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 15

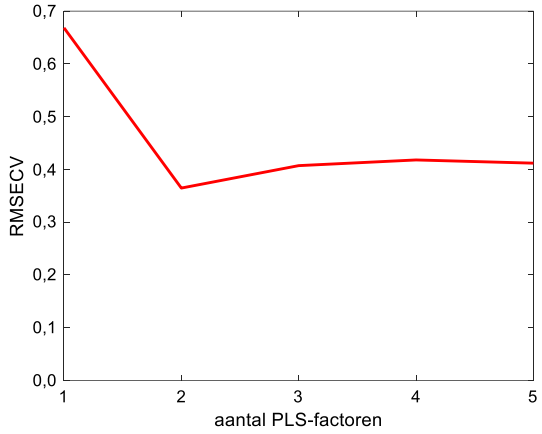
Antwoord 15.1

PLS is uitgevoerd met een klassiek kalibratiemodel na autoschaling van de LSER-factoren in de $\boldsymbol{X}$ matrix en centrering van de log k_w waarden in de $\boldsymbol{y}$ vector.

De optimale modelcomplexiteit A is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal PLS-factoren in het model. De verklaarde variantie voor elke PLS-factor, de cumulatieve verklaarde variantie en de percentages van de verklaarde varianties van de PLS-factoren zijn vermeld in de volgende tabel.

nummer j PC	$var(PLS_j)$	$\sum var(PLS_j)$	percentage verklaarde variantie
1	0,1998	0,1998	39,83
2	0,1261	0,3259	64,97
3	0,0784	0,4044	80,61
4	0,0426	0,4469	89,09
5	0,0547	0,5017	100,00

De RMSECV-curve als functie van het aantal PLS-factoren is afgebeeld in de volgende grafiek. Het minimum van de RMSECV-curve ligt bij 2 PLS-factoren dat als optimum van de modelcomplexiteit wordt beschouwd.



De loadingsmatrix $\boldsymbol{P}$ voor twee PLS-factoren is:

```
[ -0,2900      0,6823
  -0,7778      0,3594
  -0,2894      0,0569
  -0,4409      0,1599
   0,3704      0,6515]
```

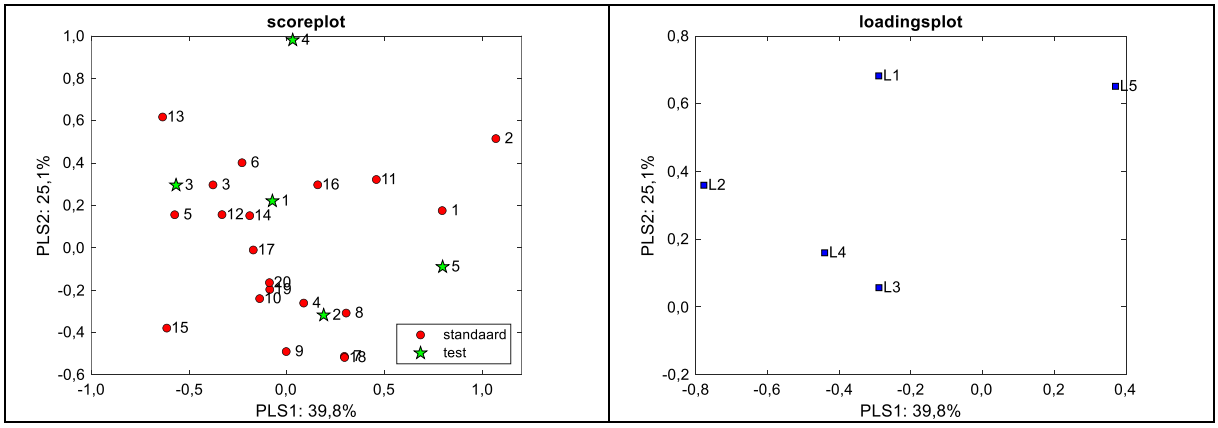
De matrix met weegfactoren $\boldsymbol{W}$ voor twee PLS-factoren is:

[-0,0952	0,6030
-0,6358	0,4394
-0,2386	0,1573
-0,4404	0,0012
0,5794	0,6470]

De getransponeerde $\boldsymbol{y}$-loadingsvector $\boldsymbol{q}$ voor twee PLS-factoren is:

[3,0530 1,5627]

De scorematrix $\boldsymbol{T}$ voor de kalibratie- en testset voor twee PLS-factoren staan in de volgende tabel. Op basis van de scorematrix $\boldsymbol{T}$ en de loadingsmatrix $\boldsymbol{P}$ kunnen respectievelijk scoreplots en loadingplots worden gemaakt. In de volgende afbeeldingen zijn deze gegeven voor de eerste twee PLS-factoren. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De getransponeerde PLS-regressievector $\boldsymbol{b}$ kan met behulp van vergelijking (15.19) worden berekend:

[0,6036 -1,5757 -0,6030 -1,5651 3,0725]

Met behulp van deze PLS-regressievector $\boldsymbol{b}$ kunnen de geschatte waarden voor de kalibratiemonsters kunnen worden berekend met vergelijking (15.21) en voor de testmonsters met vergelijking (15.28). Dit is de eenvoudigste methode, mits de berekening van de inverse $(\boldsymbol{P}_a^T\boldsymbol{W}_a)^{-1}$ in vergelijking (15.19) geen problemen oplevert.

De gemiddelde responsiewaarde voor de kalibratiemonsters is $\overline{\boldsymbol{y}}_{kal} = 2{,}2908$. De geschatte waarden voor $\hat{\boldsymbol{y}}_{\mathbf{kal}}$ en $\hat{\boldsymbol{y}}_{\mathbf{test}}$, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel.

	Log $k_w = y_i$	$\boldsymbol{T}$		$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset					
S1	5,4892	0,7937	0,1751	4,9877	-0,5015	0,2515
S2	6,0714	1,0681	0,5153	6,3571	0,2857	0,0817
S3	1,5692	-0,3786	0,2967	1,5986	0,0294	0,0009
S4	2,5650	0,0853	-0,2617	2,1423	-0,4227	0,1787
S5	1,1786	-0,5744	0,1557	0,7806	-0,3980	0,1584
S6	2,3819	-0,2300	0,4013	2,2158	-0,1661	0,0276
S7	2,0052	0,2934	-0,5136	2,3840	0,3788	0,1435
S8	2,6725	0,3026	-0,3089	2,7318	0,0593	0,0035
S9	1,0396	-0,0040	-0,4908	1,5115	0,4719	0,2227
S10	1,0938	-0,1399	-0,2407	1,4877	0,3939	0,1552
S11	4,5311	0,4569	0,3221	4,1889	-0,3422	0,1171
S12	1,6175	-0,3324	0,1561	1,5199	-0,0976	0,0095
S13	0,8923	-0,6361	0,6174	1,3136	0,4213	0,1775
S14	1,5758	-0,1906	0,1512	1,9451	0,3693	0,1364
S15	0,3015	-0,6149	-0,3799	0,1802	-0,4817	0,2320
S16	3,0787	0,1570	0,2971	3,2342	0,1555	0,0242
S17	1,9694	-0,1725	-0,0109	1,7470	-0,2224	0,0495
S18	2,6125	0,2943	-0,5188	2,3786	-0,2339	0,0547
S19	1,5680	-0,0882	-0,1975	1,7129	0,1449	0,0210
S20	1,6027	-0,0898	-0,1652	1,7586	0,1559	0,0243
	testset					
T1	2,9099	-0,0746	0,2208	2,3702	-0,5397	0,2912
T2	2,0436	0,1877	-0,3187	2,4605	0,4169	0,1738
T3	0,8308	-0,5673	0,2954	0,7339	-0,0969	0,0094
T4	4,0185	0,0292	0,9812	3,9280	-0,0905	0,0082
T5	4,8854	0,7958	-0,0904	4,9810	0,0956	0,0091

De correlatiecoëfficiënten voor de geschatte en de juiste log k_w kunnen afzonderlijk worden berekend voor de kalibratieset en testset. De kwadraten van deze correlatiecoëfficiënten zijn: $R^2_{\text{kal}} = 0{,}9522$ en $R^2_{\text{test}} = 0{,}9543$.

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

$$\sum_{i=1}^{20} (\hat{y}_i - y_i)^2 = 2{,}0699$$

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

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}} (\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{2{,}0699}{20-2}} = 0{,}3391$$

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

$$\sum_{i=1}^{5} (\hat{y}_i - y_i)^2 = 0{,}4918$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}} (\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{0{,}4918}{5}} = 0{,}3136$$

Voor de RMSEP van het SMLR-model in opgave 13.2 is gevonden 0,1630.
Voor de RMSEP van het PCR-model in opgave 14.2 is gevonden 0,2689.

Beide RMSEP's zijn lager dan die voor het PLS-model in deze opgave. De predictie van het SMLR-model in opgave 13.2 is beter dan die van het PCR-model en de predictie van het PCR-model is ook weer beter dan die van het PLS-model.

Er kan worden getest of de voorspelfout (RMSEP) van het SMLR-model significant beter (lager) is dan van het PLS-model en of de voorspelfout (RMSEP) van het PCR-model significant beter (lager) is dan van het PLS-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten.

De tests kunnen worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout. $F_{\text{krit}} = F_{(0{,}05;5;5)} = 5{,}05$ (zie tabel 4 van Bijlage 1)

Vergelijking van $\text{RMSEP}_{\text{PLS}}$ met $\text{RMSEP}_{\text{SMLR}}$:

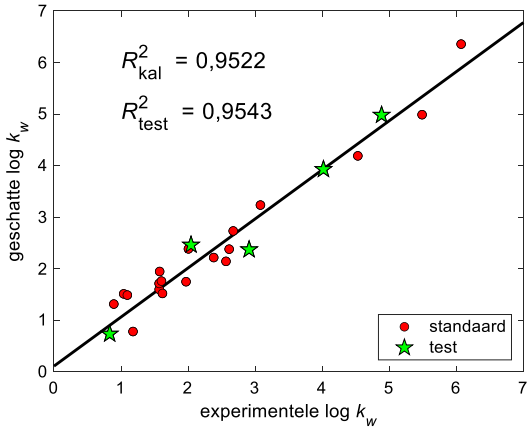
$$F_{PLS/SMLR} = \frac{RMSEP_{PLS}^2}{RMSEP_{SMLR}^2} = \frac{0{,}3136^2}{0{,}1630^2} = 3{,}70$$

Omdat $F_{\text{PLS/SMLR}} < F_{\text{krit}}$ 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 PLS-model.

Vergelijking van $\text{RMSEP}_{\text{PLS}}$ met $\text{RMSEP}_{\text{PCR}}$:

$$F_{PLS/PCR} = \frac{RMSEP_{PLS}^2}{RMSEP_{PCR}^2} = \frac{0{,}3136^2}{0{,}2689^2} = 1{,}36$$

Omdat $F_{\text{PLS/PCR}} < F_{\text{krit}}$ is er geen significant verschil tussen de varianties van de voorspelfout. De predictie van het PCR-model in opgave 14.2 is dus *niet* significant beter dan die van het PLS-model.



Antwoord 15.2

PLS is uitgevoerd met een invers kalibratiemodel na centrering van de variabelen in de $\boldsymbol{X}$ matrix en de $\boldsymbol{y}$ vector.

