

Alt det gode ved at bruge mange variable

Vi fortsætter med at se på fordelene ved multivariate metoder. Sidst handlede det om støjrreduktion, og nu tager vi håndtering af interferenser og muligheden for outlier kontrol [1].

Af Rasmus Bro, Søren Balling Engelsen, Institut for Fødevarevidenskab, Det Biovidenskabelige Fakultet, Københavns Universitet og Lars Nørgaard, FOSS

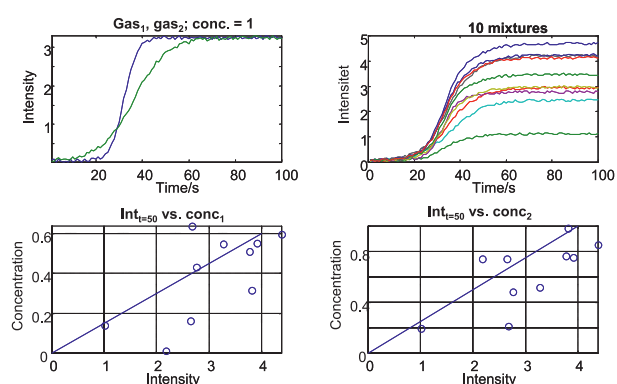
En vigtig fordel ved multivariate modeller er, at ikke-selektive signaler kan gøres selektive ad matematisk vej. Således kan interferenser håndteres, forudsat at den interfererende analyts signal ikke er helt identisk med signalet fra analytten. I figur 3 ses en situation, hvor signalet alle steder har bidrag fra både analyt og fra en (ukendt) interferens. Formen på det interfererende signal er endda meget lig analyttens (figur 3 øverst til venstre). Der kunne også være tale om mere end blot én interferens. Under alle omstændigheder er det ikke muligt at lave en traditionel univariat kalibreringsmodel, når signalet ikke er selektivt.

Det er interessant, at en multivariat kalibreringsmodel sagtens kan håndtere sådanne interferenser. Det betyder, at instrumental selektivitet ikke er nødvendig. Det er muligt at bygge kalibreringsmodeller, selv når kemisk eller fysisk selektivitet er umuligt at opnå. Det betyder også, at i forbindelse med udvikling og valg af sensorer er det ikke nødvendigt at fokusere primært på kemisk selektivitet. Andre aspekter, såsom signal-støj-forhold eller robusthed kan være lige så eller mere vigtigt, når kemisk selektivitet kan suppleres med matematisk selektivitet.

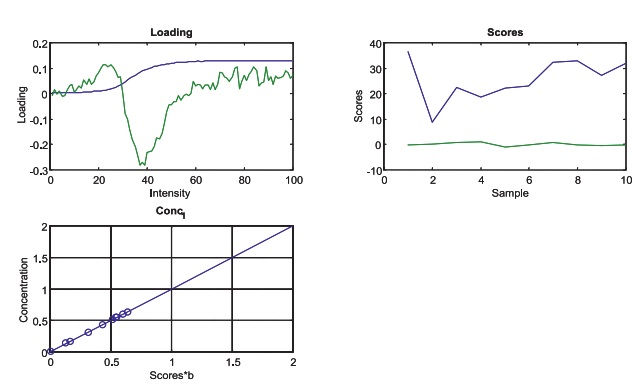
Antag, at signalerne fra de to stoffer er additive. Det betyder, at bidraget fra den ene ikke påvirker signalet fra den anden. Hvis dette holder omtrentligt, så kan man bygge en multivariat kalibreringsmodel og prædiktere indholdet af en ønsket analyt. For at opbygge kalibreringsmodellen må man have et sæt prøver - et kalibreringssæt, hvor analytkoncentrationen er kendt. Der er ikke brug for at vide noget om de andre stoffers (interferensers) koncentration. Typisk er 10-50 prøver nødvendige for at bygge en kalibreringsmodel.

