

Alt det gode ved at bruge mange variable

Vi ser i denne og næste klumme på nogle af de basale fordele ved at anvende multivariate metoder, og beskriver disse vha. nogle simple eksempler. Tre forskellige aspekter vil blive gennemgået: (1) støjreduktion, (2) håndtering af interferenser og (3) muligheden for outlier kontrol [1].

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

Antag, at vi ønsker at bygge en kalibreringsmodel for en specifik kemisk analyt. Vi har et måleapparat, der giver signalet i figur 1 (venstre) ved måling af tre prøver med hver sin koncentration. Hver profil består af hundrede målinger; i dette eksempel aflæsninger til forskellige tidspunkter. Med lidt fantasi kunne profilerne repræsentere et spektrum, en sensormåling, et kromatogram, FIA-gram, osv. Men helt generelt, kan hver enkelt profil opfattes som et sæt af hundrede forskellige univariate målinger (pH, koncentration, flow, osv).

For at opbygge en univariat kalibreringsmodel, må man vælge én af de hundrede variable. I et spektroskopisk eksempel kunne et typisk valg være aflæsning af signalet ved den bølglængde, der svarer til analyttens maksimumsignal. I dette eksempel er signalet til tiden 50 sekunder valgt. For at opbygge en kalibreringsmodel vha. univariat lineær regression skal nogle grundlæggende forudsætninger være opfyldt. To af de vigtigste er:

- **Selektivitet:** En univariat kalibreringsmodel kan kun give nøjagtige resultater, hvis det målte signal ikke har bidrag fra andre kilder. Derfor skal kun analytten bidrage til det målte signal. Hvis andre stoffer bidrager til signalet (interferens), vil kalibrering være umulig. Hvad værre er, så er der ikke nogen måde at opdage, at en ukendt analyt giver anledning til forkerte resultater.

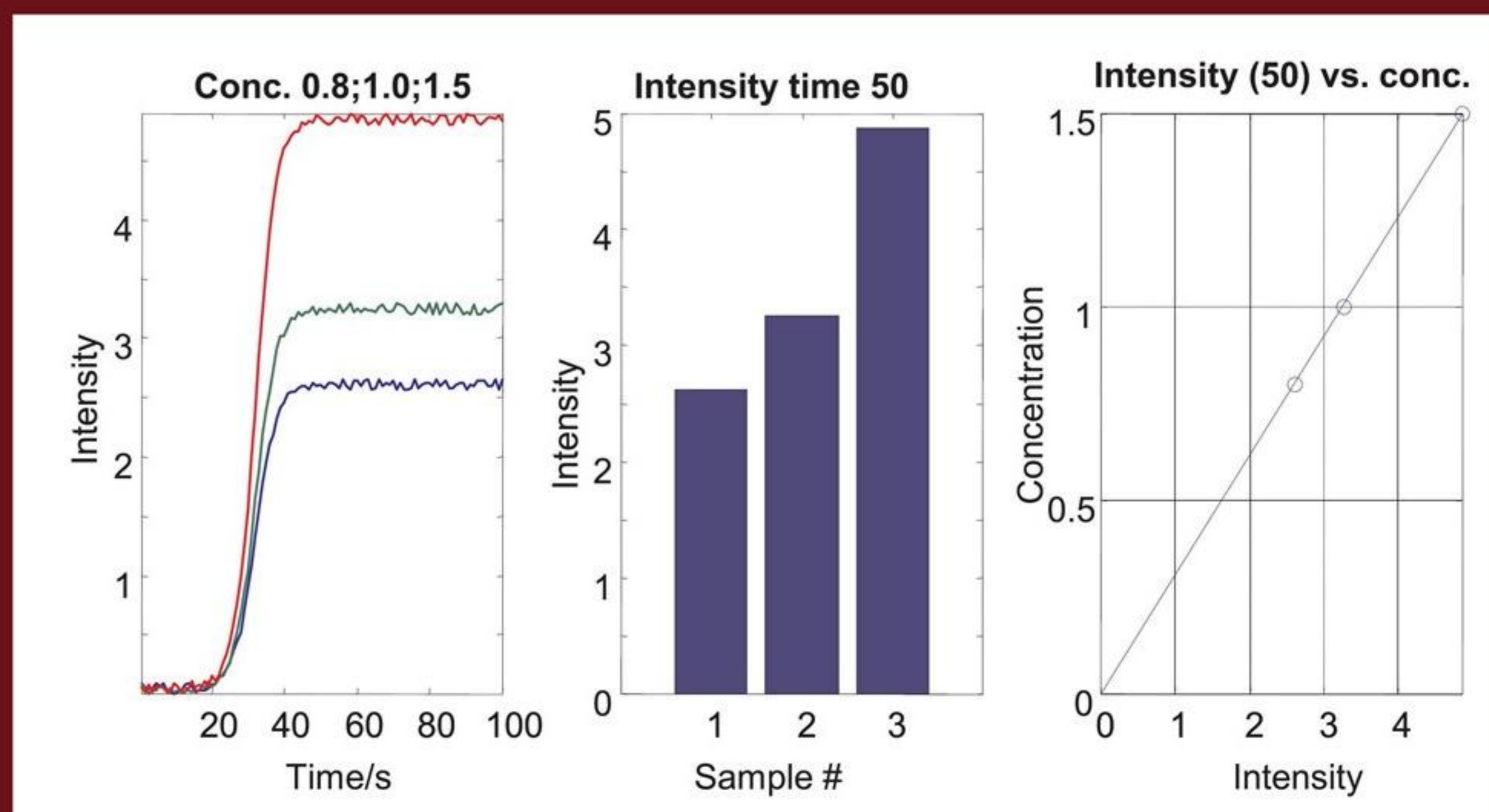
- **Linearitet:** Der skal være en lineær sammenhæng mellem analyt-koncentration og signal. Ikke-linearitet kan dog korrigeres forholdsvis nemt, når det er erkendt.