De optimale modelcomplexiteit A is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal PLS-factoren in het model. De verklaarde variantie voor elke PLS-factor, de cumulatieve verklaarde variantie en de percentages van de verklaarde varianties van de PLS-factoren zijn vermeld in de volgende tabel. Daaruit blijkt dat er slechts één PLS-factor nodig is om 99,99 % van de totale variantie in de dataset te

verklaren (A = 1). Er is net zoals bij PCR maar één PLS-factor nodig omdat er ook maar één chemische component aanwezig is in de zuivere kobaltoplossingen.

nummer j PLS	$var(PLS_j)$	$\sum var(PLS_j)$	percentage verklaarde variantie
1	0,0977	0,0977	99,99
2	0,0000	0,0977	100,00
3	0,0000	0,0977	100,00

De getransponeerde loadingsmatrix $\boldsymbol{P}$ voor één PLS-factor 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]

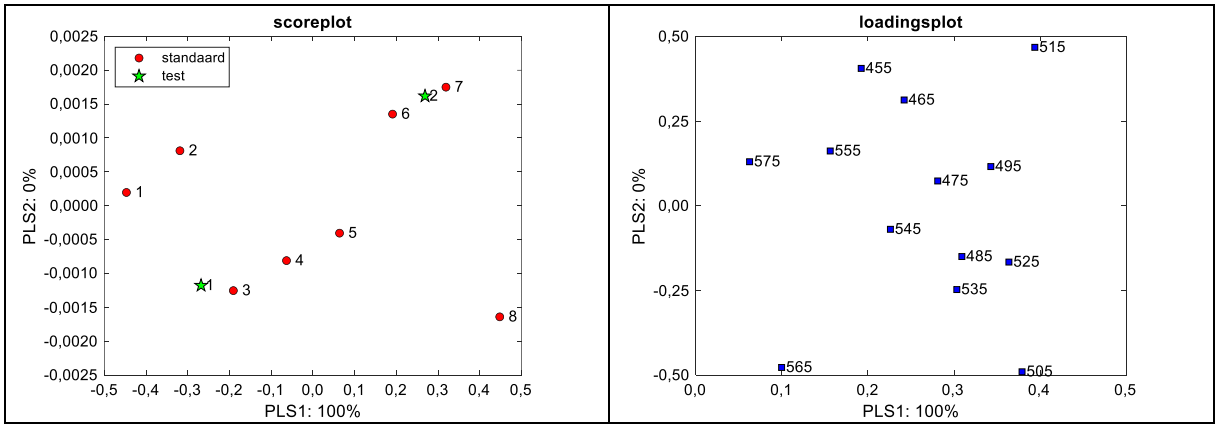
De getransponeerde matrix met weegfactoren $\boldsymbol{W}$ voor één PLS-factor is:

[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]

Merk op dat $\boldsymbol{W}$ identiek is aan $\boldsymbol{P}$.

De $\boldsymbol{y}$-loading q voor één PLS-factor is 0,0784.

De scorematrix $\boldsymbol{T}$ voor de kalibratie- en testset voor twee PLS-factoren staan in de volgende tabel. Op basis van de scorematrix $\boldsymbol{T}$ en de loadingsmatrix $\boldsymbol{P}$ kunnen respectievelijk scoreplots en loadingplots worden gemaakt. In de volgende afbeeldingen zijn deze gegeven voor de eerste twee PLS-factoren. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De getransponeerde PLS-regressievector $\boldsymbol{b}$ kan met behulp van vergelijking (15.19) worden berekend:

[0,0151 0,0190 0,0220 0,0242 0,0269 0,0297 0,0309 0,0285 0,0238
0,0177 0,0122 0,0078 0,0049]

Met behulp van deze PLS-regressievector $\boldsymbol{b}$ kunnen de geschatte waarden voor de kalibratiemonsters kunnen worden berekend met vergelijking (15.21) en voor de testmonsters met vergelijking (15.28). Dit is de eenvoudigste methode, mits de berekening van de inverse $(\boldsymbol{P}_a^T\boldsymbol{W}_a)^{-1}$ in vergelijking (15.19) geen problemen oplevert.

De gemiddelde responsiewaarde voor de kalibratiemonsters is $\overline{\boldsymbol{y}}_{\boldsymbol{kal}} = 0{,}0450$. De geschatte waarden voor $\widehat{\boldsymbol{y}}_{\mathbf{kal}}$ en $\widehat{\boldsymbol{y}}_{\mathbf{test}}$, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel.

	Co-concentratie	***T*** voor PLS1		$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset					
1	0,010	-0,4471		0,0100	-2,748E-05	7,550E-10
2	0,020	-0,3189		0,0200	1,286E-05	1,654E-10
3	0,030	-0,1909		0,0300	4,474E-05	2,001E-09
4	0,040	-0,0635		0,0400	2,632E-05	6,929E-10
5	0,050	0,0636		0,0500	-1,726E-05	2,979E-10
6	0,060	0,1907		0,0599	-6,134E-05	3,763E-09
7	0,070	0,3186		0,0700	-3,806E-05	1,449E-09
8	0,080	0,4475		0,0801	6,022E-05	3,626E-09
	testset					
1	0,024	-0.2682		0,0240	-1,467E-05	2,151E-10
2	0,066	0.2684		0,0660	3,292E-05	1,084E-09

Het kwadraat van de correlatiecoëfficiënt voor de geschatte en de juiste kobaltconcentraties voor de kalibratieset is $R^2_{\text{kal}} = 1{,}0000$. Voor de testset wordt voor de geschatte en de juiste kobaltconcentraties 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 $p = a = 1$:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}} (\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}275 \cdot 10^{-8}}{8-1}} = 4{,}27 \cdot 10^{-5}$$

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):

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

Voor de RMSEP van het SMLR-model met b_0-term in opgave 13.3 is gevonden $1{,}85 \cdot 10^{-4}$. Voor de RMSEP van het PCR-model in opgave 14.3 is gevonden $2{,}55 \cdot 10^{-5}$.

De RMSEP voor het PLS-model is lager dan die voor het SMLR-model en gelijk aan die voor het PCR-model. De predictie van het PLS-model is dus beter dan die van het SMLR-model en gelijk aan die van het PCR-model.

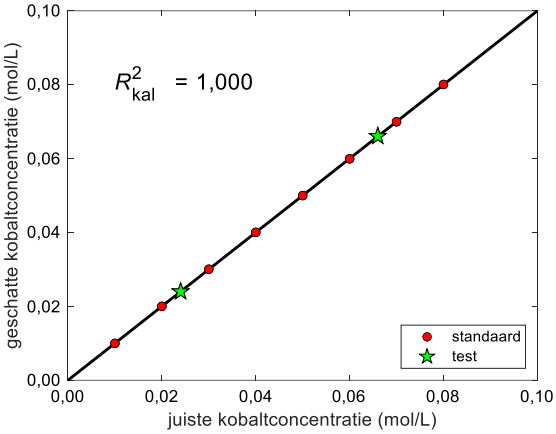
Er kan worden getest of de voorspelfout (RMSEP) van het PLS-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten.

De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout.
$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Vergelijking van $\text{RMSEP}_{\text{PLS}}$ met $\text{RMSEP}_{\text{SMLR}}$:

$$F_{SMLR/PLS} = \frac{RMSEP_{SMLR}^2}{RMSEP_{PLS}^2} = \frac{(1{,}85 \cdot 10^{-4})^2}{(2{,}55 \cdot 10^{-5})^2} = 53$$

Omdat $F_{\text{SMLR/PLS}} > F_{\text{krit}}$ is er een significant verschil tussen de varianties van de voorspelfout. De predictie van het PLS-model is dus *significant beter* dan die van het SMLR-model in opgave 13.3.



Antwoord 15.3

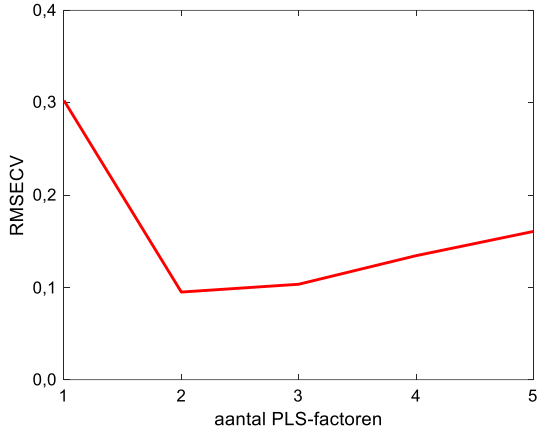
Correctie antwoorden
RMSEC = $8{,}18 \cdot 10^{-2}$ moet zijn RMSEC = $7{,}53 \cdot 10^{-2}$

PLS is uitgevoerd met een invers kalibratiemodel na centrering van de variabelen in de $\boldsymbol{X}$ matrix en de $\boldsymbol{y}$ vector.

De optimale modelcomplexiteit A is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal PLS-factoren in het model. RMSECV, de verklaarde variantie voor elke PLS-factor, de cumulatieve verklaarde variantie en de percentages van de verklaarde varianties van de PLS-factoren zijn vermeld in de volgende tabel.

Aantal PLS-factoren	RMSECV	Verklaarde variantie	Cumulative verklaarde variantie	Percentage verklaarde variantie
1	0,3023	0,2889	0,2889	99,69
2	0,0949	0,0005	0,2894	99,88
3	0,1034	0,0003	0,2898	100,00
4	0,1343	0,0000	0,2898	100,00
5	0,1605	0,0000	0,2898	100,00

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal PLS-factoren 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 PLS-model.



De loadingsmatrix ***P*** voor twee PLS-factoren is:

[0,2228	-0,4725
0,2409	-0,4982
0,2509	-0,3470
0,2621	-0,2878
0,3062	-0,1773
0,3181	-0,0576
0,3259	0,0888
0,3240	0,2343
0,3152	0,2447
0,3207	0,2712
0,2832	0,2508
0,2778	0,1803]

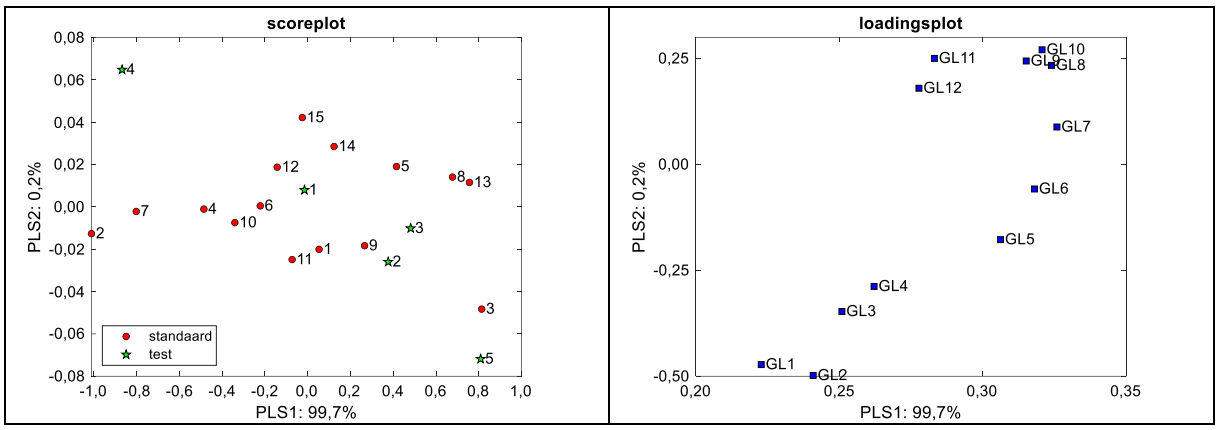
De matrix met weegfactoren ***W*** voor twee PLS-factoren is:

[0,1888	-0,4836
0,2054	-0,5054
0,2262	-0,3511
0,2424	-0,2811
0,2948	-0,1618
0,3150	-0,0433
0,3327	0,0965
0,3405	0,2353
0,3322	0,2425
0,3394	0,2662
0,3000	0,2398
0,2896	0,1687]

De getransponeerde ***y***-loadingsvector ***q*** voor twee PLS-factoren is:

[0,3008 11,3301]

De scorematrix ***T*** voor de kalibratie- en testset voor twee PLS-factoren staan in de volgende tabel. 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 PLS-factoren. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De getransponeerde PLS-regressievector $\boldsymbol{b}$ kan met behulp van vergelijking (15.19) worden berekend:

[-5,2725 -5,5005 -3,7297 -2,9191 -1,5098 -0,1444 1,4584 3,0396 3,1126
3,3887 3,0462 2,2298]

Met behulp van deze PLS-regressievector $\boldsymbol{b}$ kunnen de geschatte waarden voor de kalibratiemonsters kunnen worden berekend met vergelijking (15.21) en voor de testmonsters met vergelijking (15.28). Dit is de eenvoudigste methode, mits de berekening van de inverse $(\boldsymbol{P}_a^T\boldsymbol{W}_a)^{-1}$ in vergelijking (15.19) geen problemen oplevert.

