Metoder og ideer i moderne kemometri

Her gennemgås nogle af de grundlæggende ideer og metoder, der anvendes i kemometrien. Fremstillingen er noget matematisk, men det er ikke nødvendigt at forstå matematikken for at få indblik i den arbejdsmetodik, der anvendes i den moderne kemometri

Af Agnar Höskuldsson, IPL, DTU, ah@akp.dtu.dk

Kemometri er et relativt nyt fagområde. Interessen for faget er steget en del de seneste år især pga. den forretningsmæssige succes som mange virksomheder, både i Danmark og i udlandet, har opnået ved at bruge disse metoder.

Mange fagfolk, både inden for kemi og statistik, er og har været kritiske over for de anvendte metoder. Det skyldes i høj grad et manglende kendskab til faget.

Store datamængder

Inden for industrien er der i øjeblikket en kraftig udvikling inden for dataopsamling. Det skyldes, at det er blevet billigt at indsamle data on-line, og at der er kommet nyt avanceret måleudstyr, der giver mange værdier for hver prøve, der måles på. Virksomheder har brug for at analysere store datamængder med variable, der ofte er meget korrelerede. Her har kemometrien vist sig som et effektivt værktøj til at skabe indhold i data, som umiddelbart er uoverskuelige at håndtere.

NIR-data

Der er stor interesse for at arbejde med NIR (Near-Infra Red)-data [1,2].

De foreliggende data er hentet fra ølbrygning (Carlsberg). Opgaven er, at undersøge om NIR-data kan benyttes til at forudsige kvaliteten af det producerede øl. Opgaven kan formuleres som at løse et lineært ligningssystem, $\mathbf{XB}\cong\mathbf{Y}$. Dataværdierne i en række i $\mathbf{X}$ er resultaterne af en prøve af NIR-målinger. Den tilhørende værdi i $\mathbf{Y}$ er den observerede kvalitetsparameter, der fokuseres på. Det anvendte måleudstyr giver automatisk 1050 dataværdier for en prøve svarende til de 1050 bølgelængder. Da dataværdierne i starten og slutningen af signalet ikke altid er pålidelige, afkortes det i begge ender til 926 dataværdier.

I alt er der undersøgt 61 prøver af det producerede øl. Det betyder, at matricen $\mathbf{X}$ er 61x926 og $\mathbf{Y}$ er en 61-vektor. Tegnes rækkerne i $\mathbf{X}$ op, vil punkterne danne 61 bløde kurver. Det er karakteristisk for mange moderne måleinstrumenter, at dataene vist på denne måde giver en samling bløde matematiske kurver. I praksis betyder det ofte, at bølgelængder (variable) i nærheden af hinanden i $\mathbf{X}$ er stærkt korrelerede (måleværdier umiddelbart efter hinanden har næsten samme værdi). En nærmere analyse af data viser, at det er fordelagtigt kun at arbejde med dataværdier i søjlerne (bølgelængder) i $\mathbf{X}$ mellem 401 og 500.

I det følgende vil $\mathbf{X}$ have størrelsen 61x100. Opgaven er at bestemme en løsningsvektor $\mathbf{B}$ til $\mathbf{XB}\cong\mathbf{Y}$, så der opnås så god en forudsigelse $\hat{Y}$, $\hat{Y}=\mathbf{B}^T\mathbf{x}$ af kvalitetsparameteren som muligt, når resultatet $\mathbf{x}$ af en prøve fra NIR-instrumentet foreligger.

Data i kemometrien er ofte *flade*, som her ved NIR-dataene. Dvs. der er flere variable end prøver. Raman-spektroskopi kan give 3300 værdier for hver prøve. Moderne spektroskopiske måleinstrumenter dækker brede bølgelængdeområder og giver derfor mange måleværdier for hver prøve.

