Hoofdstuk

1

Inleiding

Chemometrics is the chemical disciplin that uses mathematical and statistical methods:
to design and select optimal measurement procedures and experiments, and
to provide maximum chemical information by analyzing chemical data.

De analytische chemie is de laatste 60 jaar sterk veranderd. Vóór 1960 werden de meeste chemische problemen opgelost door middel van eenvoudige nat-chemische analysemethodes: volumetrie en gravimetrie. Deze methodes leveren directe informatie over de concentratie of hoeveelheid van een te bepalen component na eenvoudig rekenwerk.

Tegenwoordig wordt er gebruik gemaakt van analyseapparatuur die grote hoeveelheden data geven. Technieken als NMR, UV/VIS, FTIR, MS, AAS, ICP, HPLC, GC, LC-MS, GC-MS, leveren spectra en chromatogrammen, waaruit informatie over de aard en hoeveelheid (concentratie) van de te bepalen componenten moet worden gehaald. Een IR-spectrum levert bijvoorbeeld al gauw 2000 datapunten en een GC-MS analyse geeft voor een enkele run al 600.000 datapunten. Vaak worden de kwalitatieve en kwantitatieve analysemethodes gebaseerd op gegevens uit verschillende delen van de spectra of chromatogrammen. Als meetgegevens van meerdere variabelen worden gebruikt dan spreek je van multivariate data analyse.

De software van veel analyseapparatuur maakt gebruik van multivariate data analyse technieken om meetresultaten te berekenen. Denk bijvoorbeeld aan de FTIR: als je een IR-spectrum van een onbekende stof hebt opgenomen, dan kun je in de bibliotheek van het apparaat zoeken in de al bekende spectra en zo bepalen welk referentiespectrum het meest lijkt op het spectrum van de onbekende stof. Dit vergelijken van de spectra op gelijkenis doet het apparaat voor je, op basis van multivariate data analysetechnieken. Het apparaat geeft dan een lijstje met mogelijke stoffen en geeft ook aan hoe goed de gelijkenis is (hit percentage).

Multivariate data analyse kan ook erg handig zijn voor monsters met complexe matrices, waarbij het isoleren van de te bepalen componenten arbeidsintensief, tijdrovend en dus kostbaar is. De invloed van storingen door de aanwezigheid van andere componenten is door een goede proefopzet en computerberekeningen te elimineren.

Multivariate data analyse wordt dus toegepast om objecten (monsters, componenten of materialen) te groeperen en classificeren of relaties tussen verschillende analytische data te beschrijven. Enkele voorbeelden:

- Het groeperen (clusteren) van monsters op basis van vergelijkbare chemische samenstelling;
- Het classificeren van monsters op basis van analytische data (spectrum, chromatogram of elementsamenstelling) op basis van bekende clusters;
- Kalibratie van een chemische component op basis van een compleet spectrum of van meerdere componenten tegelijk op basis van meercomponentenanalyse.
- Selectie van geschikte katalysatoren voor een synthese op basis van dipoolmoment, enthalpie, bindingsenergie, etc.

Ook bij het sorteren van afval kan gebruik gemaakt worden van multivariate data analysetechnieken om afval geautomatiseerd te sorteren. Materialen worden dan aan de hand van infraroodspectra. herkend, waarbij de robots die het afval sorteren een leerproces doorlopen. Aan de hand van spectra van bijvoorbeeld verschillende plasticsoorten wordt een model opgesteld en met dit model kunnen dan nieuwe objecten worden geclassificeerd.

Een ander voorbeeld is het controleren van de productkwaliteit van bijvoorbeeld polyester garens. Voor toepassing in autobanden als versterkend element is het belangrijk om te weten in hoeverre het materiaal krimpt tijdens de hittebepaling bij het fabriceren van de banden. De afname in lengte van het garen is een maat hiervoor, maar dit is een tijdrovende analyse. De krimp kan ook geschat worden uit de Raman-spectra, omdat er een relatie is tussen het spectrum en materiaaleigenschappen. Aan de hand van garens met bekende krimp en de bijbehorende Raman-spectra kan er een model opgesteld worden voor deze relatie. Op basis van dit model kan dan tijdens het productieproces gecontroleerd worden of de kwaliteit van het garen nog voldoet aan de gestelde eisen door een Raman-spectrum op te nemen.

Inhoud

In dit dictaat wordt eerst ingegaan op variabelen en relaties tussen variabelen. Vervolgens komen diverse voorbewerkingen op de datamatrix aan bod die zeer waarschijnlijk nodig zijn bij het uitvoeren van de verschillende data analysetechnieken.
Daarna komen de twee multivariabele data analysetechnieken, te weten clusteranalyse en principale componentenanalyse, aan bod. Bij alle hoofdstukken wordt ingegaan op hoe dit in BlueSky statistics of RStudio uitgevoerd wordt (korte beschrijving met screenshots) en zijn er datasets om dit mee te oefenen.