De gemiddelde responsiewaarde voor de kalibratiemonsters is $\bar{\boldsymbol{y}}_{\boldsymbol{kal}} = 41{,}9267$. De geschatte waarden voor $\hat{\boldsymbol{y}}_{\mathbf{kal}}$ en $\hat{\boldsymbol{y}}_{\mathbf{test}}$, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel.

	vetpercentage	$\boldsymbol{T}$			$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset						
1	41,7	0,0530	-0,0200		41,7157	0,0157	2,46E-04
2	41,5	-1,0112	-0,0126		41,4793	-0,0207	4,27E-04
3	41,7	0,8132	-0,0483		41,6243	-0,0757	5,73E-03
4	41,7	-0,4856	-0,0010		41,7688	0,0688	4,74E-03
5	42,2	0,4148	0,0191		42,2676	0,0676	4,57E-03
6	41,9	-0,2221	0,0005		41,8653	-0,0347	1,20E-03
7	41,7	-0,8018	-0,0021		41,6612	-0,0388	1,51E-03
8	42,2	0,6768	0,0141		42,2900	0,0900	8,10E-03
9	41,7	0,2658	-0,0183		41,7991	0,0991	9,81E-03
10	41,8	-0,3407	-0,0074		41,7399	-0,0601	3,61E-03
11	41,6	-0,0731	-0,0249		41,6228	0,0228	5,21E-04
12	42,0	-0,1428	0,0188		42,0963	0,0963	9,27E-03
13	42,4	0,7560	0,0116		42,2850	-0,1150	1,32E-02
14	42,3	0,1230	0,0286		42,2873	-0,0127	1,62E-04
15	42,5	-0,0254	0,0422		42,3973	-0,1027	1,05E-02
	testset						
1	41,9	-0,0155	0,0080		42,0001	0,1001	1,00E-02
2	42,0	0,3761	-0,0260		42,0452	0,0452	2,04E-03
3	42,3	0,4813	-0,0101		42,3404	0,0404	1,63E-03
4	41,6	-0,8679	0,0648		41,7082	0,1082	1,17E-02
5	42,1	0,8083	-0,0718		42,0000	-0,1000	1,00E-02

De correlatiecoëfficiënten voor de geschatte en de juiste vetpercentages kunnen afzonderlijk worden berekend voor de kalibratieset en testset. De kwadraten van deze correlatiecoëfficiënten zijn: $R^2_{\text{kal}} = 0{,}9477$ en $R^2_{\text{test}} = 0{,}9029$.

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

$$\sum_{i=1}^{15}(\hat{y}_i - y_i)^2 = 7{,}365 \cdot 10^{-2}$$

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

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}}(\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{7{,}365 \cdot 10^{-2}}{15-2}} = 7{,}53 \cdot 10^{-2}$$

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

$$\sum_{i=1}^{5}(\hat{y}_i - y_i)^2 = 3{,}540 \cdot 10^{-2}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}}(\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{3{,}540 \cdot 10^{-2}}{5}} = 8{,}41 \cdot 10^{-2}$$

Voor de RMSEP van het SMLR-model in opgave 13.4 is gevonden $1{,}20 \cdot 10^{-1}$.
Voor de RMSEP van het PCR-model in opgave 14.4 is gevonden $8{,}19 \cdot 10^{-2}$.

De RMSEP voor het PLS-model is lager dan die voor het SMLR-model en vergelijkbaar met die voor het PCR-model. De predictie van het PLS-model is dus beter dan die van het SMLR-model en vergelijkbaar met die van het PCR-model.

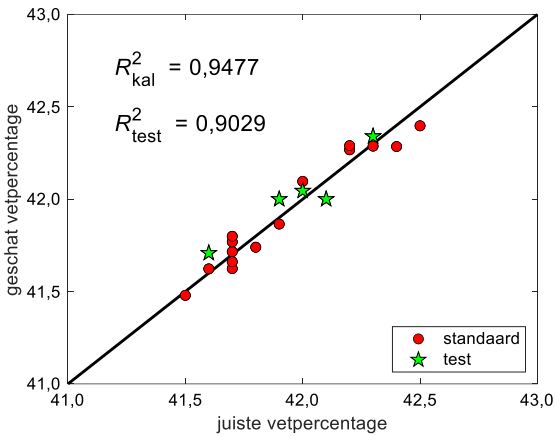
Er kan worden getest of de voorspelfout (RMSEP) van het PLS-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten.

De test kan worden uitgevoerd met een eenzijdige *F*-test op de varianties van de voorspelfout.
$F_{\text{krit}} = F_{(0{,}05;5;5)} = 5{,}05$ (zie tabel 4 van Bijlage 1)

Vergelijking van $\text{RMSEP}_{\text{PLS}}$ met $\text{RMSEP}_{\text{SMLR}}$:

$$F_{SMLR/PLS} = \frac{RMSEP_{SMLR}^2}{RMSEP_{PLS}^2} = \frac{(1{,}20 \cdot 10^{-1})^2}{(8{,}41 \cdot 10^{-2})^2} = 2{,}04$$

Omdat $F_{\text{SMLR/PLS}} < F_{\text{krit}}$ is er geen significant verschil tussen de varianties van de voorspelfout. De predictie van het PLS-model is dus *niet* significant beter dan die van het SMLR-model in opgave 13.4.



Antwoord 15.4

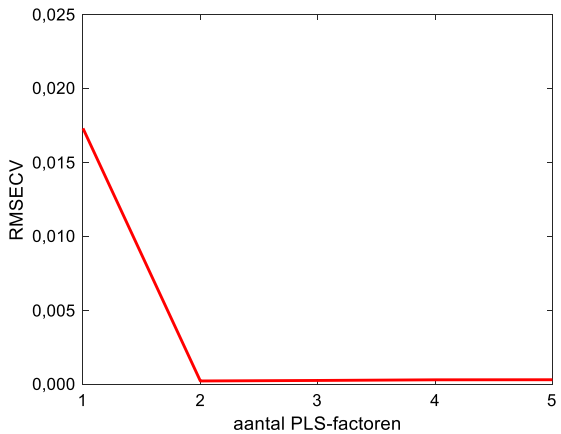
PLS is uitgevoerd met een invers kalibratiemodel na centrering van de variabelen in de ***X*** matrix en de ***y*** vector.

<u>Kobalt</u>

De optimale modelcomplexiteit *A* is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal PLS-factoren in het model. RMSECV, de verklaarde variantie voor elke PLS-factor, de cumulatieve verklaarde variantie en de percentages van de verklaarde varianties van de PLS-factoren zijn vermeld in de volgende tabel.

Aantal PLS-factoren	RMSECV	Verklaarde variantie	Cumulative verklaarde variantie	Percentage verklaarde variantie
1	0,0173	0,5298	0,5298	85,13
2	0,0002	0,0925	0,6223	99,99
3	0,0003	0,0000	0,6223	99,99
4	0,0003	0,0000	0,6223	99,99
5	0,0003	0,0000	0,6223	100,00

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal PLS-factoren 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 PLS-model.



De loadingsmatrix ***P*** voor twee PLS-factoren is:

[0,0790	-0,1501
0,2174	-0,3932
0,3286	-0,5820
0,2870	-0,4799
0,1846	-0,2326
0,1557	-0,0341
0,2163	0,0596
0,2771	0,1238
0,3142	0,1703
0,3661	0,2117
0,3954	0,2282
0,3395	0,1915
0,2346	0,1183
0,1384	0,0556
0,0820	0,0056]

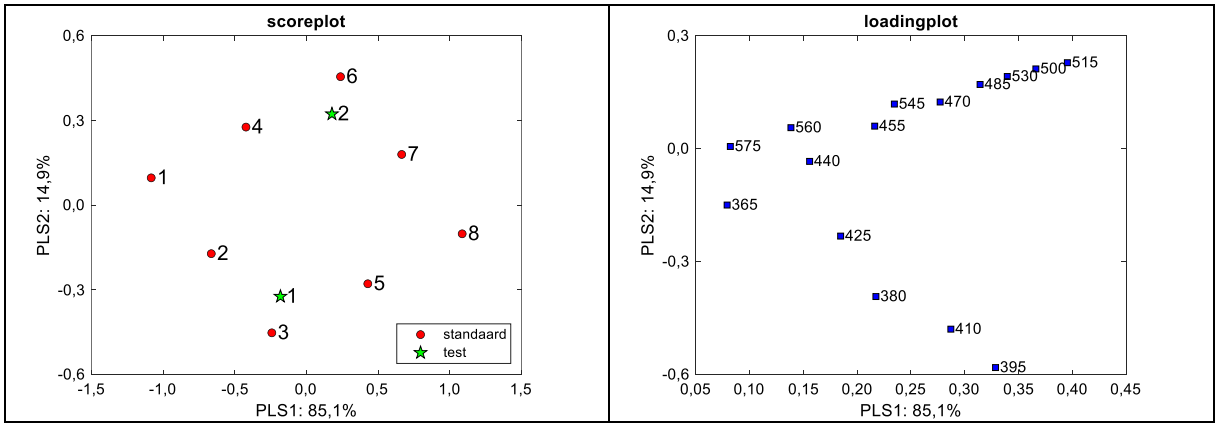
De matrix met weegfactoren ***W*** voor twee PLS-factoren is:

[0,0633	-0,1501
0,1760	-0,3932
0,2674	-0,5820
0,2366	-0,4799
0,1601	-0,2326
0,1522	-0,0340
0,2225	0,0596
0,2901	0,1237
0,3322	0,1704
0,3884	0,2117
0,4194	0,2282
0,3597	0,1916
0,2470	0,1183
0,1442	0,0556
0,0826	0,0055]

De getransponeerde ***y***-loadingsvector ***q*** voor twee PLS-factoren is:

[0,0780 0,0470]

De scorematrix ***T*** voor de kalibratie- en testset voor twee PLS-factoren staan in de volgende tabel. 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 PLS-factoren. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De getransponeerde PLS-regressievector $\boldsymbol{b}$ kan met behulp van vergelijking (15.19) worden berekend:

[-0,0018	-0,0039	-0,0052	-0,0029	0,0023	0,0110	0,0213	0,0299	0,0356
0,0422	0,0455	0,0388	0,0261	0,0146	0,0071]			

Met behulp van deze PLS-regressievector $\boldsymbol{b}$ kunnen de geschatte waarden voor de kalibratiemonsters kunnen worden berekend met vergelijking (15.21) en voor de testmonsters met vergelijking (15.28). Dit is de eenvoudigste methode, mits de berekening van de inverse $(\boldsymbol{P_a^T W_a})^{-1}$ in vergelijking (15.19) geen problemen oplevert.

De gemiddelde responsiewaarde voor de kalibratiemonsters is $\overline{\boldsymbol{y}}_{\boldsymbol{kal}} = 0{,}1400$. De geschatte waarden voor $\hat{\boldsymbol{y}}_{\mathbf{kal}}$ en $\hat{\boldsymbol{y}}_{\mathbf{test}}$, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel.

	Co-concentratie	$\boldsymbol{T}$			$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset						
1	0,060	-1,0834	0,0963		0.0600	1.33e-05	1.78e-10
2	0,080	-0,6644	-0,1724		0.0801	6.19e-05	3.84e-09
3	0,100	-0,2422	-0,4529		0.0998	-1.86e-04	3.45e-08
4	0,120	-0,4225	0,2761		0.1200	2.09e-05	4.37e-10
5	0,160	0,4271	-0,2790		0.1602	2.02e-04	4.09e-08
6	0,180	0,2368	0,4547		0.1798	-1.52e-04	2.30e-08
7	0,200	0,6634	0,1792		0.2002	1.74e-04	3.03e-08
8	0,220	1,0852	-0,1020		0.2199	-1.35e-04	1.82e-08
	testset						
1	0,110	-0,1822	-0,3240		0.1097	-3.42e-04	1.17e-07
2	0,170	0,1771	0,3225		0.1699	-1.48e-04	2.19e-08

Het kwadraat van de correlatiecoëfficiënt voor de geschatte en de juiste kobaltconcentraties voor de kalibratieset is $R^2_{\text{kal}} = 1{,}0000$. Voor de testset wordt voor de geschatte en de juiste kobaltconcentraties 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 $p = a = 2$:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}} (\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}513 \cdot 10^{-7}}{8-2}} = 1{,}59 \cdot 10^{-4}$$

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):

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

Voor de RMSEP van het SMLR-model in opgave 13.5 is gevonden $4{,}20 \cdot 10^{-4}$.
Voor de RMSEP van het PCR-model in opgave 14.5 is gevonden $2{,}63 \cdot 10^{-4}$.

De RMSEP voor het PLS-model is lager dan die voor het SMLR-model en gelijk aan die voor het PCR-model. De predictie van het PLS-model is dus beter dan die van het SMLR-model en gelijk aan die van het PCR-model.

