Ti år med kemometri

Kemometri er som begreb efterhånden godt indarbejdet i danske virksomheder, hvorimod en tilsvarende indtrængen på universiteterne stadig går trægt. Den filosofiske baggrund for kemometrien beskrives, og der gives et simpelt flervariabelt taleksempel, som demonstrerer principperne ved kalibrering, prædiktion og outlier-detektion

Af Carsten Ridder, formand for Dansk Selskab for Kemometri samt Kemometricentret på Foss Electric A/S, cr@foss-electric.dk

Da jeg blev færdig som kemiingeniør fra DTU i 1985, spurgte den daværende professor i analytisk kemi Jarda Ruzicka fra Kemisk Laboratorium A mig, om jeg ikke kunne tænke mig at undersøge potentialet i at kombinere *kemometri* og *Flow Injection Analysis* (FIA), en metode, som han havde udviklet ti år tidligere sammen med Elo H. Hansen. Ruzickas hyppige ophold på University of Washington i Seattle bragte ham sammen med den fremtrædende kemometriker professor Bruce Kowalski. På trods af at Jarda havde »nedprioriteret færdigheder inden for matematik og statistik« (han har så rigeligt med andre færdigheder) indså han, at de kemometriske metoders mulighed for at trække information ud af store og uoverskuelige data kunne være en styrke i forbindelse med FIA-metoden, hvor det analytiske signal ofte er multivariabelt. Da jeg syntes, at kemometri lød spændende, og da jeg havde specialiseret mig i FIA, accepterede jeg tilbuddet og blev ansat som ph.d.-studerende.

I de følgende år opstod der på Kemisk Laboratorium A en lille »græsrodsbevægelse«, bestående af undertegnede samt entusiastiske studerende, som også havde fornemmet potentialet ved kemometrien. Når jeg kalder det for en græsrodsbevægelse, skyldes det, at jeg med undren måtte konstatere, at den kemometriske filosofi virkede provokerende på »den traditionelle videnskab« (mere om dette skisma kan læses i [1]). Det mentale skift fra den traditionelle reduktionistiske og deduktive tankegang til den kemometriske holistiske og induktive tankegang (eller groft sagt: fra hovmod til ydmyghed overfor »naturen«) frembød store vanskeligheder. I dag udbydes der kurser i kemometri på DTU, KVL og på Aalborg Universitet Esbjerg, hvorimod det stadig er særdeles vanskeligt at trænge igennem med budskabet på de »gamle« universiteter i København, Odense og Århus.

I 1991 blev *Dansk Selskab for Kemometri* (DSK) dannet med det overordnede formål at udbrede kendskabet til kemometri i Danmark, og i dag har industrien efterhånden taget kemometrien til sig. En af de store opgaver for DSK er nu at få overbevist de resterende relevante lektorer og professorer på universiteterne om fordelene ved kemometrien, så de studerende - i studietiden vel at mærke - får chancen for at tilegne sig disse begreber.

Kemometrien introduceres

På internationalt plan blev begrebet kemometri introduceret i begyndelsen af 1970'erne. Da ændrede karakteren af de kemiske analyseinstrumenter sig fra kun at give et signal pr. prøve til at give mange signaler pr. prøve (f.eks. NIR-spektre). Kemikerne, som ville analysere disse multivariable data, søgte forgæves hjælp i de gængse statistikbøger, hvor rådet konsekvent var, at man måtte fjerne variable, indtil man havde få nok til, at de statistiske metoder var brugbare (antallet af målte variable skal være mindre end antallet af prøver). Nogle stædige kemikere, som ikke ville finde sig i blot at kassere værdifuld information, fandt større hjælp inden for psykometrien. Her havde man længe opereret med begrebet »latente variable«, et begreb som skulle vise sig at blive af fundamental betydning for analysen af multivariable kemiske data. Latente variable kan opfattes som »the missing link« mellem den multivariable virkelighed og den klassiske statistiks begrænsninger.

Da kemometrien tager sit udgangspunkt i analytisk kemi, vil jeg først berøre nogle vigtige aspekter ved denne disciplin.