Datasets

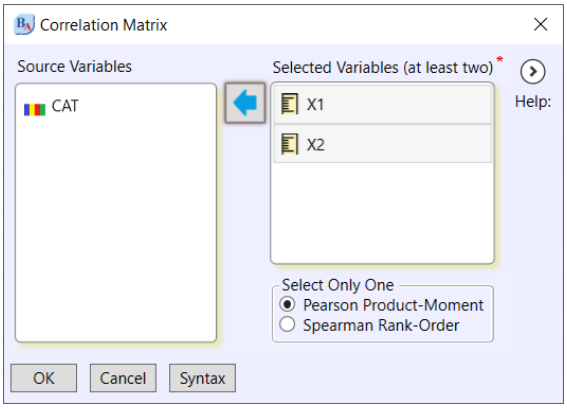
Alle oefendatasets uit het dictaat zijn als excel-bestand op Blackboard terug te vinden, deze zijn in BlueSky statistics te openen. Het gaat om de volgende datasets:

- Voorbeeld twee variabelen
- Archeologische dataset groep 1 en 2
- FeTi dataset
- Archeologische dataset
- Levensmiddelenpatroon dataset
- Peas onderzoek
- Iris dataset 75 monsters
- Wijn dataset

2.5 Uitvoering met BlueSky Statistics

Correlatie

- Open de dataset **Voorbeeld twee variabelen.sav** .
- Klik in de menublak op"**Analysis**", "**Summary Analysis**" en dan "**Correlation Matrix**".
- Selecteer de variabelen in het linkervenster en klik op de [**pijl**] voor "**Selected Variables**".



- Klik op [**Ok**]. In de output verschijnt dan een tabel met de correlatie.

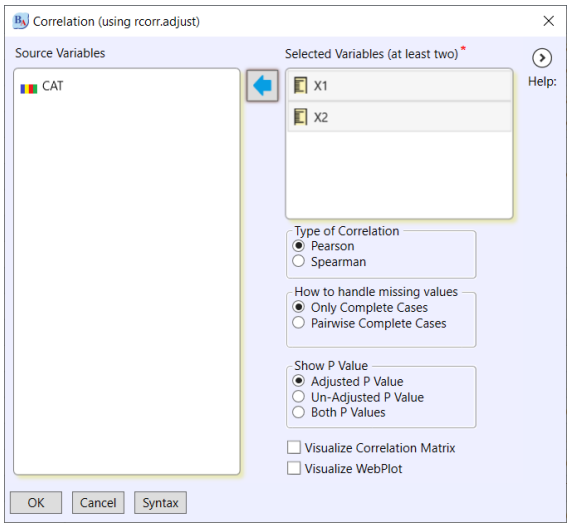
Correlation Matrix

	X1	X2
X1	1	-0.0354
X2	-0.0354	1

Correlatie inclusief significantie

- Klik in de menublak op"**Analysis**", "**Summary Analysis**" en dan "**Correlation Test (Multi-Variable)**".

- Selecteer de variabelen in het linkervenster en klik op de [**pijl**] voor "**Selected Variables**".



- Klik op [**Ok**]. In de output verschijnt dan een tabel met de correlatie. De Adj-P is groter dan 0,05, dus de correlatie is niet significant

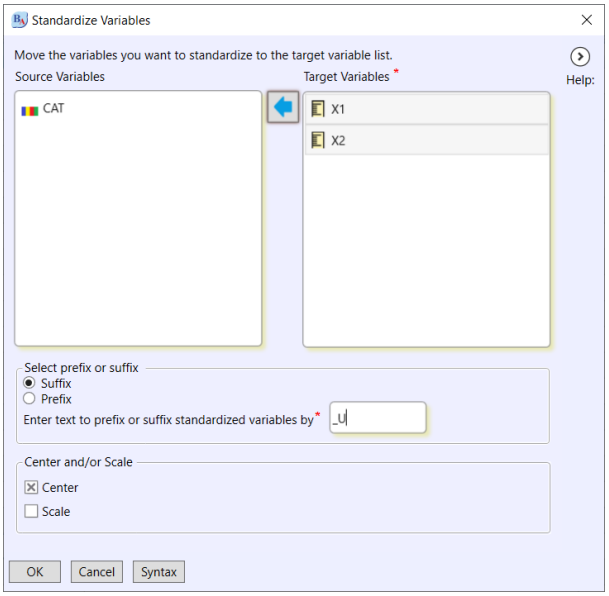
Pearson correlation

		X1	X2
X1	Correlation	1	-0.0354
	Adj-P		0.8636
	n	26	26
X2	Correlation	-0.0354	1
	Adj-P	0.8636	
	n	26	26

3.5 Uitvoering met BlueSky Statistics

Transformatie van de dataset

- Klik in de menublak op"**Data**" en dan "**Standardize variable(s)**".
- Selecteer de variabelen in het linkervenster en klik op de [**pijl**] voor "**Target Variables**".
- Selecteer Suffix als je een toevoeging achter de variabele naam wilt doen of Prefix als je een toevoeging voor de variabele naam wilt doen en geef op wat er moet worden toegevoegd.
- Selecteer "**Center**" voor een U-transformatie en zowel "**Center**" als "**Scale**" voor Z-transformatie.



- In de dataset zijn nu extra kolommen verschenen met de waarden voor elk object op basis van de U-transformatie of de Z-transformatie.
- Het is ook mogelijk om elke meetwaarde alleen te delen door de standaarddeviatie door alleen de optie "**Scale**" te selecteren.

