Chemometrie draagt in laboratoria bij aan kwaliteitszorg

en aan de beheersing van Big Data (Deel 2)

Dr. dr. J.P.M. Andries

Deel 1 is beschreven dat in de laboratoria door de moderne analyse-instrumenten en door geautomatiseerde analysesystemen grote hoeveelheden data worden gegenereerd. Bovendien komen er steeds meer grote hoeveelheden data beschikbaar voor onderzoek in online databases. Daardoor is in de laboratoria een *'tsunami'* aan data veroorzaakt die soms ook wordt aangeduid als een *'Big Data'* probleem. Daarnaast moeten de onderzoeksresultaten voldoen aan strenge kwaliteitseisen. De laboratoria moeten daarom beschikken over een goed kwaliteitszorgsysteem. *Chemometrie* draagt in de laboratoria bij aan goede kwaliteitszorg en aan de beheersing van 'Big Data' op tal van onderzoekgebieden zoals chemisch en biomedisch onderzoek, milieuonderzoek, levensmiddelenonderzoek, biotechnologie, bio-informatica en metabolomics. Daardoor is in het onderwijs en in het werkveld behoefte aan meer kennis ontstaan op het gebied van chemometrie. Recent is een geheel herziene vierde druk van het boek *'Chemometrie'* bij Syntax Media verschenen waarin de basisprincipes worden beschreven van veelgebruikte chemometrische technieken die zich bewezen hebben, zie Ref. (1). In Deel 1 is een overzicht gegeven van bijdragen die de chemometrie levert aan de kwaliteitszorg van laboratoriumonderzoek en de beheersing van Big Data, toegelicht met voorbeelden. Dit overzicht wordt hierna vervolgd. Ook wordt informatie gegeven over het boek en hoe dit gebruikt kan worden.

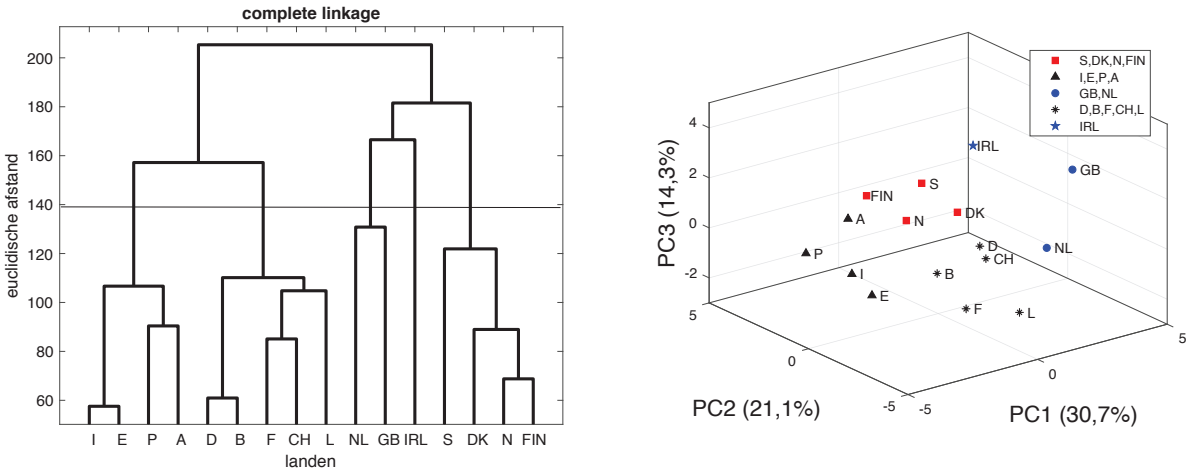
Multivariate data-analyse en Big Data

Analyses worden vaak uitgevoerd in mengsels met een complexe samenstelling zoals bijvoorbeeld monsters uit reactievaten in de industrie, uit het milieu, van natuurproducten of van lichaamsvloeistoffen. Het isoleren van zuivere componenten en de identificatie en concentratiebepaling ervan is arbeidsintensief, tijdrovend en daardoor kostbaar. Er zijn daarom chemometrische technieken ontwikkeld waarbij de isolatie van componenten uit complexe mengsels niet nodig is.