Analytisk kemi

Opgaven for en analytisk kemiker er at finde en metode, efter hvilken man kan bestemme koncentrationen af en analyt i en prøve. Koncentrationen skal gerne bestemmes korrekt, også når der er interfererende stoffer i prøven. Den generelle arbejdsgang er at underkaste prøven en eller flere former for forbehandling, hvorefter der måles et analytisk signal. For at fastlægge signalets afhængighed af analytkoncentrationen anvendes ofte en kalibreringskurve (eller standardkurve), som fremstilles ved at måle det analytiske signal af en række opløsninger (standarder) med kendt analytkoncentration.

For at demonstrere princippet vil jeg lave følgende analogi: I et lokale befinder der sig 100 personer, hvoraf de 88 er helt uinteresseret i politik, og de 12 er inkarnerede tilhængere af partiet Venstre. Hvis vi kalder de 12 sidstnævnte for analytten og de 88 politisk uinteresserede for opløsningsmidlet, er opgaven at bestemme koncentrationen af venstrefolk i lokalet. I dette simple tilfælde (en ren opløsning af venstrefolk – ingen interferenser) kan man blot stille et enkelt spørgsmål til forsamlingen, f.eks.: »Hvor mange er tilhængere af EU?«, og det analytiske signal er antallet af personer, der rækker hånden i vejret. Da det er sjældent, at man har rene opløsninger af analyt, forestiller vi os nu, at 20 af de politisk uinteresserede udskiftes med socialdemokrater. Hvis man måler det samme analytiske signal, vil man ikke finde 12% venstrefolk, men et sted mellem 12% og 32%, hvilket klart er et forkert resultat. Problemet er, at interferensen (socialdemokraterne) er *signalgivende* (rækker hånden op), og det analytiske spørgsmål er med andre ord ikke *selektivt*. Der er to klassiske løsninger på problemet:

1) foretag en *maskering* af interferensen (forbehandling af prøven), eller
2) anvend et selektivt spørgsmål.

En løsning efter metode 1 kunne være at stille en bar op i lokalet og råbe »Gratis øl i baren!«, hvorefter samtlige socialdemokrater vil have alt for travlt med at stå i kø til at høre, når man stiller det uselektive spørgsmål. Da en del af venstrefolkene muligvis også bliver fristet af det gode tilbud, er det nu ikke tilstrækkeligt blot at tælle oprakte hænder. Man kunne så lave en kalibreringskurve ved

Kalibrering

	Latent variabel	Laktose (%)	Absorbans (mAU) ved (cm-1):			
			1176	1377	1454	2799
	t = X*w	y	x1	x2	x3	x4
	0.0	0	0.0	0.0	0.0	0.0
	77.6	5	29.3	53.5	27.0	39.6
	155.2	10	58.6	107.0	54.0	79.2
b(t,y)	0.0644	b(y,X)	5.9	10.7	5.4	7.9
		w	0.38	0.69	0.35	0.51

t*w (estimat af x1, x2, x3 og x4)

0.0	0.0	0.0	0.0
29.3	53.5	27.0	39.6
58.6	107.0	54.0	79.2

Xres=X-t*w				kvadratsum
0.0	0.0	0.0	0.0	2.1E-06
0.0	0.0	0.0	0.0	4.6E-06
0.0	0.0	0.0	0.0	6.8E-06
		total kvadratsum =		1.3E-05

Boks 1.

at stille spørgsmålet til en række forsamlinger med kendt antal venstretilhængere, hvorefter en korrektionsfaktor kan beregnes. En løsning efter metode 2 kunne f.eks. være at stille det selektive spørgsmål: »Hvor mange ønsker Anders Fogh Rasmussen som statsminister?«, som med garanti giver det korrekte svar 12%. Hvordan bestemmer man så koncentrationen af venstretilhængere, hvis der i lokalet også er tilhængere af CD, SF, Konservative og Kristeligt Folkeparti? Ved tilstedeværelsen af disse nye interferenser, er det nu ikke nemt hverken at finde brugbare maskeringsmetoder eller ét selektivt spørgsmål. Det svarer til at prøven er meget kompleks, og desværre er det mere reglen end undtagelsen i praksis, at virkelige (kemiske) prøver indeholder mange interferenser.