PLS-regression

En populær metode er PLS (Partial Least Squares) regression. Den grundlæggende ide er at bestemme de latente forhold i $\mathbf{X}$-data, som genererer de responsdata, $\mathbf{Y}$-data, der er observeret. PLS var længe kendt som en algoritme, der var svær at gennemskue. Et gennembrud kom i 1988 [3], hvor det blev vist, at PLS er baseret på betragtninger, der er kendt i den statistiske teori. Det blev vist, at PLS maximerer *kovariansen* mellem en *score-vektor* for $\mathbf{X}$ og en *score-vektor* for $\mathbf{Y}$. En score-vektor $\mathbf{t}$ for $\mathbf{X}$ er givet ved $\mathbf{t}=\mathbf{Xw}=w_1\mathbf{x}_1+ w_2\mathbf{x}_2+\dots$, hvor $\mathbf{X}= (\mathbf{x}_1,\mathbf{x}_2,\dots)$. Den er således en vægtet sum af variablene (søjlerne) i $\mathbf{X}$. Tilsvarende er en score-vektor for $\mathbf{Y}$ givet ved $\mathbf{u}=\mathbf{Yq}=q_1\mathbf{y}_1+ q_2\mathbf{y}_2+\dots$

1^0 Find $\mathbf{w}$ og $\mathbf{q}$, begge af længden 1, så kovariansen mellem $\mathbf{t}=\mathbf{Xw}$ og $\mathbf{u}=\mathbf{Yq}$ er maximeret, max $(\mathbf{t}^T\mathbf{u})$.

2^0 Juster $\mathbf{X}$ for den fundne $\mathbf{t}$ skorvektor så den nye $\mathbf{X}$ bliver ortogonal til skorvektoren. Tilsvarende justeres $\mathbf{Y}$ for den fundne $\mathbf{t}$.

3^0 Hvis flere par af skorvektorer skal uddrages, startes igen på 1^0.

Boks 1. Fremgangsmåde ved PLS.

Fremgangsmåden ved PLS-regression er vist i boks 1. Ideen er at maximere kovariansen, $(\mathbf{t}^T\mathbf{u})=t_1u_1+t_2u_2+.$, der kan opnås fra $\mathbf{X}$ og $\mathbf{Y}$. Det viser, at $\mathbf{X}$-data og $\mathbf{Y}$-data får samme vægt, skønt opgaven er at forudsige Y-værdierne ud fra kendskab til X-værdierne.

I multivariabel statistik haves metoden Kanonisk Korrelationsanalyse, hvor korrelationen maximeres mellem de to score-vektorer fra $\mathbf{X}$ og $\mathbf{Y}$. Men ellers er fremgangsmåden analog. Det kan anbefales at skalere data forud for PLS-regression, så alle variable har middelværdi 0 og varians 1. Skaleres data på

denne måde (autoskalering), bliver det første par af score-vektorer, **t** og **u**, i PLS næsten det samme som i Kanonisk Korrelationsanalyse.

Balancering af »tilpasning« og »præcision«

Numerikere, som har arbejdet med ortogonale funktioner, ved, at det er svært at finde funktioner, som giver stabile forudsigelser. Problemet er, at funktionerne alligevel godt kan passe til de foreliggende data, men bevæger vi os ganske lidt fra de foreliggende værdier, kan funktionerne give estimater, som ikke er tilfredsstillende. I kemometrien ønskes modeller, som giver så stabile forudsigelser som muligt. Ved modellering af industrielle data er forudsigelsesvariansen i fokus. Ser vi på en standardregressionsanalyse, hvor data beskrives ved en lineær model og residualerne (afvigelserne $y-(\beta_1x_1+\beta_2x_2+\ldots)$) er uafhængige normalfordelte størrelser, er variansen på forudsigelser givet ved formlen i boks 2. Det betyder, at når der foreligger værdier af en ny prøve $\mathbf{x}_0$, udregnes en tilhørende y-værdi som $y(\mathbf{x}_0)$.

Virksomheden ønsker, at usikkerheden, $Var(y(\mathbf{x}_0))$, der knytter sig til bestemmelsen af y-værdien, er så lille som muligt, dvs. at variansen på en forudsigelse, $Var(y(\mathbf{x}_0))$, bliver så lille som mulig. I udtrykket for $Var(y(\mathbf{x}_0))$ indgår der to størrelser - udtryk for tilpasningen og udtryk for præcisionen. Modelleres data, fås en score-vektor **t**.

Variansen på $y(\mathbf{x}_0)=\mathbf{b}^T\mathbf{x}_0= b_1x_1+b_2x_2+..$, hvor **b** er de estimerede regressions-koefficienter, og $\mathbf{x}_0$ er værdierne på en ny prøve, er givet ved

$$Var(y(\mathbf{x}_0)) = s^2\,\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{x}_0$$

$$= |\mathbf{y}-\mathbf{X}\mathbf{b}|^2\,\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{x}_0/(N-K)$$

Boks 2. Variansen på forudsigelse af y-værdi for givne værdier af en ny prøve $\mathbf{x}_0$.