3.6 Opdracht

In Amerika is een archeologisch onderzoek uitgevoerd met als doel om de herkomst van pijlpunten te achterhalen. De pijlpunten zijn gemaakt van obsidiaan, een vulkanisch materiaal. Als de oorsprong van het materiaal bekend is, dan kan dit aanwijzingen geven over het migratiepatroon van indianenstammen. Daarom is obsidiaan van een aantal geologische vindplaatsen onderzocht en zijn hierin tien sporenelementen gemeten. De archeologen wilden weten of het mogelijk is om op basis van die tien gemeten sporenelementen onderscheid te maken tussen de verschillende geologische vindplaatsen.

In de dataset staan de gegevens voor twee geologische vindplaatsen. Deze dataset is digitaal beschikbaar (archeologische dataset groep 1 en 2.sav).

1. Bereken het variabele gemiddelde, de range en standaarddeviatie voor de 10 sporenelementen. Liggen deze op hetzelfde niveau? Moet er een voorbewerking op de dataset plaatsvinden en welke dan?

2. Voer een U-transformatie uit op de dataset en een Z-transformatie.

3. Stel de correlatiematrix op van de originele variabelen en geef commentaar. Is er een correlatie tussen de sporenelementen en welke zijn dat?

De gegevens:

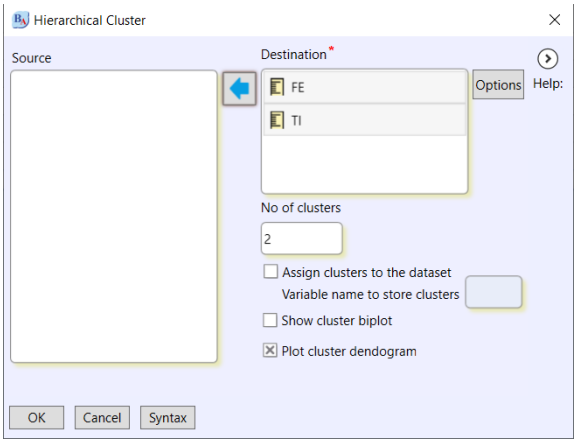
X	Fe	Ti	Ba	Ca	K	Mn	Rb	Sr	Y	Zr
1	1100	390	55	920	460	45	120	57	58	142
1	1173	417	54	961	441	47	135	55	60	145
1	1164	404	56	916	446	42	120	58	45	148
1	1030	373	59	920	487	38	128	53	58	138
1	1077	373	55	888	455	38	97	51	54	145
1	1080	403	53	919	442	41	133	60	45	155
1	1020	360	59	883	473	43	119	40	50	134
1	1050	396	56	924	482	48	140	74	71	157
1	1100	373	53	910	477	51	137	61	58	152
1	1069	375	51	958	429	42	100	51	47	128
2	863	183	8	626	452	34	121	15	58	70
2	1108	289	7	783	426	41	109	15	57	67
2	1210	276	10	966	430	44	117	20	44	73
2	1205	291	10	975	420	43	115	25	58	73
2	1100	267	10	910	500	40	145	25	65	95
2	1100	280	10	872	515	49	145	38	60	65
2	689	114	9	534	404	26	110	25	50	55
2	1186	257	10	940	431	40	121	20	53	73
2	860	182	7	722	418	33	115	20	55	53

4.4 Hoe doe je clusteranalyse

- Start BlueSky statistics op en open de dataset **FeTi dataset**.

- Klik in de menubalk op "**Analysis**", selecteer "**Cluster Analysis**" en kies de optie "**Hierarchical Cluster**".

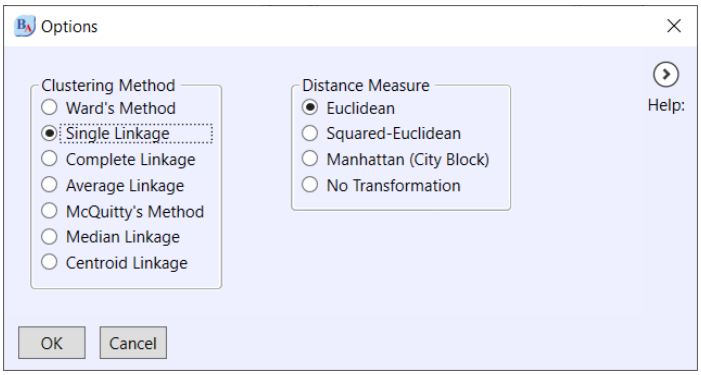
- Geef vervolgens aan met welke variabelen je de clusteranalyse wilt uitvoeren. Dit doe je door deze te selecteren in het linkerscherm en daarna op de [**pijl**] te drukken voor het invoerveld "**Destination**". Op dezelfde manier kun je variabelen ook de-selecteren (weer naar links verplaatsen).



- Geef aan hoeveel clusters er maximaal gevonden kunnen worden.

- Vink de optie "**Plot cluster dendrogram**" aan.

- Klik op [**Options**] en selecteer bij de optie "**Clustering Method**" de clustermethode in die je wilt toepassen.



- Bij de optie "**Distance Measure**" selecteer je vervolgens welke afstandsmaat je wilt toepassen.

- Klik twee keer op [**OK**] en de clusteranalyse wordt uitgevoerd. In de output wordt het dendrogram weergegeven.