Den kvikke læser vil nok indvende, at man da bare kunne spørge, hvor mange der vil stemme på partiet Venstre, men dette »avancerede analyseinstrument« er ikke opfundet i mit eksempel.

Stil flere spørgsmål ...

Den kemometriske løsning på problemet er at stille mere end et spørgsmål til forsamlingen, og det gælder generelt, at jo mere kompleks forsamlingen er, desto flere spørgsmål er det nødvendigt at stille. Det skal fremhæves, at disse spørgsmål ikke nødvendigvis skal være selektive, idet man ved udelukkelsesmetoden (og i dette eksempel megen snilde!) ville kunne finde frem til det korrekte svar. I princippet gælder det, at er der seks forskellige politiske holdninger repræsenteret i forsamlingen, er det nødvendigt med mindst seks forskellige spørgsmål. Det er indlysende, at jo flere spørgsmål man stiller, desto bedre er chancen for at nå det korrekte resultat, ligesom kravet til spørgsmålenes selektivitet mindskes. Hvis man kan finde et selektivt spørgsmål, er det naturligvis stadig tilstrækkeligt med dette ene, men denne ønskesituation er desværre sjældent opfyldt i praksis.

Jeg håber, at jeg med denne analogi har anskueliggjort, hvorfor det *multivariate* aspekt har en så fremtrædende rolle inden for kemometrien. Det erkendes simpelthen, at naturen (og dermed også en kemisk prøve) er mangfoldig og sjældent kan beskrives simpelt (énvariabelt).

Paradigmeskift

Da man indtil for et par årtier siden næsten udelukkende har rådet over énvariable analyseinstrumenter, har en af de væsentligste opgaver for analytiske kemikere gennem århundreder nødvendigvis været at opfinde prøveforbehandlingsmetoder som led i kampen mod signalgivende interferenser. Set i det lys repræsenterer kemometrien et paradigmeskift, hvilket nok er en væsentlig del af forklaringen på den træge indtrængen på de videnskabelige højborge. En anden del af forklaringen er, at multivariate - tilsyneladende kaotiske - data medfører mentalt overflow, hvis man forsøger at tolke disse med de traditionelle deduktive metoder. Med andre ord mistes det umiddelbare overblik over eksperimentelle data, som en traditionel videnskabsmand gerne vil karakteriseres ved at have. At de kemometriske metoder, som giver indblik i og overblik over dette - tilsyneladende – kaos, kræver anvendelse af matematiske metoder (primært lineær algebra på et niveau svarende til første semester på DTU) kan også virke afskrækkende på mange. Personligt mener jeg, at det er tilstrækkeligt at forstå de grundlæggende principper (filosofien), hvorefter man kan overlade de matematiske beregninger til pc'en. Man behøver ikke at kunne programmere et tekstbehandlingsprogram selv, før man kan skrive et brev!

Da der efter min mening er tale om et paradigmeskift, gælder det ikke kun om *uddannelse* af nye generationer, men i høj grad også om *opdragelse*. Kemometri er en anden måde at tænke på, nemlig en erkendelse af, at multivariable data ofte er manifestationen af en unik kombination af nogle få underliggende (latente) fænomener, og at de kemometriske metoder kan kaste lys over og udnytte disse. Skal kemometri så erstatte den klassiske analytiske kemi? Absolut nej, f.eks. er multivariable kalibreringsmetoder

afhængige af eksistensen af kemiske referencemetoder. På samme måde kan kemometri heller ikke erstatte den klassiske statistik. Kemometri skal opfattes som en supplerende ny og holistisk indfaldsvinkel til analyse af multivariable kemiske data.