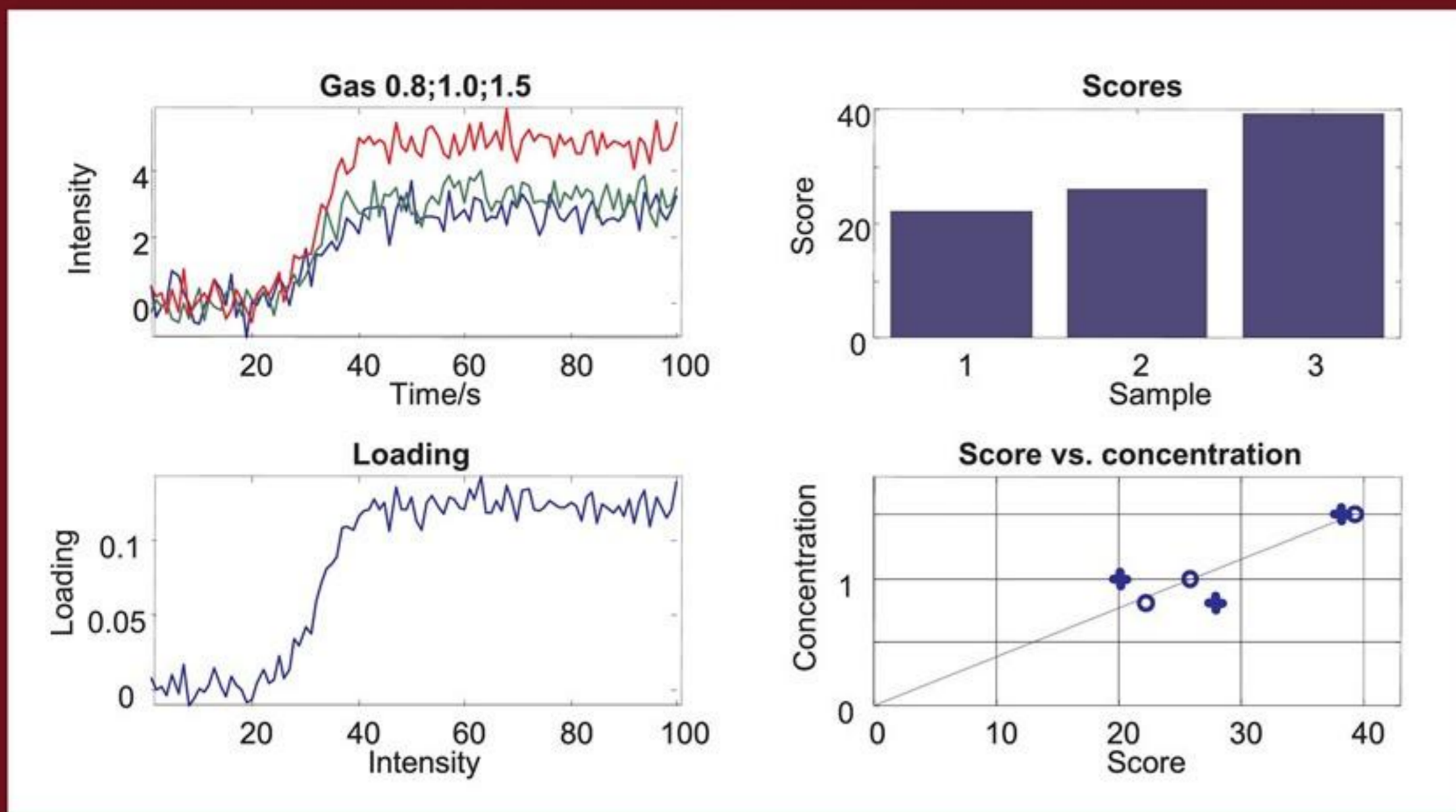
I figur 1 ses det, at signalet til tiden 50 sekunder giver en glimrende lineær kalibreringsmodel, som kan anvendes til at forudsige analyttens koncentration.