4.5 Opdrachten

Archeologische dataset
Open de dataset van het archeologische onderzoek, **Archeologische dataset.sav**. Selecteer in eerste instantie alleen objecten die bij de 4 geologische bronnen horen (gecodeerd 1 t/m 4). Dit doe je door daarvan een subset te maken. De vraagstelling bij dit archeologisch onderzoek is of de vier geologische bronnen op basis van de gemeten sporenelementen van elkaar te onderscheiden zijn.

1. Bekijk de dataset. Is er een voorbewerking nodig? Waarom wel of niet?

2. Voer een clusteranalyse uit op de objecten met als afstandsmaat de euclidische afstand en als methode single linkage. Bekijk het dendrogram. Zijn de objecten van de vier bronnen in vier clusters te onderscheiden?

3. Voer een clusteranalyse uit met als afstandsmaat de euclidische afstand, maar neem nu verschillende methoden (complete linkage, average linkage between groups, ward). Vergelijk de dendrogrammen. Zijn de objecten van de vier bronnen in vier clusters te onderscheiden?

4. Trek een conclusie. Zijn de vier geologische bronnen van elkaar te onderscheiden en zo ja, met welke afstandsmaat en clustermethode(n).

5. Herhaal de clusteranalyse met de meest geschikte afstandsmaat en clustermethode voor de gehele dataset, dus ook met de objecten van de twee archeologische vindplaatsen 6 en 7. Zijn de objecten van deze twee archeologische bronnen terug te voeren op een van de vier geologische bronnen? Kun je een uitspraak doen over het migratiepatroon van de indianenstammen?
   Eventueel kun je de clusteranalyse nog een keer uitvoeren met de andere afstandsmaten en clustermethoden om te kijken of dit nog van invloed is op de resultaten.

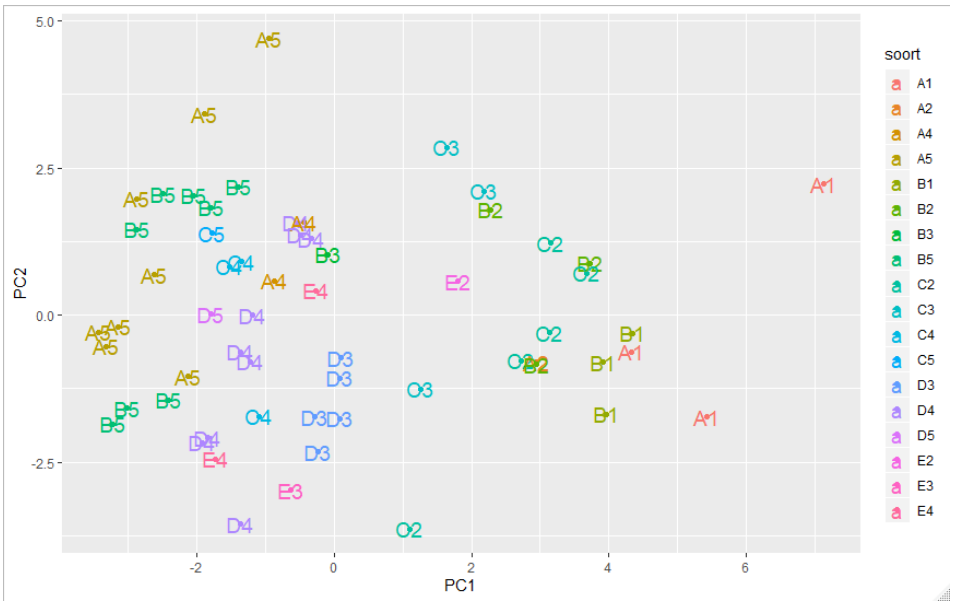
Opmerking:
Als de geologische bronnen niet van elkaar te onderscheiden zijn op basis van de gemeten sporenelementen dan moet er verder onderzoek plaatsvinden.

5.6 Scoreplot en loadingplot

Principale componentenanalyse wordt vaak gebruikt om relaties tussen objecten te vinden en om te onderzoeken of variabelen sterk gecorreleerd zijn (dezelfde informatie leveren). In dit laatste geval is het mogelijk om bij vervolgonderzoek minder variabelen te meten.

Informatie over relaties tussen objecten zijn zichtbaar in het zogenaamde scoreplot. Hierin zijn de scores (coördinaten) van de objecten voor twee of drie principale componenten in een grafiek uitgezet. Informatie over de variabelen is te vinden in het loadingplot.