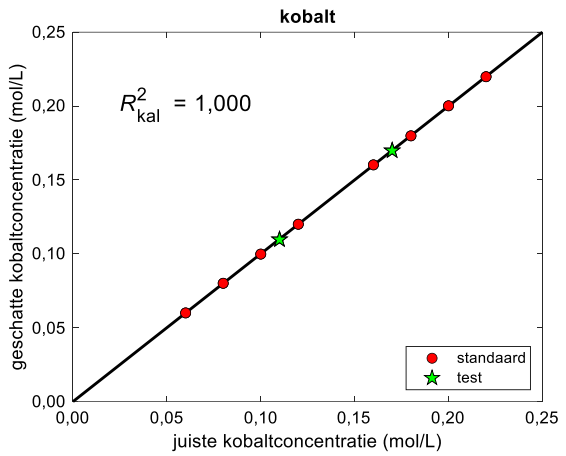
Er kan worden getest of de voorspelfout (RMSEP) van het PLS-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten.

De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout.
$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Vergelijking van $\text{RMSEP}_{\text{PLS}}$ met $\text{RMSEP}_{\text{SMLR}}$:

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

Omdat $F_{\text{SMLR/PLS}} < F_{\text{krit}}$ is er geen significant verschil tussen de varianties van de voorspelfout. De predictie van het PLS-model is dus *niet* significant beter dan die van het SMLR-model in opgave 13.5.

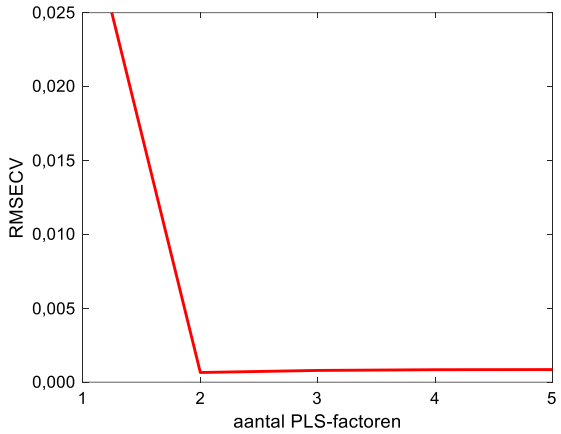


<u>Nikkel</u>

De optimale modelcomplexiteit A is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal PLS-factoren in het model. RMSECV, de verklaarde variantie voor elke PLS-factor, de cumulatieve verklaarde variantie en de percentages van de verklaarde varianties van de PLS-factoren zijn vermeld in de volgende tabel.

Aantal PLS-factoren	RMSECV	Verklaarde variantie	Cumulative verklaarde variantie	Percentage verklaarde variantie
1	0,0329	0,5002	0,5002	83,62
2	0,0007	0,0980	0,5982	99,99
3	0,0008	0,0000	0,5982	100,00
4	0,0008	0,0000	0,5982	100,00
5	0,0009	0,0000	0,5982	100,00

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal PLS-factoren 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 PLS-model.



De loadingsmatrix ***P*** voor twee PLS-factoren is:

[0,0914	0,1116
0,2498	0,2881
0,3766	0,4236
0,3269	0,3428
0,2045	0,1477
0,1604	-0,0305
0,2147	-0,1447
0,2716	-0,2308
0,3057	-0,2904
0,3551	-0,3510
0,3836	-0,3787
0,3297	-0,3209
0,2288	-0,2083
0,1360	-0,1093
0,0827	-0,0386]

De matrix met weegfactoren ***W*** voor twee PLS-factoren is:

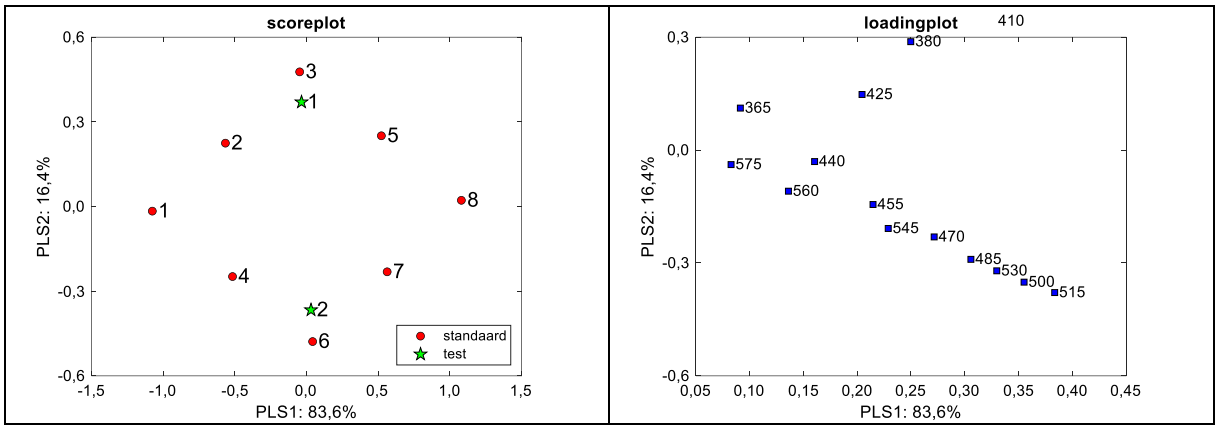
[0,1187	0,1116
0,3203	0,2881
0,4804	0,4236
0,4108	0,3428
0,2407	0,1477
0,1529	-0,0305
0,1792	-0,1448
0,2150	-0,2308
0,2346	-0,2904
0,2692	-0,3509
0,2909	-0,3786
0,2511	-0,3209
0,1778	-0,2083
0,1093	-0,1093
0,0733	-0,0386]

De getransponeerde ***y***-loadingsvector ***q*** voor twee PLS-factoren is:

[0,0724	0,0906]

De scorematrix ***T*** voor de kalibratie- en testset voor twee PLS-factoren staan in de volgende tabel. 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 PLS-factoren. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De getransponeerde PLS-regressievector $\boldsymbol{b}$ kan met behulp van vergelijking (15.19) worden berekend:

[0,0213 0,0564 0,0838 0,0699 0,0362 0,0117 0,0038 -0,0006 -0,0041
-0,0063 -0,0068 -0,0053 -0,0020 0,0004 0,0034]

Met behulp van deze PLS-regressievector $\boldsymbol{b}$ kunnen de geschatte waarden voor de kalibratiemonsters kunnen worden berekend met vergelijking (15.21) en voor de testmonsters met vergelijking (15.28). Dit is de eenvoudigste methode, mits de berekening van de inverse $(\boldsymbol{P}_a^T\boldsymbol{W}_a)^{-1}$ in vergelijking (15.19) geen problemen oplevert.

De gemiddelde responsiewaarde voor de kalibratiemonsters is $\overline{\boldsymbol{y}}_{\boldsymbol{kal}} = 0{,}1400$. De geschatte waarden voor $\hat{\boldsymbol{y}}_{\mathbf{kal}}$ en $\hat{\boldsymbol{y}}_{\mathbf{test}}$, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel.

	Ni-concentratie	$\boldsymbol{T}$			$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset						
1	0,060	-1,0756	-0,0164		0.0606	6.03e-04	3.64e-07
2	0,120	-0,5658	0,2245		0.1193	-6.51e-04	4.24e-07
3	0,180	-0,0482	0,4771		0.1797	-2.74e-04	7.49e-08
4	0,080	-0,5162	-0,2482		0.0801	1.33e-04	1.76e-08
5	0,200	0,5217	0,2507		0.2005	5.04e-04	2.54e-07
6	0,100	0,0423	-0,4785		0.0997	-2.80e-04	7.86e-08
7	0,160	0,5621	-0,2313		0.1598	-2.36e-04	5.57e-08
8	0,220	1,0797	0,0220		0.2202	2.01e-04	4.05e-08
	testset						
1	0,170	-0,0352	0,3700		0.1702	1.90e-04	3.61e-08
2	0,110	0,0312	-0,3666		0.1097	-2.52e-04	6.36e-08

Het kwadraat van de correlatiecoëfficiënt voor de geschatte en de juiste nikkelconcentraties voor de kalibratieset is $R^2_{\text{kal}} = 0{,}9999$. Voor de testset wordt voor de geschatte en de juiste nikkelconcentraties 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 $p = a = 2$:

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}}(\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}309 \cdot 10^{-6}}{8-2}} = 4{,}67 \cdot 10^{-4}$$

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

$$\sum_{i=1}^{2}(\hat{y}_i - y_i)^2 = 9{,}972 \cdot 10^{-8}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}}(\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{9{,}972 \cdot 10^{-8}}{2}} = 2{,}23 \cdot 10^{-4}$$

Voor de RMSEP van het SMLR-model in opgave 13.5 is gevonden $4{,}79 \cdot 10^{-3}$.
Voor de RMSEP van het PCR-model in opgave 14.5 is gevonden $2{,}23 \cdot 10^{-4}$.

De RMSEP voor het PLS-model is lager dan die voor het SMLR-model en gelijk aan die voor het PCR-model. De predictie van het PLS-model is dus beter dan die van het SMLR-model en gelijk aan die van het PCR-model.

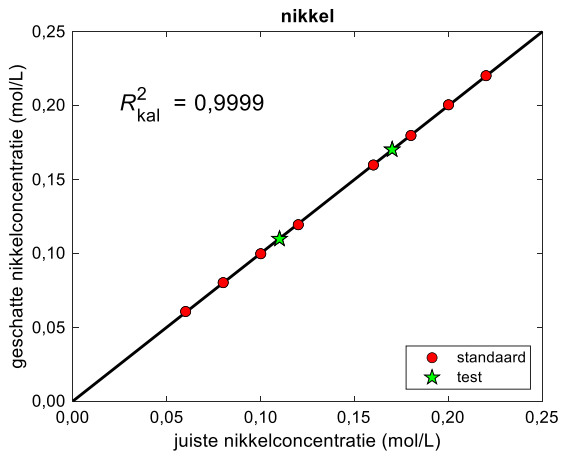
Er kan worden getest of de voorspelfout (RMSEP) van het PLS-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten.

De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout.
$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Vergelijking van $\text{RMSEP}_{\text{PLS}}$ met $\text{RMSEP}_{\text{SMLR}}$:

$$F_{SMLR/PLS} = \frac{RMSEP_{SMLR}^2}{RMSEP_{PLS}^2} = \frac{(4{,}79 \cdot 10^{-3})^2}{(2{,}23 \cdot 10^{-4})^2} = 461$$

Omdat $F_{\text{SMLR/PLS}} > F_{\text{krit}}$ is er een significant verschil tussen de varianties van de voorspelfout. De predictie van het PLS-model is dus *significant beter* dan die van het SMLR-model in opgave 13.5.



Antwoord 15.5

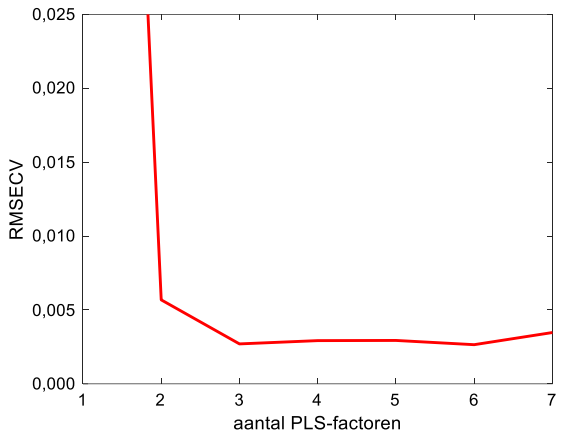
PLS is uitgevoerd met een invers kalibratiemodel na centrering van de variabelen in de ***X*** matrix en de ***y*** vector.

<u>Antraceen</u>

De optimale modelcomplexiteit *A* is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal PLS-factoren in het model. RMSECV, de verklaarde variantie voor elke PLS-factor, de cumulatieve verklaarde variantie en de percentages van de verklaarde varianties van de PLS-factoren zijn vermeld in de volgende tabel.

Aantal PLS-factoren	RMSECV	Verklaarde variantie	Cumulative verklaarde variantie	Percentage verklaarde variantie
1	0,1197	0,2905	0,2905	88,08
2	0,0057	0,0368	0,3273	99,25
3	0,0027	0,0025	0,3298	99,99
4	0,0029	0,0000	0,3298	100,00
5	0,0029	0,0000	0,3298	100,00

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal PLS-factoren 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 PLS-model.