Første multivariate fordel – fjerne støj

I stedet for at bruge blot én ud af de hundrede variable giver det mening at bruge alle de målte oplysninger. Dette kan føre til en række fordele. Den mest åbenbare fordel er, at en støjreduktion opnås ved at bruge flere (redundante) målinger af samme fænomen. Dette kan illustreres vha. principal komponent analyse på data fra figur 1.

Figur 1. Univariat kalibrering. Eksempel på et multivariat signal. I dette tilfælde et signal fra en sensor målt over tid, men det kunne også være et spektrum eller et kromatogram. Til venstre ses signalet fra tre prøver med analytkoncentrationer 0,8; 1,0 og 1,5. I midten ses signalet fra de tre prøver til tiden 50 sekunder. Til højre ses sammenhængen mellem signalet til tiden 50 sekunder og analytkoncentration.





Figur 2. Støjreduktion ved multivariat kalibrering. Principal komponent analyse af et datasæt med tre prøver og 100 variable. De tre signaler øverst til venstre kan beskrives vha. én PCA-komponent. De kvantitative forskelle mellem de tre profiler er beskrevet af de forskellige scores (øverst til højre) og af loading-vektoren (nederst til venstre). Sammenhængen mellem scores og koncentration er vist med cirkler nederst til højre og med krydser er vist signalet ved tid 50 sekunder mod koncentrationen.

I figur 2 er de samme data vist men tilsat meget mere støj for at fremhæve problemstillingen. I dette eksempel med tre prøver og hundrede variable vil PCA resultere i én komponent, der beskriver de væsentligste variationer i data – den overordnede form af signalerne. Modellen er givet ved tre forskellige dele: Loading-vektoren (figur 2 nederst til venstre), som er den grundlæggende form, der beskriver alle målte signalers form. Dernæst giver modellen scores, som er de prøvespecifikke oplysninger, nemlig hvor ”meget” loading, der er i hver prøve (figur 2 øverst til højre). Den sidste del af PCA-modellen er så residualerne; den del af de enkelte profiler, der afviger fra den fælles form (ikke vist).

En alternativ og vigtig fortolkning af en sådan PCA-model er at opfatte den nye PCA-komponent, som en helt ny variabel, der erstatter de oprindelige hundrede variable. Denne nye variabel er den vigtigste variation i data, og den kan behandles på fuldstændig samme måde som enhver anden variabel. Definitionen af en original variabel, f.eks. intensitet ved tid 50 sekunder, er givet ved sin definition (intensitet til tid 50 sekunder) og de tilsvarende værdier (aflæsning af instrumentet for de tre prøver på det definerede tids-punkt). Altså en definition og en konkret måling. På samme måde er PCA-komponenten givet ved sin definition (det vægtede gennemsnit af 100 variable – loading-vektoren) og de tilsvarende aflæsninger (score-værdierne). Den vigtige forskel mellem de oprindelige variable og PCA-komponenten er, at PCA-komponenten giver information om alle originale variable samtidigt [Dansk Kemi 2 2008].

I nederste højre del af figur 2 er score-værdierne (o) plottet mod koncentrationen, hvilket viser en god korrelation. Det signal, der måles på tidspunktet 50 sekunder (+), kan ses at korrelere meget dårligt med koncentrationen. Dette skyldes støj. Ingen af de oprindelige variable korrelerer godt med koncentrationen. Det er kun vha. et (vægtet) gennemsnit af alle målinger, at en effektiv minimering af støjen opnås og giver en robust model. Denne støjreduktion er en af de klare fordele ved multivariat data-analyse.

En klippehænger:

Anden multivariate fordel – håndtere interferenser

I næste klumme viser vi, at de multivariate metoder ikke blot kan minimere støjen. De kan også give fundamentalt nye muligheder såsom at kalibrere, selv når der er interferenser og at detektere fejltagte prøver.

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.

DENIOS.

Termoteknik fra specialisten
Effektiv opvarmning, smeltning eller afkøling

- Europas største udvalg
- Udvikling og produktion
- Mere end 1.000 individuelle løsninger om året
- Professionel rådgivning
- International i 15 lande

www.denios.dk

Miljøets partner