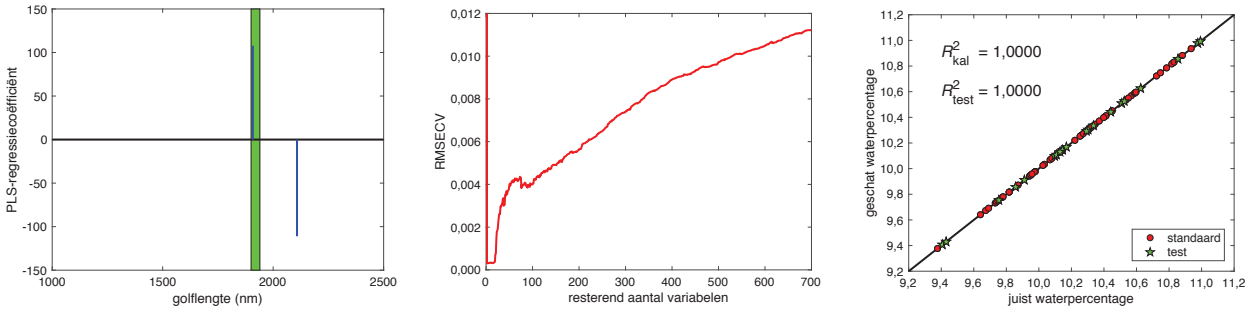
De gegenereerde signalen zijn vaak multivariaat van aard. Dat wil zeggen dat ze bestaan uit meetsignalen die afkomstig zijn van meerdere variabelen, zoals signalen bij verschillende golflengten in spectra en bij verschillende retentietijden in chromatogrammen. De multivariate data moeten worden omgezet in informatie over de aard en hoeveelheid of concentraties van componenten in complexe monsters. Daarvoor zijn in de chemometrie *data-gedreven* multivariate data-analysemethoden ontwikkeld. Daarbij wordt niet gewerkt met vooraf opgestelde theoretische modellen maar worden de modellen ontwikkeld op basis van de data zelf. Multivariate data-analysemethoden spelen een belangrijke rol bij het oplossen van het Big Data probleem. Ze hebben twee belangrijke toepassingsgebieden, patroonherkenning en multivariate modelleringstechnieken. Bij patroonherkenning wordt gezocht naar verborgen structuren in datasets door het aantal dimensies te verminderen. Mogelijk aanwezige structuren in een dataset worden bij minder dimensies duidelijker zichtbaar, zoals in een dendrogram of in een scoreplot van principale componenten, zie hierna. Bij multivariate modelleringstechnieken worden modellen ontwikkeld die daarna voor predictie of voorspellingen kunnen worden gebruikt voor de bepaling van componenten in complexe monsters.

Er bestaan verschillende methoden voor multivariate data-analyse. De keuze voor de beste methode is afhankelijk van het doel van de analyse en van de structuur van de data. In de chemometrie zijn de meest toegepaste multivariate technieken clusteranalyse, Meervoudige Lineaire Regressie (MLR), Principale Componenten Analyse (PCA), Principale Componenten Regressie (PCR) en Partial Least Squares (PLS). Als er geen afhankelijke variabele is, kan patroonherkenning worden uitgevoerd door middel van clusteranalyse of Principale Componenten Analyse. Als er wel één of meer afhankelijke variabelen zijn, kan MLR, PCR of PLS worden toegepast voor het opstellen van kalibratiemodellen die kunnen worden gebruikt voor de predictie of voorspelling van de samenstelling van nieuwe monsters.

Figuur 1. Onderzoek naar de overeenkomst in het gebruik van 20 verschillende soorten levensmiddelen in 16 Europese landen; (A) Dendrogram; (B) 3D-scoreplot na Principale Componentenanalyse.



Figuur 2. FCAM-variabelenselectie; (A) Voorspelfout uitgezet tegen het resterend aantal variabelen; (B) NIR waterband en twee geselecteerde golflengten bij 1908 en 2108 nm met bijbehorende PLS-regressiecoëfficiënten; (C) Voorspellend vermogen van het PLS-model voor de bepaling van het waterpercentage in mais op basis van de twee geselecteerde golflengten.



Bij multivariate datasets is het aantal variabelen *p* meestal veel groter dan het aantal objecten *n*. Dat maakt het noodzakelijk om het aantal variabelen te reduceren. Bij MLR gebeurt dat door een selectie te maken uit de set van oorspronkelijk gemeten variabelen. Bij PCA en PLS wordt de dataset eerst getransformeerd naar een ander stelsel met verborgen of latente variabelen, de principale componenten of PLS-factoren. Hierna wordt het aantal dimensies gereduceerd door de belangrijkste latente variabelen te selecteren. PCA en PLS hebben meestal de voorkeur boven MLR omdat daarbij alle informatie van de gemeten variabelen kan worden gebruikt.