De loadingsmatrix ***P*** voor drie PLS-factoren is:

[0,2017	-0,1841	0,1530
0,1730	-0,0172	0,0150
0,2463	0,1123	0,0676
0,3413	0,2586	0,1062
0,4730	0,4085	0,1536
0,3528	-0,1109	-0,1080
0,2838	-0,2475	-0,3280
0,3887	-0,3486	-0,4639
0,3635	-0,3170	-0,3588
0,3284	-0,5179	0,5379
0,2470	-0,3945	0,4301]

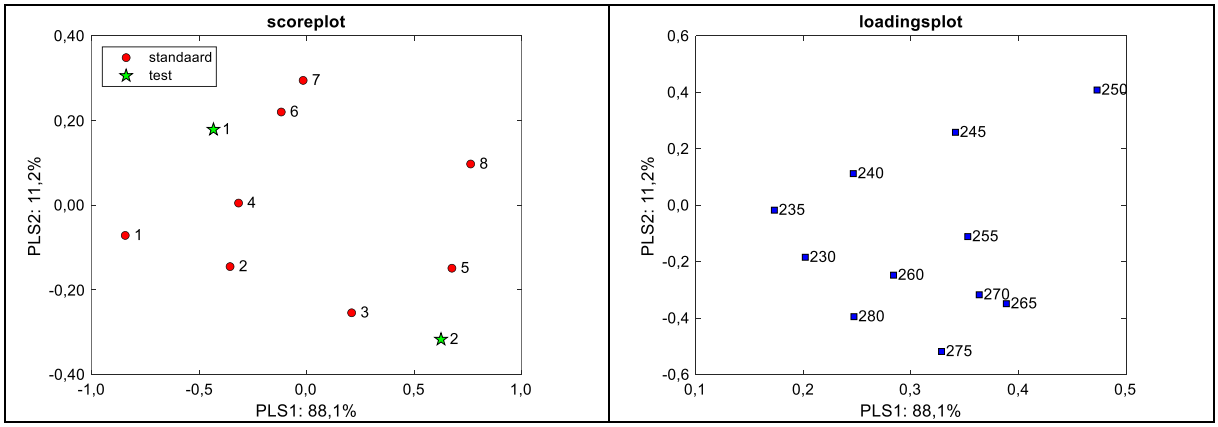
De matrix met weegfactoren $\boldsymbol{W}$ voor drie PLS-factoren is:

[0,1362	-0,1826	0,1437
0,1668	-0,0171	0,0137
0,2869	0,1131	0,0738
0,4345	0,2597	0,1035
0,6201	0,4101	0,1576
0,3126	-0,1121	-0,1217
0,1938	-0,2508	-0,3189
0,2620	-0,3534	-0,4615
0,2485	-0,3207	-0,3602
0,1447	-0,5123	0,5481
0,1071	-0,3902	0,4224]

De getransponeerde $\boldsymbol{y}$-loadingsvector $\boldsymbol{q}$ voor drie PLS-factoren is:

[0,1913 0,5410 0,0840]

De scorematrix $\boldsymbol{T}$ voor de kalibratie- en testset voor drie PLS-factoren staan in de volgende tabel. Op basis van de scorematrix $\boldsymbol{T}$ en de loadingsmatrix $\boldsymbol{P}$ kunnen respectievelijk scoreplots en loadingplots worden gemaakt. In de volgende afbeeldingen zijn deze gegeven voor de eerste twee PLS-factoren. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De getransponeerde PLS-regressievector $\boldsymbol{b}$ kan met behulp van vergelijking (15.19) worden berekend:

[-0,0344 0,0562 0,1781 0,3169 0,4746 0,0496 -0,0880 -0,1293 -0,1083
-0,1758 -0,1347]

Met behulp van deze PLS-regressievector $\boldsymbol{b}$ kunnen de geschatte waarden voor de kalibratiemonsters kunnen worden berekend met vergelijking (15.21) en voor de testmonsters met vergelijking (15.28). Dit is de eenvoudigste methode, mits de berekening van de inverse $(\boldsymbol{P}_a^T\boldsymbol{W}_a)^{-1}$ in vergelijking (15.19) geen problemen oplevert.

De gemiddelde responsiewaarde voor de kalibratiemonsters is $\bar{\boldsymbol{y}}_{kal} = 0{,}3500$. De geschatte waarden voor $\hat{\boldsymbol{y}}_{\mathbf{kal}}$ en $\hat{\boldsymbol{y}}_{\mathbf{test}}$, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel.

Nr.	*Antraceen concentratie y_i*	***T***				$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset							
1	0,150	-0,8429	-0,0712	-0,0101		0,1494	-5,82E-04	3,39E-07
2	0,200	-0,3556	-0,1448	-0,0328		0,2009	8,95E-04	8,02E-07
3	0,250	0,2096	-0,2539	-0,0350		0,2498	-2,06E-04	4,23E-08
4	0,300	-0,3161	0,0053	0,0691		0,2982	-1,80E-03	3,25E-06
5	0,400	0,6757	-0,1486	0,0339		0,4017	1,71E-03	2,93E-06
6	0,450	-0,1177	0,2204	0,0584		0,4516	1,65E-03	2,71E-06
7	0,500	-0,0159	0,2951	-0,0729		0,5005	5,03E-04	2,53E-07
8	0,550	0,7629	0,0976	-0,0107		0,5478	-2,17E-03	4,70E-06
	testset							
1	0,275	-0,4328	0,1791	-0,0373		0,2771	2,06E-03	4,24E-06
2	0,425	0,6252	-0,3168	0,0875		0,4268	1,75E-03	3,07E-06

Het kwadraat van de correlatiecoëfficiënt voor de geschatte en de juiste antraceen-concentraties voor de kalibratieset is $R^2_{\text{kal}} = 0{,}9999$. Voor de testset wordt voor de geschatte en de juiste antraceenconcentraties geen correlatiecoëfficiënt berekend omdat er slechts twee testmonsters zijn.

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

$$\sum_{i=1}^{8} (\hat{y}_i - y_i)^2 = 1{,}503 \cdot 10^{-5}$$

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

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}} (\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}503 \cdot 10^{-5}}{8-3}} = 1{,}73 \cdot 10^{-3}$$

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

$$\sum_{i=1}^{2} (\hat{y}_i - y_i)^2 = 7{,}306 \cdot 10^{-6}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}} (\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{7{,}306 \cdot 10^{-6}}{2}} = 1{,}91 \cdot 10^{-3}$$

Voor de RMSEP van het SMLR-model in opgave 13.6 is gevonden $3{,}21 \cdot 10^{-3}$.
Voor de RMSEP van het PCR-model in opgave 14.6 is gevonden $1{,}92 \cdot 10^{-3}$.

De RMSEP voor het PLS-model is lager dan die voor het SMLR-model en vrijwel gelijk aan die voor het PCR-model. De predictie van het PLS-model is dus beter dan die van het SMLR-model en vrijwel gelijk aan die van het PCR-model.

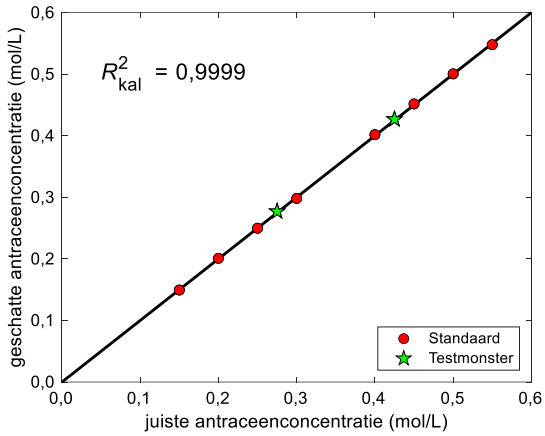
Er kan worden getest of de voorspelfout (RMSEP) van het PLS-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten.

De test kan worden uitgevoerd met een eenzijdige *F*-test op de varianties van de voorspelfout.
$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Vergelijking van $RMSEP_{PLS}$ met $RMSEP_{SMLR}$:

$$F_{SMLR/PLS} = \frac{RMSEP_{SMLR}^2}{RMSEP_{PLS}^2} = \frac{(3{,}21 \cdot 10^{-3})^2}{(1{,}92 \cdot 10^{-3})^2} = 2{,}82$$

Omdat $F_{SMLR/PLS} < F_{krit}$ is er geen significant verschil tussen de varianties van de voorspelfout. De predictie van het PLS-model is dus *niet* significant beter dan die van het SMLR-model in opgave 13.6.

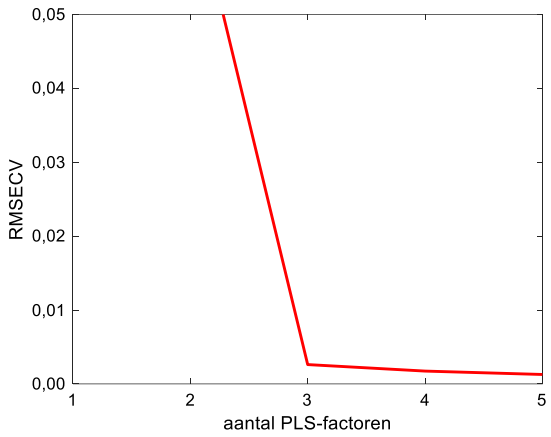


<u>Benzo(a)Antraceen</u>

De optimale modelcomplexiteit *A* is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal PLS-factoren in het model. RMSECV, de verklaarde variantie voor elke PLS-factor, de cumulatieve verklaarde variantie en de percentages van de verklaarde varianties van de PLS-factoren zijn vermeld in de volgende tabel.

Aantal PLS-factoren	RMSECV	Verklaarde variantie	Cumulative verklaarde variantie	Percentage verklaarde variantie
1	0,0562	0,3322	0,3322	94,80
2	0,0684	0,0072	0,3394	96,86
3	0,0026	0,0110	0,3503	99,99
4	0,0017	0,0000	0,3503	100,00
5	0,0013	0,0000	0,3504	100,00

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal PLS-factoren in het model en hiervan is de volgende grafiek getekend. RMSECV daalt nog maar zeer weinig na $A = 3$. Daarom wordt $A = 3$ beschouwd als de optimale modelcomplexiteit voor het PLS-model.



De loadingsmatrix ***P*** voor drie PLS-factoren is:

[0,1946	-0,2686	0,0498
0,1613	0,0723	-0,0485
0,2242	0,3475	-0,2214
0,3066	0,6724	-0,4125
0,4228	1,0246	-0,6231
0,3320	0,0476	0,0304
0,2734	-0,1857	0,2755
0,3749	-0,2679	0,3925
0,3502	-0,2570	0,3283
0,3255	-0,9407	0,1794
0,2450	-0,7250	0,1296]

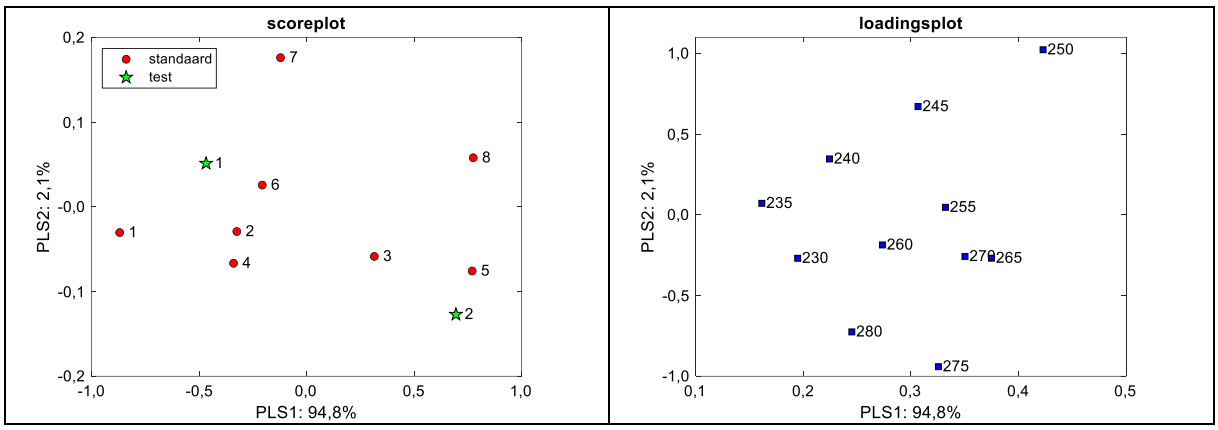
De matrix met weegfactoren ***W*** voor drie PLS-factoren is:

[0,1884	-0,1926	0,0503
0,1613	-0,0005	-0,0481
0,2246	0,0114	-0,2223
0,3082	0,0497	-0,4119
0,4255	0,0822	-0,6234
0,3351	0,0946	0,0311
0,2809	0,2300	0,2750
0,3855	0,3260	0,3929
0,3580	0,2386	0,3279
0,3037	-0,6694	0,1795
0,2277	-0,5296	0,1293]

De getransponeerde ***y***-loadingsvector ***q*** voor drie PLS-factoren is:

[0,2398	0,3596	0,3575]

De scorematrix ***T*** voor de kalibratie- en testset voor drie PLS-factoren staan in de volgende tabel. 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 PLS-factoren. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De getransponeerde PLS-regressievector $\boldsymbol{b}$ kan met behulp van vergelijking (15.19) worden berekend:

[-0,1047 0,0257 -0,0087 -0,0196 -0,0344 0,1864 0,3809 0,5376 0,4282
-0,4566 -0,3691]

Met behulp van deze PLS-regressievector $\boldsymbol{b}$ kunnen de geschatte waarden voor de kalibratiemonsters kunnen worden berekend met vergelijking (15.21) en voor de testmonsters met vergelijking (15.28). Dit is de eenvoudigste methode, mits de berekening van de inverse $(\boldsymbol{P}_a^T\boldsymbol{W}_a)^{-1}$ in vergelijking (15.19) geen problemen oplevert.