Jeg vil til sidst vove den påstand, at en betragtelig del af de analyseresultater, baseret på énvariable metoder, der dagligt produceres i alverdens laboratorier er decideret forkerte. Det opdages bare ikke, fordi énvariable metoder lider af den alvorlige skavank ikke at kunne detektere afvigende prøver: når man har brugt den ene frihedsgrad til at bestemme analyseresultatet, er der ikke ingen tilbage til at teste for resultatets validitet. Det er endnu en væsentlig grund til, at vi advokerer for udbredelse af kemometriske metoder. I næste afsnit vil jeg give et lille taleksempel, som viser, hvordan man kan analysere data, hvor der er flere målte variable end prøver, og hvordan den multivariable måling kan udnyttes til at bestemme afvigende prøver. Det demonstrerer samtidigt, hvordan man kan finde en robust og brugbar løsning på tre ligninger med fem ubekendte.

Spektroskopisk måling af laktose

De ovenfor nævnte principper demonstreres nu ved at overføre dem til spektroskopisk måling af laktose i vandige opløsninger (f.eks. mælk). Et »spørgsmål« kan i dette tilfælde eksempelvis være: »Hvor meget lys absorberer prøven ved bølgelængden 1176 cm^{-1}?«. Opgaven er her at bestemme laktosekoncentrationen i fire »ukendte« prøver a, b, c og d. Prøve a og c indeholder kun laktose (2% og 6%), prøve b indeholder ud over 4% laktose også 0.5% fedt, og prøve d indeholder ikke laktose men 2% fedt.

Den klassiske fremgangsmåde er at fremstille en række standardopløsninger med kendt indhold af analytten (laktose), udmåle disse ved en given bølgelængde og på basis heraf fremstille en kalibreringskurve. I dette tilfælde er der fremstillet tre standarder med henholdsvis 0%, 5% og 10% laktose, og disse tre standarder er målt ved de fire bølgelængder i det infrarøde område: 1176, 1377, 1454 og 2799 cm^{-1}. Man udvælger typisk den bølgelængde, som har størst sensitivitet over for analytten, men jeg vil her vise, hvordan man kalibrerer med alle fire bølgelængder på en gang.

Kalibrering

Resultaterne af standardudmålingerne er vist i boks 1 (»Kalibrering«), hvor der er en vandret datamatrix på 3x4 elementer.

Tabellen af de fire målte variable x_1 til x_4 kaldes **X**. Det betyder umiddelbart, at man ikke kan analysere denne med en klassisk statistisk metode (f.eks. MLR: multipel lineær regression), hvor det kræves, at antallet af prøver er mindst det samme som antallet af variable. Denne begrænsning kan tilskrives det forhold, at man i realiteten er interesseret i at bestemme de fem koefficienter b_j, j=0..5 i formlen:

$$\%\text{laktose} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \qquad (1)$$

og at man til dette formål kun råder over tre ligninger. Finder man blandt uendeligheden af løsninger et godt bud på b_j, vil man kunne bruge formel (1) til at bestemme laktose i ukendte prøver, hvis man måler absorbansen ved de samme fire bølgelængder, som standarderne er målt ved.

Hvordan analyserer man så denne »håbløse« datatabel? Det virker intuitivt fornuftigt at bestemme de enkelte variables respons på ændringen i laktosekoncentrationen, hvilket kan oversættes til hældningskoefficienten i afbildningen af absorbansen mod laktosekoncentrationen. I boks 1 er disse anført som »b(y,X)«, og som det ses, ligger sensitiviteterne mellem 5.4 og 10.7 mAU pr. %laktose. Ifølge den kemometriske filosofi opfatter man herefter disse hældningskoefficienter som »kvalitetsmål«, idet de anvendes som vægte på de oprindeligt målte variable. Af numeriske grunde skaleres de først ned, så kvadratsummen af vægtene bliver en, hvilket her er gjort ved at dividere alle b(y,X) med ca. 15.5. De herved fremkomne skalerede vægte kaldes for **w**, og er fremhævet i boks 1. Vha. disse udregnes der herefter for hver prøve en *latent variabel* (kaldet **t** i boks 1), som er vægtede summer af den originale variabel og den tilhørende w. F.eks. fås for standarden med 10% laktose: t = 0.38*58.6 + 0.69*107.0 + 0.35*54.0 + 0.51*79.2 = 155.2. Hermed er problemet reduceret fra at være 4-dimensionalt til at være 1-dimensionalt, og man kan nu lave sædvanlig univariat lineær regression mellem t og y (laktosekoncentrationen). Hældningskoefficienten i afbildningen y mod t kan udregnes til 0.0644 og kaldes for »b(t,y)« i boks 1. Det skal her bemærkes, at den latente variabel **t** indeholder bidrag fra *samtlige* originale variable, og at man derfor *ikke* har kasseret potentiel vigtig kemisk information. Skulle en målt variabel udvise ringe effekt af ændringer i analytkoncentrationen, ville denne variabel blot få en lavere vægt.