Bij PCA wordt een dataset ***X*** door een lineaire combinatie van de oorspronkelijke variabelen getransformeerd naar een set met minder variabelen, de principale componenten, die de meeste relevante informatie bevatten. De principale componenten worden berekend in volgorde van verklaarde variantie. Ze staan loodrecht op elkaar en zijn niet gecorreleerd. De eerste principale component loopt in de richting van de grootste spreiding in de dataset en verklaart de meeste variantie. De tweede principale component staat loodrecht op de eerste principale component, loopt in de richting van de op één na grootste spreiding en verklaart de op één na grootste hoeveelheid variantie. Dit proces wordt herhaald totdat uit alle *p* oorspronkelijke variabelen ook *p* principale componenten zijn berekend.

Oorspronkelijke variabelen die onderling sterk zijn gecorreleerd, combineren meestal in één principale component. De meeste informatie die in de oorspronkelijke variabelen zit, komt terecht in de eerste *a* (*a* < *p*) principale componenten. Deze zijn daarom het belangrijkste. Ze worden gebruikt voor patroonherkenning en voor multivariate modellering. Omdat *a* < *p* vindt bij toepassing van PCA dimensiereductie plaats. De kracht van PCA kan worden verklaard door het feit dat vrijwel alle informatie uit de oorspronkelijke dataset geconcentreerd wordt in een beperkt aantal niet-gecorreleerde principale componenten.
Bij Principale Componenten Regressie worden multivariate regressiemodellen ontwikkeld op basis van de belangrijkste principale componenten. Bij PCA en PCR worden de latente variabelen (de principale componenten) bepaald zonder gebruik te maken van de informatie in de afhankelijke variabelen. Bij Partial Least Squares wordt daar wel gebruik van gemaakt.

Een niet-chemisch voorbeeld van een toepassing van patroonherkenning is afgebeeld in Figuur 1. Hierin wordt een onderzoek gepresenteerd naar het gebruik van 20 verschillende soorten levensmiddelen in 16 Europese

landen. De landen zijn hierin aangegeven met afkortingen die worden gebruikt op kentekenplaten van auto's. Met behulp van clusteranalyse kan een dendrogram of boomdiagram worden gemaakt, zie Figuur 1A, waarin de overeenkomst in het levensmiddelenpatroon tussen groepen van landen kan worden afgelezen. In het dendrogram kan op de verticale as de afstand tussen clusters worden afgelezen. Hoe kleiner de afstand tussen clusters is, des te groter is de overeenkomst. Een cluster wordt in het dendrogram gevormd als twee verticale lijnen met elkaar worden verbonden door een horizontale lijn. Landen die sterk overeenkomen in het gebruik van levensmiddelen liggen dicht bij elkaar in de 20-dimensionale ruimte van de levensmiddelen zoals Italië (I) en Spanje (E), en Duitsland (D) en België (B). Zij vormen in het dendrogram al op geringe afstand (laag in het dendrogram) clusters.

In het dendrogram zijn vijf clusters te onderscheiden. Deze clusters worden gevonden door een horizontale lijn te trekken op een verticale afstand van ±140. Er worden dan vijf verticale lijnen doorgesneden. Alle landen die onder een doorgesneden lijn of onder een van de vertakkingen ervan liggen, vormen een cluster. De volgende combinaties van landen vormen clusters: (*i*) Italië, Spanje, Portugal en Oostenrijk (Zuid-Europese landen), (*ii*) Duitsland, België, Frankrijk, Zwitserland en Luxemburg (Midden-Europese landen), (*iii*) Nederland en Groot-Brittannië, (*iv*) Ierland, en (*v*) Zweden, Denemarken, Noorwegen en Finland (Scandinavische landen). De landen in de clusters komen onderling overeen in het levensmiddelengebruik. Deze overeenkomst was niet te zien in de oorspronkelijke tabel met gegevens van 20 levensmiddelen voor de 16 landen. Het betrof dus verborgen informatie in de dataset. De gevonden clustering is goed te begrijpen vanwege de geografische ligging van de landen.

In dit voorbeeld zijn gegevens voor 20 variabelen (de levensmiddelen), vanuit de 20-dimensionale (20D)-ruimte, door clustering teruggebracht naar de 2D-ruimte van het dendrogram waarin duidelijk patronen zijn te herkennen. Dit is dus een mooi voorbeeld van dimensie-reductie. Voor dezelfde dataset kan ook een principale componentenanalyse worden uitgevoerd. De eerste drie principale componenten verklaren samen 66,1% van de totale variantie in de dataset. De coördinaten van de landen op de principale componentenassen kunnen worden weergegeven in een 3D-scoreplot van de eerste drie principale componenten, zie Figuur 1B. Hierin zijn ook de clusters van de landen die gevonden zijn met het dendrogram weergegeven. In deze scoreplot heeft een dimensie-reductie van 20D naar 3D plaatsgevonden.