De gemiddelde responsiewaarde voor de kalibratiemonsters is $\bar{\boldsymbol{y}}_{\boldsymbol{kal}} = 0{,}3500$. De geschatte waarden voor $\hat{\boldsymbol{y}}_{\mathbf{kal}}$ en $\hat{\boldsymbol{y}}_{\mathbf{test}}$, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel.

Nr.	*Benzo(a)antraceen concentratie y_i*	$\boldsymbol{T}$				$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset							
1	0,150	-0,8689	-0,0303	0,0496		0,1485	-1,49E-03	2,23E-06
2	0,300	-0,3243	-0,0291	0,1050		0,2993	-6,57E-04	4,31E-07
3	0,450	0,3150	-0,0585	0,1336		0,4523	2,26E-03	5,10E-06
4	0,200	-0,3396	-0,0666	-0,1203		0,2016	1,64E-03	2,68E-06
5	0,500	0,7698	-0,0757	-0,0260		0,4981	-1,93E-03	3,74E-06
6	0,250	-0,2061	0,0258	-0,1670		0,2501	1,45E-04	2,10E-08
7	0,400	-0,1211	0,1763	0,0443		0,4002	2,03E-04	4,13E-08
8	0,550	0,7750	0,0580	-0,0191		0,5498	-1,60E-04	2,56E-08
	testset							
1	0,275	-0,4670	0,0514	0,0099		0,2741	-8,67E-04	7,52E-07
2	0,425	0,6944	-0,1270	0,0222		0,4305	5,48E-03	3,00E-05

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

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

$$\sum_{i=1}^{8}(\hat{y}_i - y_i)^2 = 1{,}427 \cdot 10^{-5}$$

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

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}}(\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}427 \cdot 10^{-5}}{8-3}} = 1{,}69 \cdot 10^{-3}$$

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

$$\sum_{i=1}^{2}(\hat{y}_i - y_i)^2 = 3{,}077 \cdot 10^{-5}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}}(\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{3{,}077 \cdot 10^{-5}}{2}} = 3{,}92 \cdot 10^{-3}$$

Voor de RMSEP van het SMLR-model in opgave 13.6 is gevonden $6{,}86 \cdot 10^{-2}$.
Voor de RMSEP van het PCR-model in opgave 14.6 is gevonden $3{,}92 \cdot 10^{-3}$.

De RMSEP voor het PLS-model is lager dan die voor het SMLR-model en gelijk aan die voor het PCR-model. De predictie van het PLS-model is dus beter dan die van het SMLR-model en gelijk aan die van het PCR-model.

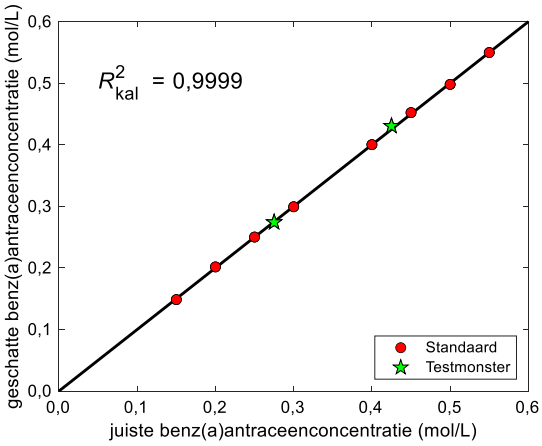
Er kan worden getest of de voorspelfout (RMSEP) van het PLS-model significant beter (lager) is dan van het SMLR-model. Dit is niet gevraagd in de opgave maar wel interessant om te weten.

De test kan worden uitgevoerd met een eenzijdige F-test op de varianties van de voorspelfout.
$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Vergelijking van $\text{RMSEP}_{\text{PLS}}$ met $\text{RMSEP}_{\text{SMLR}}$:

$$F_{SMLR/PLS} = \frac{RMSEP_{SMLR}^2}{RMSEP_{PLS}^2} = \frac{(6{,}86 \cdot 10^{-2})^2}{(3{,}92 \cdot 10^{-3})^2} = 306$$

Omdat $F_{\text{SMLR/PLS}} > F_{\text{krit}}$ is er een significant verschil tussen de varianties van de voorspelfout. De predictie van het PLS-model is dus *significant beter* dan die van het SMLR-model in opgave 13.6.



Antwoord 15.6

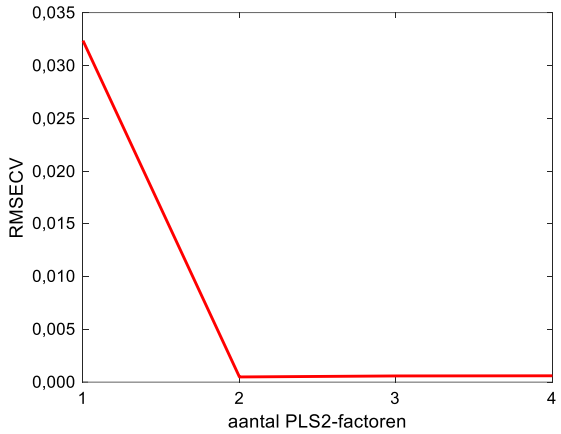
PLS2 is uitgevoerd met een invers kalibratiemodel na centrering van de variabelen in de ***X*** en ***Y*** matrix.

Vergelijkingen (15.30) en (15.31) beschrijven het PLS2-model. In dit geval is de ***Y***-matrix een 8 x 2 matrix bestaande uit twee responsievectoren. In de eerste kolom staan de kobaltconcentraties en in de tweede kolom de nikkelconcentraties.

De optimale modelcomplexiteit *A* is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal PLS-factoren in het model. RMSECV, de verklaarde variantie voor elke PLS2-factor, de cumulatieve verklaarde variantie en de percentages van de verklaarde varianties van de PLS2-factoren zijn vermeld in de volgende tabel.

Aantal PLS-factoren	RMSECV	Verklaarde variantie	Cumulative verklaarde variantie	Percentage verklaarde variantie
1	0,0324	76,9872	76,99	92,43
2	0,0005	6,2989	83,29	99,99
3	0,0006	0,0013	83,29	99,99
4	0,0006	0,0057	83,29	100,00

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal PLS-factoren 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 PLS2-model.



De loadingsmatrix ***P*** voor twee PLS-factoren is:

```
[ 0,0066    0,0182
  0,0180    0,0476
  0,0273    0,0705
  0,0238    0,0582
  0,0153    0,0282
  0,0129    0,0041
  0,0179   -0,0072
  0,0230   -0,0150
  0,0261   -0,0206
  0,0304   -0,0257
  0,0328   -0,0277
  0,0282   -0,0232
  0,0195   -0,0143
  0,0115   -0,0067
  0,0068   -0,0007]
```

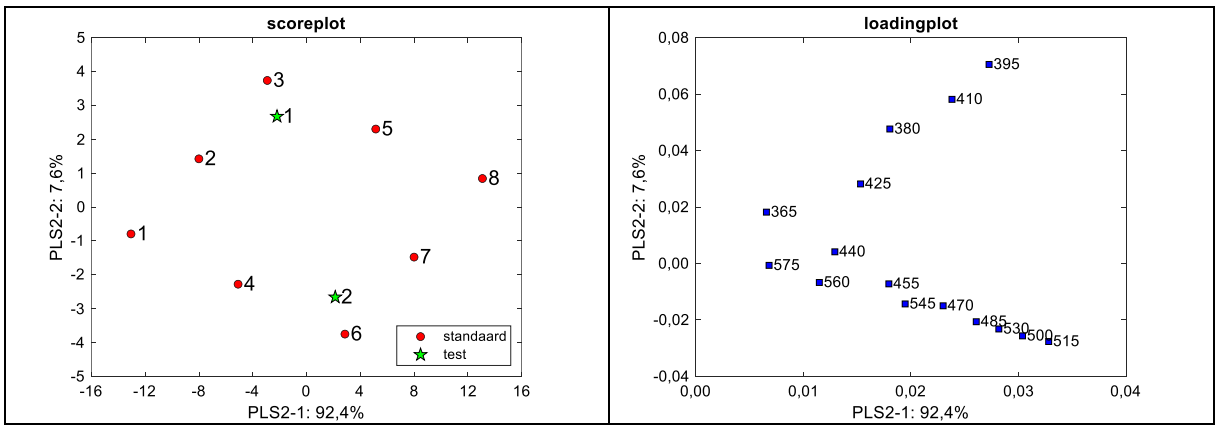
De matrix met weegfactoren $\boldsymbol{W}$ voor twee PLS-factoren is:

$$
\begin{bmatrix}
0{,}7625 & 1{,}2388 \\
2{,}1217 & 3{,}2448 \\
3{,}2233 & 4{,}8027 \\
2{,}8517 & 3{,}9608 \\
1{,}9300 & 1{,}9196 \\
1{,}8342 & 0{,}2812 \\
2{,}6825 & -0{,}4925 \\
3{,}4975 & -1{,}0218 \\
4{,}0042 & -1{,}4056 \\
4{,}6817 & -1{,}7469 \\
5{,}0558 & -1{,}8830 \\
4{,}3358 & -1{,}5805 \\
2{,}9775 & -0{,}9766 \\
1{,}7383 & -0{,}4586 \\
0{,}9958 & -0{,}0462
\end{bmatrix}
$$

De $\boldsymbol{Q}$-loadingsmatrix voor twee PLS-factoren is:

$$
\begin{bmatrix}
0{,}0065 & -0{,}0057 \\
0{,}0052 & 0{,}0147
\end{bmatrix}
$$

De scorematrix $\boldsymbol{T}$ voor de kalibratie- en testset voor twee PLS-factoren staan in de volgende tabel. Op basis van de scorematrix $\boldsymbol{T}$ en de loadingsmatrix $\boldsymbol{P}$ kunnen respectievelijk scoreplots en loadingplots worden gemaakt. In de volgende afbeeldingen zijn deze gegeven voor de eerste twee PLS-factoren. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De matrix $\boldsymbol{B}$, met regressiecoëfficiënten op basis van een PLS2-model met a = 2 PLS2-factoren, kan met behulp van een aangepaste vergelijking (15.19) worden berekend met:

$\boldsymbol{B_a} = \boldsymbol{W_a}(\boldsymbol{P_a^T}\boldsymbol{W_a})^{-1}\boldsymbol{Q_a}$ met a = 2.

$$
\begin{bmatrix}
-0{,}0018 & 0{,}0213 \\
-0{,}0039 & 0{,}0564 \\
-0{,}0052 & 0{,}0838 \\
-0{,}0029 & 0{,}0699 \\
0{,}0023 & 0{,}0362 \\
0{,}0110 & 0{,}0117 \\
0{,}0213 & 0{,}0038 \\
0{,}0299 & -0{,}0006 \\
0{,}0356 & -0{,}0041 \\
0{,}0422 & -0{,}0063 \\
0{,}0455 & -0{,}0068 \\
0{,}0388 & -0{,}0053 \\
0{,}0261 & -0{,}0020 \\
0{,}0146 & 0{,}0004 \\
0{,}0071 & 0{,}0034
\end{bmatrix}
$$

Met behulp van deze PLS2-regressiematrix $\boldsymbol{B}$ kunnen de geschatte waarden voor de *kalibratiemonsters* kunnen worden berekend met een aangepaste vergelijking (15.21),

$$\hat{\boldsymbol{Y}}_{\mathbf{kal}} = \boldsymbol{X}_{\mathbf{0}}\boldsymbol{B}_{\boldsymbol{a}} + \mathbf{1}\bar{\boldsymbol{y}}_{\mathbf{kal}}^{\boldsymbol{T}}$$

waarin $\hat{\boldsymbol{Y}}_{\mathbf{kal}}$ een n_{kal} x m matrix is met geschatte responsiewaarden, $\boldsymbol{X}_{\mathbf{0}}$ een gecentreerde $n_{\text{kal}} \times p$-datamatrix is, $\boldsymbol{B}_{\boldsymbol{a}}$ een $p \times m$ matrix is met regressiecoëfficiënten op basis van een PLS2-model met a PLS2-factoren, $\mathbf{1}$ een n_{kal} x 1 kolomvector met enen, $\bar{\boldsymbol{y}}_{\mathbf{kal}}^{\boldsymbol{T}}$ een 1 x m rijvector is met kolomgemiddelden van de $\boldsymbol{Y}_{\mathbf{kal}}$-matrix, n_{kal} het aantal kalibratiemonsters, m het aantal componenten, p het aantal variabelen en a het aantal PLS2-factoren.