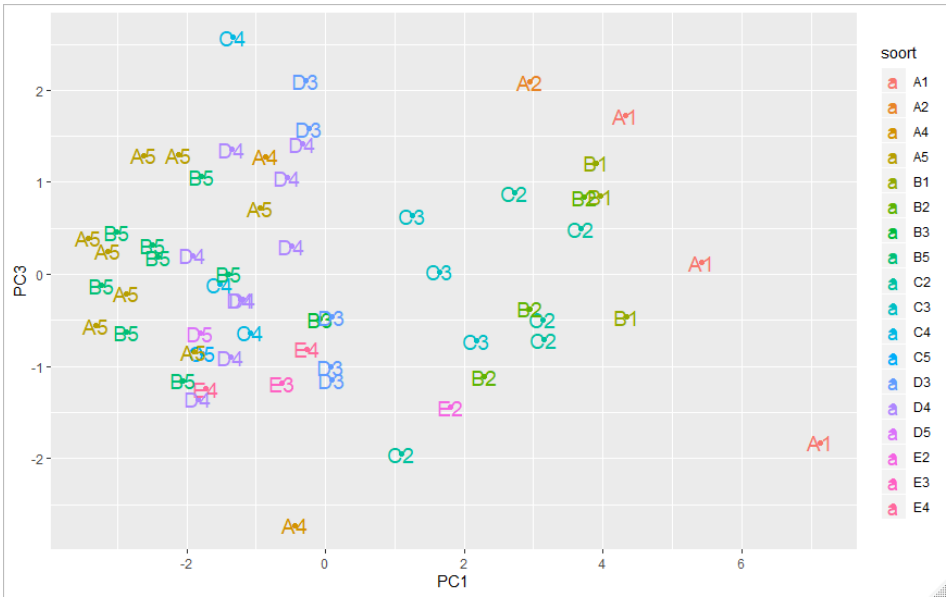
Om de interpretatie van de informatie in scoreplot en loadingplot te verduidelijken wordt het volgende voorbeeld gebruikt. Er is een onderzoek geweest naar de kwaliteit van een aantal erwtensoorten en oogsttijden. De erwtensoorten zijn gecodeerd A, B, C, D en E; de oogsttijden zijn gecodeerd met 1, 2, 3, 4 en 5. Van de erwten zijn een aantal eigenschappen bepaald, zoals kleur, geur, malsheid, etc. Het scoreplot van PC1 en PC2 voor het voorbeeld ziet er als volgt uit.



Vraag:
Wat zie je in de grafiek? Zijn er duidelijk groepen objecten waar te nemen? Is er onderscheid tussen de verschillende erwtensoorten A t/m E? Is er onderscheid tussen de verschillende oogsttijden 1 t/m 5? Heb je een idee wat PC1 zou kunnen voorstellen?

Het scoreplot van PC1 tegen PC2 levert de meeste informatie op, maar ook andere scoreplots kunnen relevante informatie bevatten. Met de informatie van een scoreplot is het mogelijk om uitbijters op te sporen. Deze uitbijters moeten uit de dataset worden verwijderd waarna de principale componentenanalyse opnieuw wordt uitgevoerd.

Hieronder de scoreplot van PC1 en PC3.



De loading is de cosinus van de hoek tussen een principale componenten-as en een variabele-as. Als een principale component sterk overeenkomt met een variabele, dan is de hoek tussen de PC-as en de variabele-as klein en de cosinus groot (maximaal 1). Een variabele kan een grote, een gemiddelde of een kleine invloed op een principale component hebben. De grenzen die hiervoor gehanteerd worden zijn:

grote invloed	loading tussen 0,7 – 1,0
gemiddelde invloed	loading tussen 0,5 – 0,7
kleine invloed	loading tussen 0 – 0,5

Deze invloed kan zowel positief als negatief zijn. Dat wil zeggen dat een variabele een directe relatie of een inverse relatie heeft met de principale component. Uit het loadingplot is die relatie af te lezen. Soms hebben variabelen een gemiddelde loading op twee principale componenten. Omdat twee principale componenten orthogonaal zijn, betekent dit dat deze variabelen van invloed zijn op twee verschillende factoren die de variantie veroorzaken.

Als er slechts één variabele een hoge loading heeft op een principale component, dan betekent dit dat deze variabele niet is gecorreleerd met de overige variabelen. De variabele levert dus unieke informatie, maar dit kan ook ruis zijn! De principale component kan dus ook een afwijking in een meting of een object beschrijven en zo kunnen afwijkingen opgespoord worden.

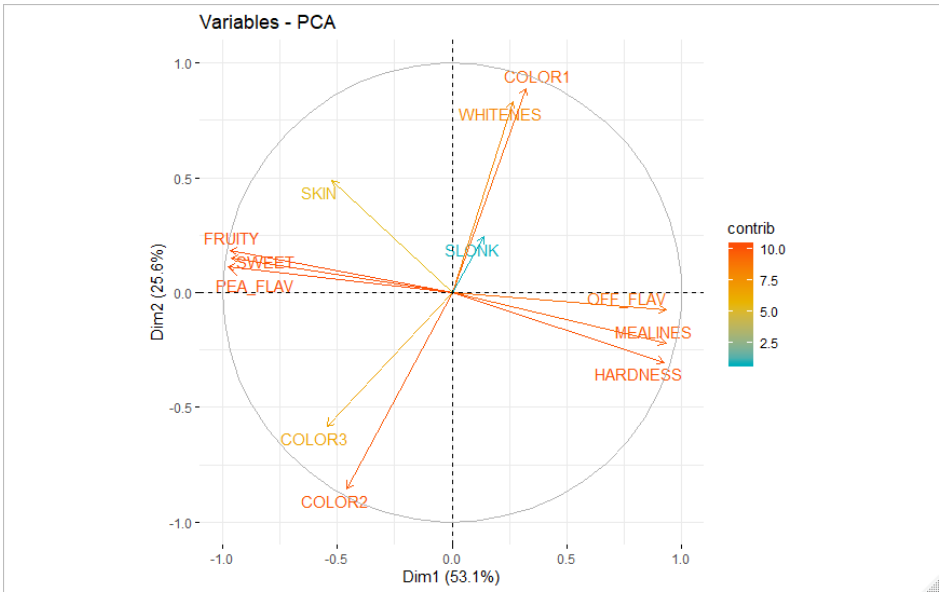
De loadings van de Peas onderzoek dataset voor de eerste drie principale componenten zijn in de tabel hieronder te vinden.

