Exceluitvoer blijkt dat $R^2_{\text{kal}} = 0{,}9999$, $RMSEC = s_r = 1{,}95 \cdot 10^{-3}$. Voor de geschatte antraceenconcentraties in de testmonsters, zie twee tabellen terug.

De correlatiecoëfficiënt voor de geschatte en juiste antraceenconcentraties voor de kalibratieset is $R^2_{\text{kal}} = 0{,}9999$. Voor de testset wordt voor de geschatte $\hat{y}$ en de juiste y geen correlatiecoëfficiënt berekend omdat er slechts twee testmonsters zijn.

Voor de kalibratieset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{8} (\hat{y}_i - y_i)^2 = 1{,}519 \cdot 10^{-5}$$

RMSEC kan worden berekend met (12.14) met $p = 4$:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}} (\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}519 \cdot 10^{-5}}{8-4}} = 1{,}95 \cdot 10^{-3}$$

Voor de testset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{2} (\hat{y}_i - y_i)^2 = 7{,}355 \cdot 10^{-6}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}} (\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{7{,}355 \cdot 10^{-6}}{2}} = 1{,}92 \cdot 10^{-3}$$

RMSEP van het SMLR-model in opgave 13.6 is $3{,}21 \cdot 10^{-3}$. Dit is hoger dan voor het PCR-model in deze opgave. De predictie van het PCR-model in deze opgave is beter dan die van het SMLR-model in opgave 13.6. Dit kan worden verklaard door het feit dat bij PCR de aanwezige spectrale informatie in alle golflengten kan worden benut.

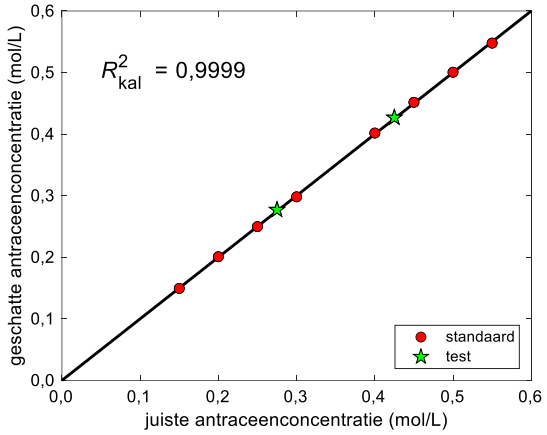
Er kan ook worden getest of de voorspelfout (RMSEP) van het PCR-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten. De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout:

$$F_{SMLR/PCR} = \frac{RMSEP^2_{SMLR}}{RMSEP^2_{PCR}} = \frac{(3{,}21 \cdot 10^{-3})^2}{(1{,}92 \cdot 10^{-3})^2} = 2{,}81$$

$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Omdat $F_{\text{SMLR/PCR}} < F_{\text{krit}}$ is er geen significant verschil tussen de varianties van de voorspelfout.

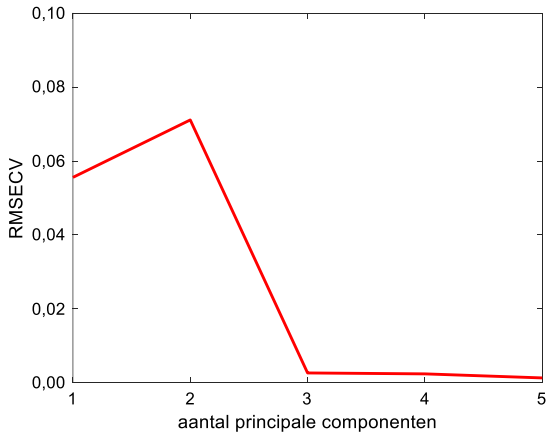
De predictie van het PCR-model voor antraceen is dus *niet* significant beter dan die van het SMLR-model in opgave 13.6.



<u>Benzo(a)antraceen</u>

Het PCR model voor benzo(a)antraceen wordt ontwikkeld na de PCA die hiervoor is beschreven bij antraceen. De optimale modelcomplexiteit *A* voor het PCR-model met benzo(a)antraceen kan anders zijn dan die voor antraceen. Echter, omdat het oplossingen betreft met drie componenten (antraceen, benzo(a)antraceen en een interferent) mag worden verwacht dat de modelcomplexiteit opnieuw zal liggen bij *A* = 3.

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal principale componenten in het model en hiervan is de volgende grafiek getekend. De afname van RMSECV na *A* = 3 is zeer gering. Omdat het streven is naar een zo klein mogelijke modelcomplexiteit wordt in dit geval *A* =3 als optimaal beschouwd.



De gereduceerde scorematrix ***T*** voor de kalibratie- en testset met drie principale componenten staat in de volgende tabel.

Met behulp van de gereduceerde scorematrix $\boldsymbol{T}_{\mathbf{kal}}$ met drie principale componenten kan voor de kalibratieset een invers PCR-model worden ontwikkeld met de benzo(a)antraceen-concentraties in de $\boldsymbol{y}_{\mathbf{kal}}$ -vector. Deze regressie kan worden uitgevoerd met Excel, zie de volgende regressie-uitvoer van Excel. Met de regressiecoëfficiënten van dit model kunnen geschatte benzo(a)antraceenconcentraties, de bijbehorende residuen en kwadraten van de residuen voor de kalibratieset en de testset worden berekend, zie volgende tabel.