En voor de *testmonsters* met een aangepaste vergelijking (15.28),

$$\hat{\boldsymbol{Y}}_{\mathbf{test}} = \boldsymbol{X}_{\mathbf{0,test}}\boldsymbol{B}_{\boldsymbol{a}} + \mathbf{1}\bar{\boldsymbol{y}}_{\mathbf{kal}}^{\boldsymbol{T}}$$

waarin $\hat{\boldsymbol{Y}}_{\mathbf{test}}$ een n_{test} x m matrix is met geschatte responsiewaarden, $\boldsymbol{X}_{\mathbf{0,test}}$ een gecentreerde $n_{\text{test}} \times p$-datamatrix is en n_{test} het aantal testmonsters.

Dit is de eenvoudigste methode, mits de berekening van de inverse $(\boldsymbol{P}_{\boldsymbol{a}}^{\boldsymbol{T}}\boldsymbol{W}_{\boldsymbol{a}})^{-1}$ in vergelijking (15.19) geen problemen oplevert.

De gemiddelde responsiewaarde voor de kalibratiemonsters is $\bar{\boldsymbol{y}}_{\mathbf{kal}}^{\boldsymbol{T}} = [0{,}1400 \quad 0{,}1400]$. Hierna worden de resultaten voor kobalt en nikkel apart beschreven.

<u>Kobalt</u>

De geschatte waarden voor de kobaltconcentraties in de kalibratie- en testmonsters, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel. De geschatte waarden voor de kobaltconcentraties in de kalibratieset staan in de eerste kolom van $\hat{\boldsymbol{Y}}_{\mathbf{kal}}$ en die voor de testset staan in de eerste kolom van $\hat{\boldsymbol{Y}}_{\mathbf{test}}$. De scorematrix voor de kalibratie- ($\boldsymbol{T}_{\mathbf{kal}}$) en testset ($\boldsymbol{T}_{\mathbf{test}}$) zijn gemeenschappelijk met die voor nikkel. $\boldsymbol{T}_{\mathbf{kal}}$ is de $\boldsymbol{T}_{\boldsymbol{a}}$-matrix in de vergelijkingen (15.30) en (15.31) voor het PLS2-model.

	Co-concentratie	$\boldsymbol{T}$			$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset						
1	0,060	-13,0595	-0,7947		0,0600	1,33E-05	1,78E-10
2	0,080	-8,0096	1,4228		0,0801	6,19E-05	3,84E-09
3	0,100	-2,9200	3,7375		0,0998	-1,86E-04	3,45E-08
4	0,120	-5,0930	-2,2787		0,1200	2,09E-05	4,38E-10
5	0,160	5,1484	2,3025		0,1602	2,02E-04	4,09E-08
6	0,180	2,8549	-3,7522		0,1798	-1,52E-04	2,30E-08
7	0,200	7,9968	-1,4787		0,2002	1,74E-04	3,02E-08
8	0,220	13,0819	0,8416		0,2199	-1,35E-04	1,82E-08
	testset						
1	0,110	-2,1958	2,6739		0,1097	-3,42E-04	1,17E-07
2	0,170	2,1352	-2,6611		0,1699	-1,48E-04	2,19E-08

Het kwadraat van de correlatiecoëfficiënt voor de geschatte en de juiste kobaltconcentraties voor de kalibratieset is $R_{\text{kal}}^2 = 1{,}0000$. Voor de testset wordt voor de geschatte en de juiste kobaltconcentraties geen correlatiecoëfficiënt berekend omdat er slechts twee testmonsters zijn.

De som van de kwadraten van de residuen voor de kalibratieset en de testset, RMSEC en RMSEP, berekend voor kobalt op basis van het PLS2-model, zijn identiek aan die voor het PLS1-model in opgave 15.4.

Nikkel

De geschatte waarden voor de nikkelconcentraties in de kalibratie- en testmonsters, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel. De geschatte waarden voor de nikkelconcentraties in de kalibratieset staan in de tweede kolom van $\hat{\boldsymbol{Y}}_{\mathbf{kal}}$ en die voor de testset staan in de tweede kolom van $\hat{\boldsymbol{Y}}_{\mathbf{test}}$. De scorematrix voor de kalibratie- ($\boldsymbol{T}_{\mathbf{kal}}$) en testset ($\boldsymbol{T}_{\mathbf{test}}$) gelijk zijn gemeenschappelijk met die voor kobalt. $\boldsymbol{T}_{\mathbf{kal}}$ is de $\boldsymbol{T}_{\boldsymbol{a}}$-matrix in de vergelijkingen (15.30) en (15.31) voor het PLS2-model.

	Ni-concentratie	$\boldsymbol{T}$			$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset						
1	0,060	-13,0595	-0,7947		0,0606	6,03E-04	3,64E-07
2	0,120	-8,0096	1,4228		0,1193	-6,51E-04	4,24E-07
3	0,180	-2,9200	3,7375		0,1797	-2,74E-04	7,49E-08
4	0,080	-5,0930	-2,2787		0,0801	1,33E-04	1,76E-08
5	0,200	5,1484	2,3025		0,2005	5,04E-04	2,54E-07
6	0,100	2,8549	-3,7522		0,0997	-2,80E-04	7,86E-08
7	0,160	7,9968	-1,4787		0,1598	-2,36E-04	5,57E-08
8	0,220	13,0819	0,8416		0,2202	2,01E-04	4,05E-08
	testset						
1	0,170	-2,1958	2,6739		0,1702	1,90E-04	3,61E-08
2	0,110	2,1352	-2,6611		0,1097	-2,52E-04	6,36E-08

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

De som van de kwadraten van de residuen voor de kalibratieset en de testset, RMSEC en RMSEP, berekend voor nikkel op basis van het PLS2-model, zijn identiek aan die voor het PLS1-model in opgave 15.4.

Antwoord 15.7

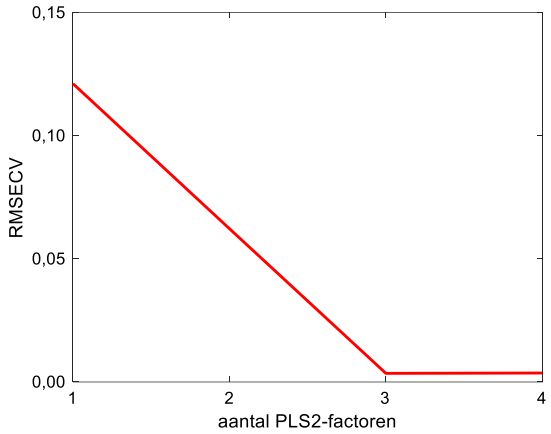
PLS2 is uitgevoerd met een invers kalibratiemodel na centrering van de variabelen in de $\boldsymbol{X}$ en $\boldsymbol{Y}$ matrix.

Vergelijkingen (15.30) en (15.31) beschrijven het PLS2-model. In dit geval is de $\boldsymbol{Y}$-matrix een 8 x 2 matrix met twee responsievectoren. In de eerste kolom staan de antraceenconcentraties en in de tweede kolom de benzo(a)antraceenconcentraties.

De optimale modelcomplexiteit A is bepaald door een leave-one-out (LOO) kruisvalidatie herhaald uit te voeren met een oplopend aantal PLS-factoren in het model. RMSECV, de verklaarde variantie voor elke PLS2-factor, de cumulatieve verklaarde variantie en de percentages van de verklaarde varianties van de PLS2-factoren zijn vermeld in de volgende tabel.

Aantal PLS-factoren	RMSECV	Verklaarde variantie	Cumulative verklaarde variantie	Percentage verklaarde variantie
1	0,1211	1,1896	1,1896	93,93
2	0,0623	0,0748	1,2643	99,83
3	0,0035	0,0021	1,2665	100,00
4	0,0036	0,0000	1,2665	100,00

Bij de LOO kruisvalidatie is RMSECV berekend als functie van het aantal PLS-factoren in het model en hiervan is de volgende grafiek getekend. Het minimum van deze curve ligt bij *A* = 3. Dit is de optimale modelcomplexiteit voor het PLS2-model.



De loadingsmatrix ***P*** voor drie PLS-factoren is:

[0,1037	-0,0601	0,1591
0,0848	0,0458	0,0128
0,1166	0,1599	0,0685
0,1586	0,2928	0,1091
0,2182	0,4403	0,1583
0,1751	0,0410	-0,1228
0,1456	-0,0765	-0,3581
0,1996	-0,1114	-0,5063
0,1864	-0,0980	-0,3930
0,1754	-0,2489	0,5668
0,1321	-0,1906	0,4538]

De matrix met weegfactoren ***W*** voor drie PLS-factoren is:

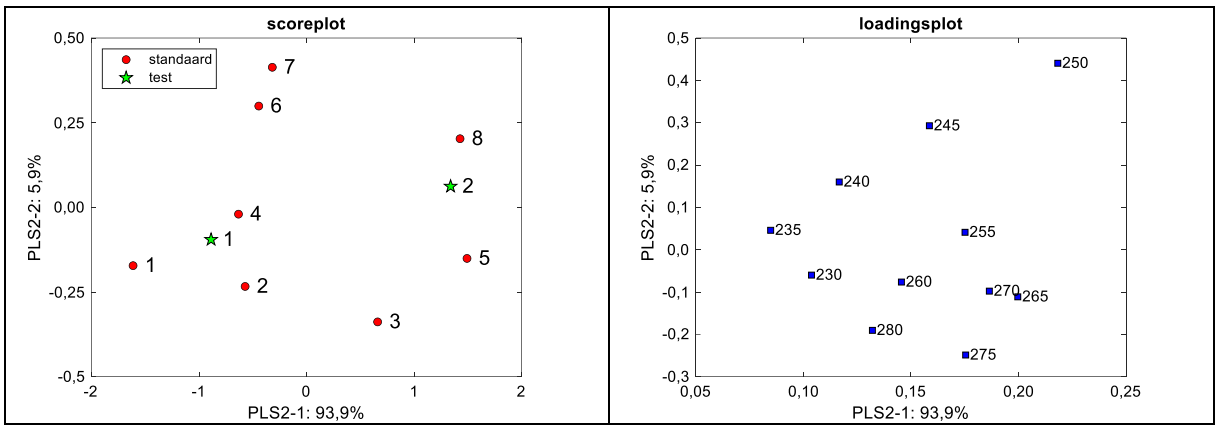
[0,4064	-0,1350	0,1391
0,2922	0,1045	0,0102
0,3673	0,3655	0,0604
0,4713	0,6688	0,0944
0,6348	1,0060	0,1380
0,6191	0,0917	-0,1071
0,5478	-0,1791	-0,3120
0,7534	-0,2607	-0,4413
0,7039	-0,2286	-0,3414
0,7527	-0,5596	0,4925
0,5689	-0,4286	0,3964]

De ***Q***-loadingsmatrix voor drie PLS-factoren is:

[0,0771	0,4381	0,1261
0,0458	0,2252	-0,3714]

De scorematrix ***T*** voor de kalibratie- en testset voor drie PLS-factoren staan in de volgende tabel. 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 PLS-factoren. Deze plots zijn informatief en spelen geen rol bij de uitwerking van deze opgave.