Da **t** som nævnt beregnes som **X***w, kan man også regne baglæns og få et estimat for **X** ved at gange **t** med **w** (det ydre

Prædiktion

Modelparametre:

w	0.38	0.69	0.35	0.51
b(t,y)	0.0644			
residual	1.3E-05			

Absorbans (mAU) ved (cm-1):

	1176	1377	1454	2799			
tpred = x*w	x1	x2	x3	x4	prøve	Laktose	Fedt
31.0	11.7	21.4	10.8	15.9	a	2.0	0.0
67.2	30.2	42.6	24.1	35.2	b	4.0	0.5
93.1	35.2	64.2	32.4	47.5	c	6.0	0.0
20.3	26.9	-0.7	9.9	14.0	d	0.0	2.0
0.00	0	0	0	0	b0		
0.38	1	0	0	0	b1		
0.69	0	1	0	0	b2		
0.35	0	0	1	0	b3		
0.51	0	0	0	1	b4		

Beregnet laktosekoncentration:

prøve	tpred*b(t,y)	
a	2.0	OK
b	4.3	outlier
c	6.0	OK
d	1.3	outlier
b0	0.0000	
b1	0.0243	
b2	0.0444	
b3	0.0224	
b4	0.0329	

%laktose = b0 + b1*x1 + b2*x2 + b3*x3 + b4*x4

prøve	tpred*w (estimat af x1, x2, x3 og x4)			
a	11.7	21.4	10.8	15.8
b	25.4	46.3	23.4	34.3
c	35.2	64.2	32.4	47.5
d	7.7	14.0	7.1	10.4

prøve	Xres=X-tpred*w				residual	
a	0.0	0.0	0.0	0.0	2.4E-06	OK
b	4.8	-3.7	0.7	0.9	37.97	outlier
c	0.0	0.0	0.0	0.0	5.1E-06	OK
d	19.3	-14.7	2.9	3.6	607.70	outlier

Boks 2.

produkt). Det er gjort i boks 1, og i dette simple tilfælde (simulering af en analyt med 1e-5 støj lagt oveni) finder man, at **t*****w** næsten er identisk med **X**. Kvadratsummen af elementerne i residualmatricen **X** – **t*****w** er derfor tæt på nul og af samme størrelsesorden, som den tilfældige støj jeg har lagt på, nemlig 1.3e-5 (for overskuelighedens skyld er decimalerne udeladt i tabellen). Havde der i standarderne været mere end et signalgivende stof, ville residualmatricen have været forskellig fra nul, og man måtte beregne en ekstra **w**, **t** og b(t,y). Det gøres på samme måde som netop beskrevet, men man baserer blot beregningerne på denne residualmatrix. Nu er den multivariable kalibreringsmodel færdigberegnet og består af de i boks 1 fremhævede parametre.

Prædiktion

I boks 2 (»Prædiktion«) er modelparametrene samlet øverst, og herunder er signalerne målt på de ukendte prøver anført (samt

nogle »prøver« b0 til b4, som jeg vender tilbage til nedenfor). Ved simuleringen er Lambert-Beers lov udnyttet samt at en opløsning af 1% laktose absorberer henholdsvis 5.86, 10.70, 5.40 og 7.92 mAU ved de fire bølgelængder, og at en opløsning af 1% fedt absorberer henholdsvis 13.46, -0.34, 4.96 og 7.01 mAU. Den negative absorption ved 1377 cm^{-1} skyldes, at fedt er mere transparent for stråling af denne bølgelængde end nulstillingsvæsken, som er almindeligt vand.