Nr.	*Benzo(a)antraceen concentratie y_i*	***T*** (PC1)	***T*** (PC2)	***T*** (PC3)		$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset							
1	0,150	-0,86820317	0,10078417	-0,01015241		0,1485	-1,506E-03	2,268E-06
2	0,300	-0,32334223	0,14862749	-0,03283074		0,2993	-6,659E-04	4,434E-07
3	0,450	0,31741464	0,22637978	-0,03523001		0,4523	2,281E-03	5,202E-06
4	0,200	-0,33753920	0,00900144	0,06892035		0,2017	1,652E-03	2,729E-06
5	0,500	0,77287672	0,11022333	0,03385461		0,4980	-1,950E-03	3,804E-06
6	0,250	-0,20729925	-0,19885976	0,05875725		0,2501	1,460E-04	2,133E-08
7	0,400	-0,12726035	-0,27392110	-0,07269030		0,4002	2,053E-04	4,216E-08
8	0,550	0,77335284	-0,12223535	-0,01062875		0,5498	-1,620E-04	2,624E-08
	testset							
1	0,275	-0,46839081	-0,00472525	-0,03697741		0,2741	-8,777E-04	7,703E-07
2	0,425	0,69831468	0,06103279	0,08663271		0,4305	5,469E-03	2,991E-05

Gegevens voor de regressie	
Meervoudige correlatiecoëfficiënt R	1,0000
R-kwadraat	0,9999
Aangepaste kleinste kwadraat	0,9998
Standaardfout	1,91E-03
Waarnemingen	8

Variantieanalyse

	Vrijheidsgraden	Kwadratensom	Gemiddelde kwadraten	F	Significantie F
Regressie	3	0,1500	0,0500	13757	1,76E-08
Storing	4	1,454E-05	3,634E-06		
Totaal	7	0,15			

	Coëfficiënten	Standaardfout	T-statistische gegevens	P-waarde	Laagste 95%	Hoogste 95%
Snijpunt	0,3500	6,740E-04	519,2959	8,251E-11	0,3481	0,3519
PC1	0,2394	1,250E-03	191,5666	4,454E-09	0,2359	0,2428
PC2	-0,0356	4,017E-03	-8,8689	8,927E-04	-0,0468	-0,0245
PC3	-0,9755	1,455E-02	-67,0473	2,965E-07	-1,0159	-0,9351

De regressiecoëfficiënten zijn af te lezen in de Excel-uitvoer onder de kop ‘Coëfficiënten’ en de significantie ervan onder de kop ‘P-waarde’. Alle regressiecoëfficiënten zijn significant omdat $p < 0{,}05$. Uit voorgaande Exceluitvoer blijkt dat $R^2_{\text{kal}} = 0{,}9999$, $RMSEC = s_r = 1{,}91 \cdot 10^{-3}$. Voor de geschatte benzo(a)antraceenconcentraties in de testmonsters, zie twee tabellen terug.

De correlatiecoëfficiënt voor de geschatte en juiste benzo(a)antraceenconcentraties voor de kalibratieset is $R^2_{\text{kal}} = 0{,}9999$. Voor de testset wordt voor de geschatte $\hat{y}$ en de juiste y geen correlatiecoëfficiënt berekend omdat er slechts twee testmonsters zijn.

Voor de kalibratieset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{8} (\hat{y}_i - y_i)^2 = 1{,}454 \cdot 10^{-5}$$

RMSEC kan worden berekend met (12.14) met p = 4:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}}(\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}454 \cdot 10^{-5}}{8-4}} = 1{,}91 \cdot 10^{-3}$$

Voor de testset is de som van de kwadraten van de residuen:

$$\sum_{i=1}^{2}(\hat{y}_i - y_i)^2 = 3{,}068 \cdot 10^{-5}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}}(\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{3{,}068 \cdot 10^{-5}}{2}} = 3{,}92 \cdot 10^{-3}$$

RMSEP van het SMLR-model in opgave 13.6 is $6{,}86 \cdot 10^{-2}$. Dit is hoger dan voor het PCR-model in deze opgave. De predictie van het PCR-model in deze opgave is beter dan die van het SMLR-model in opgave 13.6. Dit kan worden verklaard door het feit dat bij PCR de aanwezige spectrale informatie in alle golflengten kan worden benut.

Er kan ook worden getest of de voorspelfout (RMSEP) van het PCR-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten. De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout:

$$F_{SMLR/PCR} = \frac{RMSEP_{SMLR}^2}{RMSEP_{PCR}^2} = \frac{(6{,}86 \cdot 10^{-2})^2}{(3{,}92 \cdot 10^{-3})^2} = 307$$

$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Omdat $F_{\text{SMLR/PCR}} > F_{\text{krit}}$ is er een significant verschil tussen de varianties van de voorspelfout.

De predictie van het PCR-model voor antraceen is dus *significant beter* dan die van het SMLR-model in opgave 13.6.