De matrix $\boldsymbol{B}$, met regressiecoëfficiënten op basis van een PLS2-model met a PLS2-factoren, kan met behulp van een aangepaste vergelijking (15.19) worden berekend met:

$\boldsymbol{B_a} = \boldsymbol{W_a}(\boldsymbol{P_a^T}\boldsymbol{W_a})^{-1}\boldsymbol{Q_a}$ met $a = 3$.

[-0,0335	-0,1061	0,3201
0,0562	0,0261	0,0537
0,1776	-0,0079	0,0366
0,3171	-0,0200	-0,0065
0,4742	-0,0342	-0,0347
0,0507	0,1854	0,0135
-0,0887	0,3819	-0,1421
-0,1295	0,5374	-0,2042
-0,1080	0,4283	-0,1150
-0,1768	-0,4555	0,9749
-0,1340	-0,3698	0,7666]

Met behulp van deze PLS2-regressiematrix $\boldsymbol{B}$ kunnen de geschatte waarden voor de *kalibratiemonsters* kunnen worden berekend met een aangepaste vergelijking (15.21),

$$\widehat{\boldsymbol{Y}}_{\mathbf{kal}} = \boldsymbol{X_0}\boldsymbol{B_a} + \mathbf{1}\overline{\boldsymbol{y}}_{\mathbf{kal}}^{\boldsymbol{T}}$$

waarin $\widehat{\boldsymbol{Y}}_{\mathbf{kal}}$ een n_{kal} x m matrix is met geschatte responsiewaarden, $\boldsymbol{X_0}$ een gecentreerde $n_{\text{kal}} \times p$-datamatrix is, $\boldsymbol{B_a}$ een $p \times m$ matrix is met regressiecoëfficiënten op basis van een PLS2-model met a PLS2-factoren, $\mathbf{1}$ een n_{kal} x 1 kolomvector met enen, $\overline{\boldsymbol{y}}_{\mathbf{kal}}^{\boldsymbol{T}}$ een 1 x m rijvector is met kolomgemiddelden van de $\boldsymbol{Y}_{\mathbf{kal}}$-matrix, n_{kal} het aantal kalibratiemonsters, m het aantal componenten, p het aantal variabelen en a het aantal PLS2-factoren.

En voor de *testmonsters* met een aangepaste vergelijking (15.28),

$$\widehat{\boldsymbol{Y}}_{\mathbf{test}} = \boldsymbol{X}_{\mathbf{0,test}}\boldsymbol{B_a} + \mathbf{1}\overline{\boldsymbol{y}}_{\mathbf{kal}}^{\boldsymbol{T}}$$

waarin $\widehat{\boldsymbol{Y}}_{\mathbf{test}}$ een n_{test} x m matrix is met geschatte responsiewaarden, $\boldsymbol{X}_{\mathbf{0,test}}$ een gecentreerde $n_{\text{test}} \times p$-datamatrix is en n_{test} het aantal testmonsters.

Dit is de eenvoudigste methode, mits de berekening van de inverse $(\boldsymbol{P_a^T}\boldsymbol{W_a})^{-1}$ in vergelijking (15.19) geen problemen oplevert.

De gemiddelde responsiewaarde voor de kalibratiemonsters is $\bar{\boldsymbol{y}}_{\mathbf{kal}}^{T} = [0{,}3500 \quad 0{,}3500]$. Hierna worden de resultaten voor antraceen en benzo(a)antraceen apart beschreven.

<u>Antraceen</u>

De geschatte waarden voor de antraceenconcentraties in de kalibratie- en testmonsters, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel. De geschatte waarden voor de antraceenconcentraties in de kalibratieset staan in de eerste kolom van $\hat{\boldsymbol{Y}}_{\mathbf{kal}}$ en die voor de testset staan in de eerste kolom van $\hat{\boldsymbol{Y}}_{\mathbf{test}}$. De scorematrix voor de kalibratie- ($\boldsymbol{T}_{\mathbf{kal}}$) en testset ($\boldsymbol{T}_{\mathbf{test}}$) zijn gemeenschappelijk met die voor de benzo(a)antraceen-concentraties. $\boldsymbol{T}_{\mathbf{kal}}$ is de $\boldsymbol{T}_{a}$-matrix in de vergelijkingen (15.30) en (15.31) voor het PLS2-model.

Nr.	*Antraceen concentratie y_i*	***T***				$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset							
1	0,150	-1,6131	-0,1722	-0,0093		0,1494	-5,80E-04	3,37E-07
2	0,200	-0,5712	-0,2336	-0,0305		0,2009	9,03E-04	8,15E-07
3	0,250	0,6608	-0,3383	-0,0328		0,2498	-2,06E-04	4,23E-08
4	0,300	-0,6328	-0,0202	0,0644		0,2982	-1,82E-03	3,32E-06
5	0,400	1,4919	-0,1509	0,0315		0,4017	1,72E-03	2,95E-06
6	0,450	-0,4443	0,2990	0,0548		0,4517	1,66E-03	2,75E-06
7	0,500	-0,3184	0,4135	-0,0680		0,5005	5,03E-04	2,53E-07
8	0,550	1,4272	0,2026	-0,0101		0,5478	-2,17E-03	4,72E-06
	testset							
1	0,275	-0,8877	-0,0953	-0,0258		0,2771	2,07E-03	4,27E-06
2	0,425	1,3394	0,0619	0,0682		0,4268	1,75E-03	3,08E-06

De resultaten die voor de antraceenconcentraties worden verkregen met het PLS2-model worden hierna gedetailleerd beschreven. Ze zijn echter vrijwel gelijk aan die voor het PLS1-model in opgave 15.5.

Het kwadraat van de correlatiecoëfficiënt voor de geschatte en de juiste antraceen-concentraties voor de kalibratieset is $R_{\text{kal}}^2 = 0{,}9999$. Voor de testset wordt voor de geschatte en de juiste antraceenconcentraties geen correlatiecoëfficiënt berekend omdat er slechts twee testmonsters zijn.

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

$$\sum_{i=1}^{8} (\hat{y}_i - y_i)^2 = 1{,}518 \cdot 10^{-5}$$

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

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}} (\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}518 \cdot 10^{-5}}{8-3}} = 1{,}74 \cdot 10^{-3}$$

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

$$\sum_{i=1}^{2} (\hat{y}_i - y_i)^2 = 7{,}348 \cdot 10^{-6}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}}(\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{7{,}348 \cdot 10^{-6}}{2}} = 1{,}92 \cdot 10^{-3}$$

Voor de RMSEP van het SMLR-model in opgave 13.6 is gevonden $3{,}21 \cdot 10^{-3}$.
Voor de RMSEP van het PCR-model in opgave 14.6 is gevonden $1{,}92 \cdot 10^{-3}$.
Voor de RMSEP van het PLS1-model in opgave 15.5 is gevonden $1{,}91 \cdot 10^{-3}$.

De RMSEP voor het PLS2-model is lager dan die voor het SMLR-model, gelijk aan die voor het PCR-model en vrijwel gelijk aan die voor het PLS1-model. De predictie van het PLS2-model is dus beter dan die van het SMLR-model, gelijk aan die voor het PCR-model en vrijwel gelijk aan die voor het PLS1-model.

Er kan worden getest of de voorspelfout (RMSEP) van het PLS2-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.
$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Vergelijking van $\text{RMSEP}_{\text{PLS2}}$ met $\text{RMSEP}_{\text{SMLR}}$:

$$F_{SMLR/PLS2} = \frac{RMSEP_{SMLR}^2}{RMSEP_{PLS2}^2} = \frac{(3{,}21 \cdot 10^{-3})^2}{(1{,}92 \cdot 10^{-3})^2} = 2{,}80$$

Omdat $F_{\text{SMLR/PLS2}} < F_{\text{krit}}$ is er geen significant verschil tussen de varianties van de voorspelfout. De predictie van het PLS2-model is dus *niet* significant beter dan die van het SMLR-model in opgave 13.6.

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

De geschatte waarden voor de benzo(a)antraceenconcentraties in de kalibratie- en testmonsters, de bijbehorende waarden voor de residuen en de kwadraten van de residuen staan in de volgende tabel. De geschatte waarden voor de benzo(a)antraceenconcentraties in de kalibratieset staan in de tweede kolom van $\hat{\boldsymbol{Y}}_{\textbf{kal}}$ en die voor de testset staan in de tweede kolom van $\hat{\boldsymbol{Y}}_{\textbf{test}}$. De scorematrix voor de kalibratie- ($\boldsymbol{T}_{\textbf{kal}}$) en testset ($\boldsymbol{T}_{\textbf{test}}$) zijn gemeenschappelijk met die voor antraceen. $\boldsymbol{T}_{\textbf{kal}}$ is de $\boldsymbol{T}_{\boldsymbol{a}}$-matrix in de vergelijkingen (15.30) en (15.31) voor het PLS2-model.

Nr.	*Benzo(a)antraceen concentratie y_i*	***T***				$\hat{y}_i$	$r_i = (y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
	kalibratieset							
1	0,150	-1,6131	-0,1722	-0,0093		0,1485	-1,50E-03	2,24E-06
2	0,300	-0,5712	-0,2336	-0,0305		0,2993	-6,66E-04	4,43E-07
3	0,450	0,6608	-0,3383	-0,0328		0,4523	2,26E-03	5,13E-06
4	0,200	-0,6328	-0,0202	0,0644		0,2017	1,66E-03	2,75E-06
5	0,500	1,4919	-0,1509	0,0315		0,4981	-1,94E-03	3,77E-06
6	0,250	-0,4443	0,2990	0,0548		0,2501	1,30E-04	1,69E-08
7	0,400	-0,3184	0,4135	-0,0680		0,4002	2,09E-04	4,35E-08
8	0,550	1,4272	0,2026	-0,0101		0,5498	-1,57E-04	2,46E-08
	testset							
1	0,275	-0,8877	-0,0953	-0,0258		0,2741	-8,73E-04	7,61E-07
2	0,425	1,3394	0,0619	0,0682		0,4305	5,47E-03	2,99E-05

De resultaten die voor de benzo(a)antraceenconcentraties worden verkregen met het PLS2-model worden hierna gedetailleerd beschreven. Ze zijn echter vrijwel gelijk aan die voor het PLS1-model in opgave 15.5.

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

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

$$\sum_{i=1}^{8}(\hat{y}_i - y_i)^2 = 1{,}441 \cdot 10^{-5}$$

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

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n_{\text{kal}}}(\hat{y}_i - y_i)^2}{n_{\text{kal}} - p}} = \sqrt{\frac{1{,}441 \cdot 10^{-5}}{8-3}} = 1{,}70 \cdot 10^{-3}$$

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

$$\sum_{i=1}^{2}(\hat{y}_i - y_i)^2 = 3{,}069 \cdot 10^{-5}$$

RMSEP kan worden berekend met (12.15):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}}(\hat{y}_i - y_i)^2}{n_{\text{test}}}} = \sqrt{\frac{3{,}069 \cdot 10^{-5}}{2}} = 3{,}92 \cdot 10^{-3}$$

Voor de RMSEP van het SMLR-model in opgave 13.6 is gevonden $6{,}86 \cdot 10^{-2}$.
Voor de RMSEP van het PCR-model in opgave 14.6 is gevonden $3{,}92 \cdot 10^{-3}$.
Voor de RMSEP van het PLS1-model in opgave 15.5 is gevonden $3{,}92 \cdot 10^{-3}$.

De RMSEP voor het PLS2-model is lager dan die voor het SMLR-model en gelijk aan die voor het PCR- en het PLS1-model. De predictie van het PLS-model is dus beter dan die van het SMLR-model en gelijk aan die van het PCR- en het PLS1-model.

Er kan worden getest of de voorspelfout (RMSEP) van het PLS2-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.
$F_{\text{krit}} = F_{(0{,}05;2;2)} = 19{,}00$ (zie tabel 4 van Bijlage 1)

Vergelijking van $\text{RMSEP}_{\text{PLS2}}$ met $\text{RMSEP}_{\text{SMLR}}$:

$$F_{SMLR/PLS2} = \frac{RMSEP^2_{SMLR}}{RMSEP^2_{PLS2}} = \frac{(6{,}86 \cdot 10^{-2})^2}{(3{,}92 \cdot 10^{-3})^2} = 306$$

Omdat $F_{\text{SMLR/PLS2}} > F_{\text{krit}}$ is er een significant verschil tussen de varianties van de voorspelfout. De predictie van het PLS2-model is dus *significant beter* dan die van het SMLR-model in opgave 13.6.