	Dim.1	Dim.2	Dim.3
PEA_FLAV	-0.9751204	0.11463886	-0.0172817008
SWEET	-0.9614643	0.15292403	0.0847441964
FRUITY	-0.9647551	0.18563298	-0.0001528903
OFF_FLAV	0.9321122	-0.07449691	0.0553715821
MEALINES	0.9313957	-0.21905374	-0.0072343216
HARDNESS	0.9255084	-0.30343365	-0.0313663589
WHITENES	0.2632781	0.83277489	-0.3339093037
COLOR1	0.3219877	0.89078732	-0.1941174966
COLOR2	-0.4579415	-0.85302193	0.1366988102
COLOR3	-0.5446026	-0.58144030	-0.4204424875
SKIN	-0.5220562	0.48850610	0.1778006012
SLONK	0.1395219	0.24275152	0.8994782023

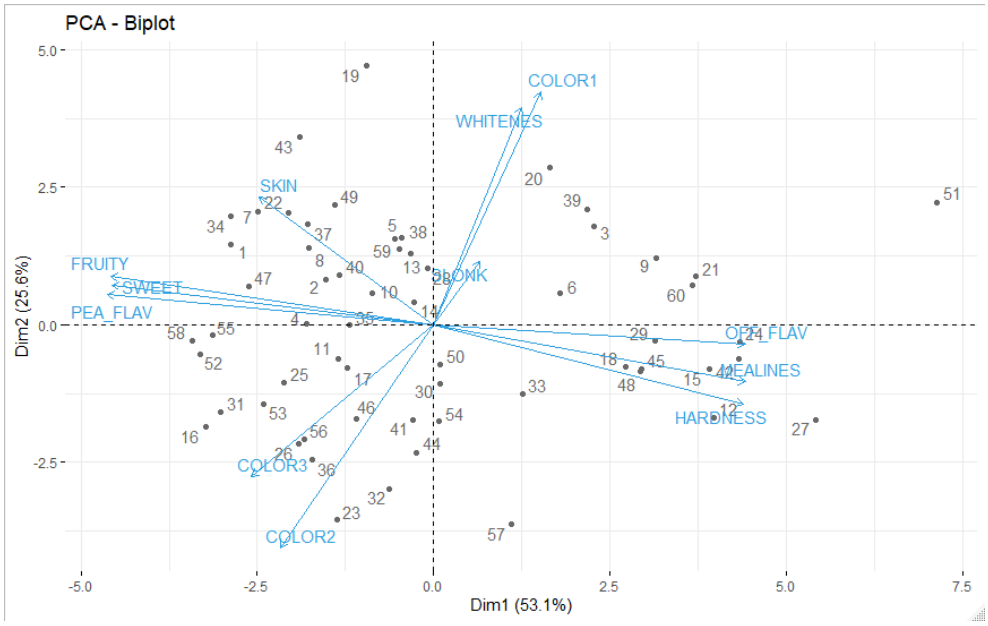
Vraag:
Welke variabelen hebben een grote invloed op PC1? En welke op PC2? En welke op PC3?
Welke variabelen hebben een gemiddelde invloed op PC1? En welke op PC2? En welke op PC3?
Zijn er variabelen die een kleine invloed hebben op alle drie principale componenten?

In het loadingplot is goed te zien welke variabelen sterk gecorreleerd zijn, want deze liggen dan dicht bij elkaar. Als variabelen sterk gecorreleerd zijn, dan kan vaak variabelenreductie worden toegepast. Van de gecorreleerde variabelen worden dan een aantal variabelen weggelaten bij vervolgonderzoek, meestal de variabelen die tijdrovend, arbeidsintensief, etc. zijn. Natuurlijk moet wel gecontroleerd worden of er bij de principale componentenanalyse met minder variabelen niet te veel informatie verloren gaat.

De loadingplot voor de Peas onderzoek dataset ziet er als volgt uit:



Door het scoreplot te combineren met het loadingplot kan zichtbaar gemaakt worden welke objecten vergelijkbare waarden hebben voor bepaalde variabelen. Soms kun je dan ook informatie halen over wat een principale component eigenlijk beschrijft (voorstelt).