Da der er to stoffer, der varierer i de målte profiler, så må man anvende to komponenter i en kalibreringsmodel for at beskrive al variationen. Det er ikke sådan, at den ene af disse komponenter giver analyt-koncentration og den anden interferensens koncentration. Af matematiske årsager er informationen blandet op i de to komponenter. Det ses også i figur 4 øverst til venstre, hvor to loadingvektorer er vist fra en PCA-model af kalibreringsdata. De ligner ikke de rene spektre af de to komponenter, men sammen rummer de den nødvendige information til at beskrive koncentrationen af analyt i en hvilken som helst tilsvarende prøve.

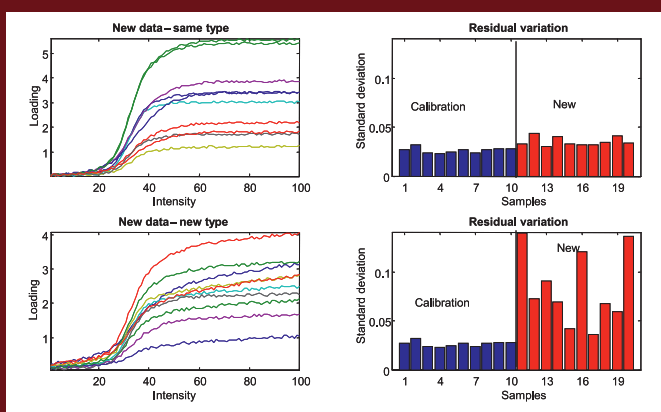
Den første komponent er givet ved den glatte kurve i figur 4 (blå kurve øverst til venstre). Denne komponent beskriver den generelle udvikling i data, og derfor vil scores for denne komponent afspejle det generelle niveau. Scoreværdierne er vist øverst til højre i



Figur 3. Illustration af interferensproblemet. Øverst til venstre vises signalet fra analyt (blå) og interferens (grøn). Øverst til højre vises profiler fra ti blandinger af disse to stoffer. Nederst til venstre vises intensitet ved tid 50 mod koncentration af analyt og til højre mod koncentration af interferens.



Figur 4. Multivariat kalibrering med interferenser. PCA på data fra figur 3 giver en to-komponent-model. Loadings vises øverst til venstre og scores til højre. Nederst til venstre er det vist, at for hver prøve kan man tage de to scoreværdier og gange med en regressionskoefficient og summere disse to tal. Denne prædiktion vil (efter at have tilføjet et offset) give et estimat af koncentrationen.



Figur 5. Outlier-kontrol. En kalibreringsmodel baseret på data i figur 3 med én analyt af og én interferens. Modellen bruges på data af samme type (øverst) og på data med en lidt anderledes ny interferens (nederst).

figur 4. Prøve to har f.eks. en lav scoreværdi på første komponent. Det betyder, at intensiteten af det målte signal må være lavt. Ganske rigtigt, så svarer denne prøve til den nederste profil i figur 3 (øverst til højre). Den anden loadingvektor er mere støjfyldt og mindre let at fortolke. Den kan tolkes som at beskrive afvigelsen fra den fælles profil afspejlet i komponent et. En kalibreringsmodel baseret på scoreværdierne kaldes en principal komponent regressionsmodel (PCR) [Dansk Kemi 8 2008].

Der findes en række teknikker til at opbygge multivariate kalibreringsmodeller. Af de mere almindelige er: partial least squares regression [Dansk Kemi 11 2008], principal komponent regression [Dansk Kemi 8 2008], feed-forward neurale netværk og ridge regression. Selvom der er forskelle mellem disse metoder, og selvom de ofte bliver præsenteret vha. meget forskellig terminologi, så er deres vigtigste fælles egenskaber, at de håndterer multivariate ikke-selektive målinger og muliggør udnyttelsen af alle målte oplysninger i stedet for at skulle ty til at udvælge få diskrete enkeltvariable.