På samme måde som før beregnes for hver prøve en latent variabel (**tpred** i boks 2), hvor **w** fra kalibreringsfasen anvendes. For prøve »a« fås f.eks.: tpred = 0.38*11.7 + 0.69*21.4 + 0.35*10.8 + 0.51*15.9 = 31.0. Da vi ved kalibreringen fandt ud af, at laktosekoncentrationen kunne bestemmes som b(t,y)*t, findes resultatet for prøve »a« til 0.0644*31.0 = 2.0% laktose. På samme måde beregnes koncentrationen i de øvrige prøver, og det fremgår af boks 2, at de rene laktoseopløsninger bestemmes korrekt, og at de to prøver med signalgivende interferens bestemmes forkert. Laktosekoncentrationen i de fem prøver b0 til b4 kan bestemmes ved samme procedure, og ved at indsætte de anførte måleværdier for x_1 til x_4 i formel (1) ovenfor kan man se, at disse prædiktioner netop svarer til de ønskede regressionskoefficienter (når $b_0 = 0$, ellers skulle man trække b_0 fra resultatet). Prøve »a« kunne med andre ord også bestemmes på den alternative – og hurtigere - måde: ypred = 0.0000 + 0.0243*11.7 + 0.0444*21.4 + 0.0224*10.8 + 0.0329*15.9 = 2.0% laktose.

Outlier-detektion

Hvis vi standsede her, ville vi stå med to rigtige og to forkerte resultater, og vi ville ikke vide, hvilke vi kunne stole på. Men på samme måde som i kalibreringsfasen kan vi nu regne baglæns og for hver prøve estimere de originale målinger ved at gange **w** med tpred. For prøve »b« fås f.eks.: 67.2*[0.38 0.69 0.35 0.51] = [25.4 46.3 23.4 34.3], og hvis dette sammenlignes med det egentligt målte, får man residualet: [30.2 42.6 24.1 35.2] - [25.4 46.3 23.4 34.3] = [4.8 -3.7 0.7 0.9], som har kvadratsummen 37.97 (se nederst i boks 2). Dette residual kan sammenlignes med residualet fundet for kalibreringsprøverne (1.3e-5) og - med et antal frihedsgrader som jeg har tillid til, at programmørerne har regnet rigtigt ud for mig – kan man statistisk teste om prøvens residual er signifikant større end 1.3e-5. Er dette tilfældet, kan man konkludere, at denne prøve ikke ligner de prøver, som kalibreringen er baseret på, og at man ikke kan stole på resultatet. Ved anvendelse af et sådan test, er der ud for prædiktionerne anført enten »OK« eller »outlier«.

Som analytisk kemiker synes jeg, at det er særdeles betryggende, at man på denne måde kan afsløre fejlagtige resultater. Omvendt ville jeg føle mig på gyngende grund, hvis jeg *ikke* havde denne mulighed for detektion af afvigende prøver.

Da det er utilfredsstillende, at man ikke kan bestemme laktosekoncentrationen i prøver, som indeholder fedt, vil man være nødt til at måle flere kalibreringsprøver og beregne en ny model, som er ufølsom over for signaler fra fedt. For disse kalibreringsprøver skal det gælde, at koncentrationen af både laktose og alle interferenser (her kun fedt) varierer i det område, man forventer, at fremtidige prøver vil ligge i. Herudover skal det gælde, at disse koncentrationer skal variere uafhængigt af hinanden. Eksempelvis kan man således ikke blot anvende en fortyndingsrække af en prøve, idet man herved kun ville modellere komponenten »analyt+interferenser«. Til gengæld er det ikke nødvendigt at kende den nøjagtige koncentration af interferenserne i de nye kalibreringsprøver, og man vil normalt løse problemet ved at måle et tilstrækkeligt stort antal naturlige prøver, således at dette krav med stor sandsynlighed er opfyldt.