Een voorbeeld van een toepassing van multivariate modellering en voorspelling is afgebeeld in Figuur 2. Dit betreft de bepaling van het waterpercentage in mais met behulp van nabij-infra-rood (NIR) spectroscopie. Daarbij wordt de tijdrovende klassieke methode voor vochtbepaling, met wegen, drogen en opnieuw wegen, vervangen door een snelle methode met een gemiddelde analysetijd van 1 minuut per monster. De dataset bestaat uit een kalibratieset met 60 maismonsters en een test set met 20 maismonsters, waarvoor NIR spectra zijn opgenomen bij 700 golflengten. De kalibratieset wordt gebruikt voor de ontwikkeling van een PLS-model. De testset wordt gebruikt voor het testen van het ontwikkelde model. Het is bekend dat het voorspellend vermogen van het PLS-model kan worden verbeterd door eerst de niet-informatieve variabelen (meetgolflengten) te verwijderen. Dat is uitgevoerd met de Final Complexity Adapted Models (FCAM)-methode. De verbetering van het voorspellend vermogen van het PLS-model is te zien in Figuur 2A. Deze grafiek geeft de voorspelfout van het PLS-model (RMSECV) als functie van het resterend aantal variabelen bij variabelenselectie. De RMSECV-waarde neemt duidelijk af naarmate het aantal resterende variabelen kleiner wordt.

Op basis van deze RMSECV-curve is een set met slechts twee variabelen geselecteerd bij 1908 en 2108 nm, zie Figuur 2B. De golflengte van 1908 nm ligt binnen de NIR waterband. Met beide geselecteerde golflengten is vervolgens een PLS model ontwikkeld met een perfect beschrijvend vermogen voor de kalibratieset en een perfect voorspellende vermogen voor de test set omdat beide een bepalingscoëfficiënt hebben van R^2 = 1,0000, zie Figuur 2C. In dit voorbeeld heeft een sterke dimensiereductie plaatsgevonden van 700 naar 2 variabelen.

Boek Chemometrie

In het boek Chemometrie worden de volgende onderwerpen behandeld: beschrijvende statistiek, signaalbewerking (digitale filters, Savitzky-Golay filters, numerieke integratie), lineaire regressie in combinatie met univariate kalibratie, bijzondere regressietechnieken zoals robuuste regressie, gewogen lineaire regressie en regressie met fouten in *x* en *y*, interne methodevalidatie, experimentele optimalisering met de meest toegepaste experimentele designs, sequentiële en simultane optimaliseringstechnieken, en de ontwikkeling van multivariate modellen voor patroonherkenning en schatting van afhankelijke variabelen met Multivariate Lineaire Regressie, Principale Componenten Analyse, Principale Componenten Regressie en Partial Least Squares.

Het boek is geschreven op basis van ervaringen van de auteur met de verzorging van chemometrieonderwijs in het hoger laboratoriumonderwijs en de uitvoering van wetenschappelijk onderzoek op verschillende toepassingsgebieden van chemometrie. Het boek is onafhankelijk van specifieke software geschreven. In de eerste plaats omdat de ervaring heeft geleerd dat daar in de loop van de tijd veranderingen in kunnen optreden. In de tweede plaats om de gebruiker vrij te laten in de keuze daarvan. Een uitzondering wordt gemaakt voor toepassingen met Excel omdat de veranderingen daarin relatief gering zijn en het beschikbaar is voor alle gebruikers.

Gebruikers vinden de wiskunde en statistiek bij data-analyse soms lastig. In het boek is daar rekening mee gehouden door bij verschillende technieken de wiskundige en statistische basis te concentreren in aparte para-

grafen. De gebruiker kan er voor kiezen om deze onderdelen al dan niet te bestuderen. Moderne software maakt het mogelijk om chemometrie functioneel te gebruiken. De gebruiker moet daarbij wel kennis hebben van de principes en toepassingsmogelijkheden van de chemometrische technieken die in het boek beschreven worden maar hoeft niet alle details te kennen van de wiskundige en statistische achtergronden van de technieken. Het rekenwerk wordt daarbij met behulp van software uitgevoerd en de gebruiker hoeft alleen te weten hoe de output geïnterpreteerd en gebruikt moet worden. Dit vereenvoudigt het toepassen van chemometrische technieken sterk.

De onderwerpen in het boek hebben hun nut in de praktijk bewezen en zijn toepasbaar op vele onderzoekgebieden. De methoden zijn beschreven met uitgewerkte voorbeelden en oefenopgaven. De afbeeldingen in dit artikel zijn afkomstig uit het boek.

Literatuur

1. *J.P.M. Andries, Chemometrie, 4e druk, Syntax Media, Utrecht, 2019, www.syntaxmedia.nl; ISBN 978 94 91764 332.*

Nieuws