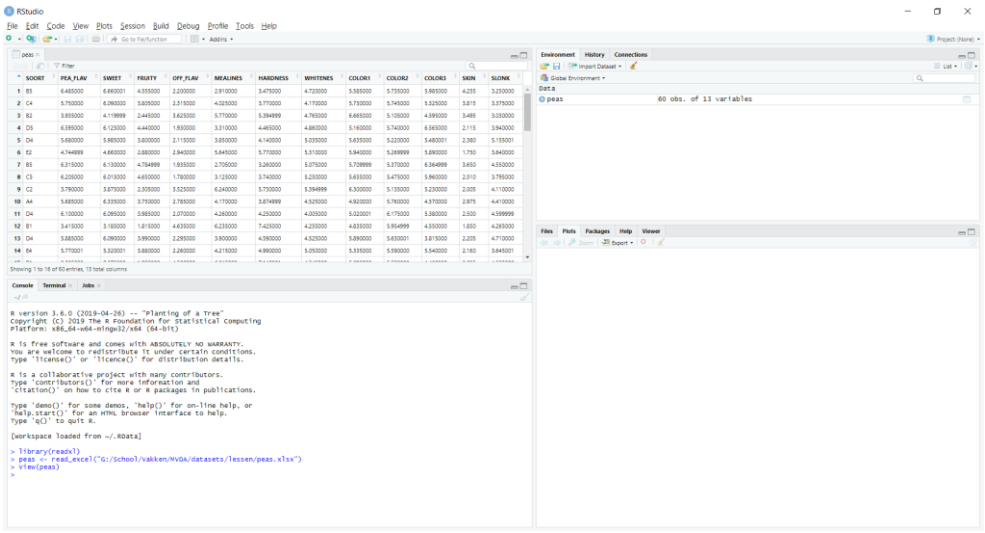


Vraag:
Wat kan je zeggen van de objecten links in het scoreplot? En van de objecten rechts in het scoreplot? Heb je nu een idee wat PC1 voorstelt?

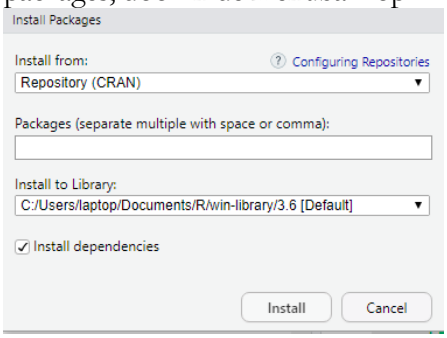
5.7 Hoe doe je principale componentenanalyse

- Start RStudio op. Als je met RStudio aan de slag gaat, dan krijg je een scherm te zien dat in vier gebieden onderverdeeld is (de allereerste keer is het gebied linksboven verborgen):

  - Linksboven de dataset die geopend is of de actuele data die je wilt bekijken.
  - Linksonder de console, waar de commando's ingevoerd worden en waar ook gegevens weergegeven worden.
  - Rechtsboven komen alle gegevens etc te staan, je kan deze aanklikken om te bekijken.
  - Rechtsonder kan je packages installeren en komen ook de grafieken te staan.



- Eenmalig:
  Kijk welke packages standaard mee geïnstalleerd zijn en installeer de benodigde packages, door in de menubalk op "Tools" en "Install packages" te klikken.



  Geef de namen van de packages op en vergeet niet om "Install dependencies" aan te vinken, dan worden alle andere packages die nodig zijn mee geïnstalleerd. Packages die je nodig hebt zijn in ieder geval:
  - factoextra
  - ggplot2
  - FrF2 (voor DoE)
  - PID (voor DoE)
  - RSM (voor DoE)

- Zorg dat je in ieder geval de volgende packages aangevinkt hebt:
  - factoextra
  - ggplot2
  - ggrepel
  - readxl
  - tools
  - utils

- Importeer het bestand **Peas.xlsx** (importeer vanuit excel).

- Als eerste moet je aangeven waaruit de dataset bestaat, dus in dit geval label (hier sample genoemd) en variabelen (hier chemical genoemd) met de betreffende kolommen uit de dataset. In dit voorbeeld zit het label in kolom 1 en de variabelen in kolommen 2 t/m 13.
  peas1 <- data.frame(sample = as.matrix(peas[,1]), chemical = I(as.matrix(peas[,2:13])))

  Met het commando [x,y] geef je met x aan om welke rij(en) het gaat (als je een aantal objecten zou willen selecteren) en met y aan om welke kolom(men) het gaat. Als je meerdere kolommen of rijen wilt selecteren dan wordt dat weergegeven met begin:eind (dus een dubbele punt ertussen).

- Om de PCA uit te voeren op basis van de Z-transformatie voer je in:
  peas.pca <- prcomp(peas1$chemical, center = TRUE, scale = TRUE)
  Als je een PCA op basis van U-transformatie wilt doen, dan wordt het scale = FALSE

- Het screeplot krijg je met: fviz_eig(peas.pca)

- Als je een scoreplot wilt met alleen de nummers van de objecten, dan kan je doen: fviz_pca_ind(peas.pca)
  Als je een scoreplot wilt met groepen/categorieën in verschillende kleuren en ook het label erbij dan doe je het volgende:
  df_out <- data.frame(peas.pca$x, soort=peas$SOORT)
  p<-ggplot(df_out,aes(x=PC1,y=PC2,color=soort, label=soort))
  p<-p+geom_point()+ geom_text(size=5)
  p
  Als je de labels groter of kleiner wilt hebben, dan pas je de size=5 aan.
  Heb je per categorie meerdere monsters, dan kan je bij color de categorie als variabele nemen en bij label de variabele met de monsternaam. Bij label vul je dan de naam van de variabele in de dataset in (bijvoorbeeld label = pijl$Bron bij de archeologische datatset).