Tredje multivariate fordel - finde outliers

Fejl opstår f.eks. pga. instrumentproblemer, stikprøvefejl, prøvehåndtering og mange, mange andre ting. Hvis disse fejl er tilstrækkelig store, enten i kvantitet eller kvalitet, kan de ødelægge enhver model. Det kan være svært at opdage outliers, når komplicerede multivariate data anvendes, men i virkeligheden er påvisning af outliers stærkt forbedret, når man benytter multivariate data frem for nogle få *a priori* udvalgte "vigtige" variable. Her følger et eksempel på dette:

Antag at en multivariat kalibreringsmodel er lavet som beskrevet. Når denne model skal bruges til at forudsige koncentrationen i nye prøver, er det nødvendigt at sikre, at disse prøver er af samme type, som de prøver kalibreringsmodellen blev bygget på. Dvs. at de nye profiler skal være linearkombinationer af de samme fænomener som de prøver, der anvendtes i kalibreringen. I figur 5 er vist et eksempel på et nyt sæt prøver, som er i overensstemmelse med kalibreringssættet (øverst til venstre) samt et datasæt "forurennet" med yderligere et stof i prøverne, som der ikke er kalibreret for (nederst til venstre).

Residualerne for en prøve er den del af det målte signal, der ikke kan beskrives af loadingvektorerne. Ideelt afspejler residualerne støj og er dermed af nogenlunde samme og forholdsvis lille størrelsesorden for alle prøver. Sådan er det også for prøver, der passer i modellen som vist i figur 5 (øverst til højre). Det ses, at residualerne for nye prøver har samme størrelse som prøverne i kalibreringssættet. Ser man i stedet på prøver, der indeholder en eller flere interferenser, så kan de gamle loadingvektorer fra kalibreringsmodellen ikke beskrive al infor-

mationen og residualerne bliver større ift. kalibreringsprøverne. I sådan en situation kan modellen ikke anvendes, fordi prøverne ikke er som kalibreringsprøverne. Heldigvis kan man let opdage dette ved at se på residualerne (figur 5 nederst til højre).

At anvende modellen, når der er nye interferenser i det målte signal, ville svare til at anvende en univariat kalibreringsmodel til prøver med interfererende signaler. Men bemærk, at kun i det multivariate tilfælde er det muligt ud fra data alene at opdage, at ikke-kendte interferenser er til stede, og kalibreringsmodellen dermed er ubrugelig. Og kun i det multivariate tilfælde kan man yderligere bygge interferensinformationen ind i modellen, så den kan håndtere det. Bemærk også at outliers ikke kan detekteres univariat f.eks. fra signalet til tiden 50 sekunder.

Outro

Vi har givet en kort opsummering af nogle af de vigtige egenskaber ved multivariate data og modeller. Vi har blandt andet erfaret at

- Sensor-selektivitet er fint, men det er ikke optimalt at udvikle eller vælge sensorer, primært baseret på fysisk selektivitet. Den matematiske selektivitet opnået med egnede modelleringsmetoder bør indgå ved vurderingen af selektivitet. Dermed kan mere fokus være rettet mod signal-støjforhold, fysisk robusthed, omkostninger mv.
- Multivariate modeller giver langt flere muligheder end univariate modeller. Man kan altid droppe variable og gå tilbage til en univariat model; så man mister ikke noget ved at satse på en multivariat tilgang.
- Multivariate modeller giver en signifikant undertrykkelse af støjniveauet i analytiske målinger, hvilket alt andet lige bidrager til forbedret følsomhed og stabilitet af kalibreringsmodeller.
- Multivariate modeller kan håndtere situationer, der ikke kan håndteres af univariate modeller. I særdeleshed er det muligt at indarbejde interferenser og få automatisk outlier-detektion. Måske er en af de vigtigste egenskaber ved de typiske kemometriske multivariate modeller, at de giver mulighed for en eksplorativ tilgang. Det har vi dog ikke berørt i denne omgang.

E-mail-adresser

Rasmus Bro: rb@life.ku.dk
Søren Balling Engelsen: se@life.ku.dk
Lars Nørgaard: lno@foss.dk

Reference

1. Rasmus Bro, Multivariate calibration. What is in chemometrics for the analytical chemist?, Anal. Chim. Acta, 2004, 2003 (500), 185-190.