Ved at anvende ovennævnte beregningsmetode på kalibreringsprøver, som udover laktosevariation har et varierende indhold af fedt kan regressionskoefficienterne findes til:

$$\%\text{laktose} = 0.0000 - 0.0171*x_1 + 0.0761*x_2 + 0.0163*x_3 + 0.0251*x_4 \quad (2)$$

I dette tilfælde skal man beregne to latente variable, svarende til antallet af signalgivende stoffer i kalibreringsprøverne. For fuldstændighedens skyld har jeg i formel (3) anført resultatet for prøvetyper, som ud over laktose og fedt også indeholder protein. Her er der – naturligvis - anvendt tre latente variable:

$$\%\text{laktose} = 0.0000 + 0.0006*x_1 + 0.0891*x_2 - 0.0862*x_3 + 0.0641*x_4 \quad (3)$$

Læseren opfordres til at beregne laktosekoncentrationen i prøverne i boks 2 vha. formel (2) og (3).

Det fremgår af ovennævnte eksempel, at antallet af »spørgsmål« (her: målte bølgelængder) sagtens kan være større end antallet af prøver, idet man fra disse mange variable udtrækker *latente variable* i et antal, der svarer til prøvernes kompleksitet. Derfor er det f.eks. muligt at udnytte samtlige 700 værdier i et NIR-spektrum, selvom man kun råder over 60 prøver.

Kalibreringsmetoden, som er gennemgået i dette afsnit kaldes PLS-regression (Partial Least Squares Regression eller Projektion på Latente Strukturer) og kan findes nærmere beskrevet i litteraturen. Jeg vil især anbefale [2], som også giver en fremragende indføring i kemometriske grundbegreber.

Kemiske referencemetoder er ofte kendetegnet ved at være tids- og arbejdskrævende og have et højt reagensforbrug, da man er nødsaget til at maskere eller fjerne interferenser. Vha. moderne multivariate analyseinstrumenter kombineret med kemometrisk dataanalyse (f.eks. PLS) kan disse erstattes af hurtige, billige og mindre miljøbelastende metoder.

Latentik

Jeg har i denne artikel primært fokuseret på udnyttelsen af multivariable data til kalibrering. Et andet stærkt aspekt ved kemometri er *eksplorativ dataanalyse*, hvor man uden først at opstille hypoteser (altså induktivt) får indblik i det multivariable kaos via afbildninger af de latente variable (**t**, kaldet »scores«) og de tilhørende vægte (**w**, kaldet »loading weights«).
Igennem de sidste 10-15 år har jeg ofte demonstreret dette eksplorative aspekt med eksempler hentet fra så forskellige områder som politik, musik, kemi, digtning, medicin, demografi, fodbold m.v. I alle disse tilfælde kan svært overskuelige datatabeller omsættes til informative afbildninger, som – ud over at bekræfte kendt teori – ofte giver anledning til udledning af hypoteser om sammenhænge, som ikke før var erkendt. Analyseprincippet er således også hypotesegenererende, hvilket ud fra en videnskabelig synsvinkel måske endda er det vigtigste.
Da det virker forkert at bruge udtrykket »kemometri« i forbindelse med analyse af f.eks. fodbold- og valgresultater, har jeg følt mig foranlediget til at indføre begrebet: »latentik«, som kan udlægges som »læren om det skjulte«. Dette begreb kunne anvendes som en overordnet betegnelse for de videnskabelige grene, som erkender og tager konsekvensen af naturens kompleksitet, og dækker således over f.eks. psykometri, økonometri og – naturligvis – kemometri. Det er mit store personlige håb, at andre videnskabelige grene efterhånden tager denne filosofi til sig.

Referencer
1. Munck L, Nørgaard L, Engelsen SB, Bro R, and Andersson CA.: *Chemometrics in Food Science – a Demonstration of the Feasibility of a Highly Exploratory, Inductive Evaluation Strategy of Fundamental Scientific Significance*. Chemometrics and Intelligent Laboratory Systems 44 (1-2):31-60, 1998.
2. Rasmus Bro, *Håndbog i multivariabel kalibrering*, Jordbrugsforlaget, 1996

Links:
KVL kemometrigruppe; software, kurser, mm: www.models.kvl.dk
Links til alt om kemometri: www.wiley.com/chemometrics
Database m. 6000 artikler: www.models.kvl.dk/ris/