- Het loadingplot krijg je met
  fviz_pca_var(peas.pca, col.var = "contrib", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)
  Door repel op TRUE te zetten, krijg je geen overlap in de tekst. De variabelen worden op basis van belangrijkheid meer verschillende kleuren weergegeven.

- Om scoreplot en loadingplot in een grafiek te laten weergeven doe je:
  fviz_pca_biplot(peas.pca, repel = TRUE, col.var = "#2E9FDF", col.ind =

"#696969")
Variabelen en objecten hebben een aparte kleur gekregen zodat ze duidelijker te herkennen zijn.

- De eigenwaarden zijn te zien met:
  eig.val <- get_eigenvalue(peas.pca)
  eig.val

- Verder kan je over de variabelen nog meer informatie krijgen met:
  res.var <- get_pca_var(peas.pca)
  res.var$coord[,1:3]
  Met coord krijg je de loadings van de variabelen voor de eerste drie principale componenten in een tabel te zien, en kan je dus zien hoe belangrijk een variabele voor een bepaalde principale component is.

- Vergelijkbare informatie (nu met de scores) krijg je voor de objecten met:
  res.ind <- get_pca_ind(peas.pca)
  res.ind$coord[,1:3]

- De commando's zijn beschikbaar in een Notepad (tekst) document, zodat je ze kunt knippen en plakken in RStudio. Je kan de commando's ook aanpassen aan nieuwe opdrachten en de data van het project chemical fingerprinting.

5.8 Opdrachten

Archeologische dataset
Importeer de dataset van het archeologische onderzoek, **Archeologische dataset**. De vraagstelling bij dit archeologisch onderzoek is of de vier geologische bronnen op basis van de gemeten sporenelementen van elkaar te onderscheiden zijn en of de twee archeologische vindplaatsen overeenkomsten met een van de geologische bronnen hebben.

1. Bekijk de dataset. Is er een voorbewerking nodig? Zijn de data van de variabelen normaal verdeeld? Waarom wel of niet?

2. Voer een principale componentenanalyse (PCA) uit op basis van de correlatiematrix en neem in eerste instantie 10 principale componenten mee.

3. Bekijk het screeplot. Hoeveel principale componenten zijn er nodig om 80 – 90 % van de informatie weer te geven?

4. Maak de scoreplots. Zijn de objecten van de vier bronnen in vier clusters te onderscheiden? Hoeveel principale componenten heb je nodig voor een goede interpretatie van de scoreplots?

5. Bekijk de loadingplots. Welke variabelen hebben een grote of gemiddelde invloed op de principale componenten die je nodig hebt? Zijn er variabelen die je eventueel weg kunt laten, omdat ze alleen informatie leveren aan een PC die je niet nodig hebt?

6. Zijn er variabelen die overeenkomsten hebben? Zo ja, welke? Komt dit overeen met de correlatiematrix? Welke variabelen kun je bij vervolgonderzoek weglaten?

7. Voer de principale componentenanalyse nogmaals uit zonder deze variabelen. Vergelijk de scoreplots met die van de eerste PCA. Is de informatie in de scoreplots duidelijker of minder duidelijk geworden.

8. Trek een conclusie. Zijn de vier geologische bronnen van elkaar te onderscheiden? Van welke geologische bron zijn de objecten van de twee archeologische vindplaatsen afkomstig? En hoeveel PCs zijn hiervoor nodig? En welke variabelen?

Opmerking:
Als de geologische bronnen niet van elkaar te onderscheiden zijn op basis van de gemeten sporenelementen dan moet er verder onderzoek plaatsvinden.

Levensmiddelen dataset
Impoteer de dataset van het levensmiddelenonderzoek, **Levensmiddelenpatroon**. De verwachting is dat mensen uit landen in een bepaald gebied van Europa eenzelfde levensmiddelenpatroon hebben.

1. Bekijk de dataset. Is er een voorbewerking nodig? Zijn de data van de variabelen normaal verdeeld? Waarom wel of niet?

2. Voer een principale componentenanalyse (PCA) uit op basis van de correlatiematrix en neem in eerste instantie 10 principale componenten mee.

3. Bekijk het screeplot. Hoeveel principale componenten zijn er nodig om 80 – 90 % van de informatie weer te geven?

4. Maak de scoreplots. Zijn de landen in verschillende clusters te onderscheiden? En zijn dit logische groepen? Hoeveel principale componenten heb je nodig voor een goede interpretatie van de scoreplots?

5. Bekijk de loadingplots. Welke variabelen hebben een grote of gemiddelde invloed op de principale componenten die je nodig hebt? Zijn er variabelen die je al weg kunt laten, omdat ze alleen informatie leveren aan een PC die je niet nodig hebt?

6. Zijn er variabelen (voedingsmiddelen) die overeenkomsten hebben? Zo ja, welke? Komt dit overeen met de correlatiematrix? Welke variabelen kun je bij vervolgonderzoek weglaten?

7. Voer de principale componentenanalyse nogmaals uit zonder deze variabelen. Vergelijk de scoreplots met die van de eerste PCA. Is de informatie in de scoreplots duidelijker of minder duidelijk geworden.

8. Trek een conclusie. Welke clusters van landen kan je onderscheiden? En hoeveel PCs zijn hiervoor nodig? En welke variabelen?