Hvad er en god score-vektor? Som vist i boks 3 er der to grundlæggende aspekter. Der ønskes så god forbedring af tilpasningen som muligt. Samtidig ønskes det, at den præcision, der fås ved benyttelse af score-vektoren bliver så god som muligt. Forbedringen ønskes så stor som mulig og udtrykket for præcisionen så lille som mulig. De to ønsker kan ikke opnås samtidig. Det kan vises, at de to ønsker statistisk set (under normale forudsætninger) er uafhængige af hinanden. Begge størrelser kan ikke optimeres samtidig, og der skal tages hensyn til begge forhold for at gøre variansen på forudsigelser så lille som mulig.

Forbedring i tilpasning: $|\mathbf{Y}^T\mathbf{t}|^2/(\mathbf{t}^T\mathbf{t})$

Tilhørende præcision: $1/(\mathbf{t}^T\mathbf{t})$

*Boks 3. To aspekter af modellering for given score-vektor **t**.*

En mulig fremgangsmåde er at maximere produktet af de størrelser, som er vist i boks 4. Det kan vises, at PLS netop maksimerer udtrykket i boks 4. Det betyder, at man i PLS søger efter en score-vektor, der sikrer balance mellem forklaring af **Y**-data og den præcision, det medfører, hvis score-vektoren benyttes. Betragtningen kan anvendes ved enhver form for matematisk modellering af data. Den kaldes H-princippet, hvor H står for Heisenberg [4], pga. den nære analogi der er mellem modelleringsopgaven og Heisenbergs usikkerhedsprincip. Ved matematisk modellering af industrielle data er princippet særlig vigtigt, fordi der er fokus på forudsigelser: når der foreligger en ny prøve, estimeres den tilhørende Y-værdi så præcist som muligt.

Find en vektor **w** således at udtrykket

{Forbedring i tilpasning}×{1/Præcision}

$$= \{|\mathbf{Y}^T\mathbf{t}|^2/(\mathbf{t}^T\mathbf{t})\}\times\{1/(1/(\mathbf{t}^T\mathbf{t}))\} = |\mathbf{Y}^T\mathbf{t}|^2$$

$$= \mathbf{w}^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}$$

bliver maksimeret.

Boks 4. Balancering af tilpasning og præcision bliver maksimeret.

Latente strukturer

De score-vektorer, som bestemmes ved analysen, samles i en matrix $\mathbf{T}=(\mathbf{t}_1, \mathbf{t}_2,..,\mathbf{t}_A)$. Matricen **T** udgør den latente struktur i data, som anvendes ved forudsigelser af Y-værdier. Man kan sige, at den viser de »profiler«, der kan dannes i data, og som er karakteristiske for de variationer, der er i **Y**-data. Ved dataanalysen tegnes score-vektorerne op imod hinanden. Prøver, som ligner hinanden spektralt, bør også ligge tæt på hinanden i disse score-plots. Score-vektorerne er ifølge konstruktionen ortogonale på hinanden. Er der normale forhold i data, bør et sådant scatter-plot vise en punktsværm, der ligger inden for en ellipse med koordinatakserne som akser. Disse scatter-plot studeres med henblik på at opdage grupperinger i

Nr	D(**X**)	SD(**X**)	D(**Y**)	SD(**Y**)
1	96.12	96.12	94.31	94.31
2	1.93	98.05	4.87	99.18
3	0.38	98.43	0.24	99.42
4	0.13	98.56	0.17	99.59
5	0.09	98.65	0.12	99.71
6	0.10	98.75	0.05	99.76
7	0.09	98.84	0.03	99.79
8	0.09	98.93	0.03	99.82
9	0.12	99.05	0.01	99.83
10	0.10	99.14	0.01	99.84

*Tabel 1. Procentvis uddraget variation af **X** og **Y**.*

data, outliers (ekstreme score-værdier), ulinearitet og lignende forhold. Det er en væsentlig del af kemometrien at arbejde med sådanne scatter-plots. De viser, hvordan prøverne forholder sig i forhold til hinanden i lyset af den opgave, der foreligger (regressionsanalyse eller andre typer opgaver).

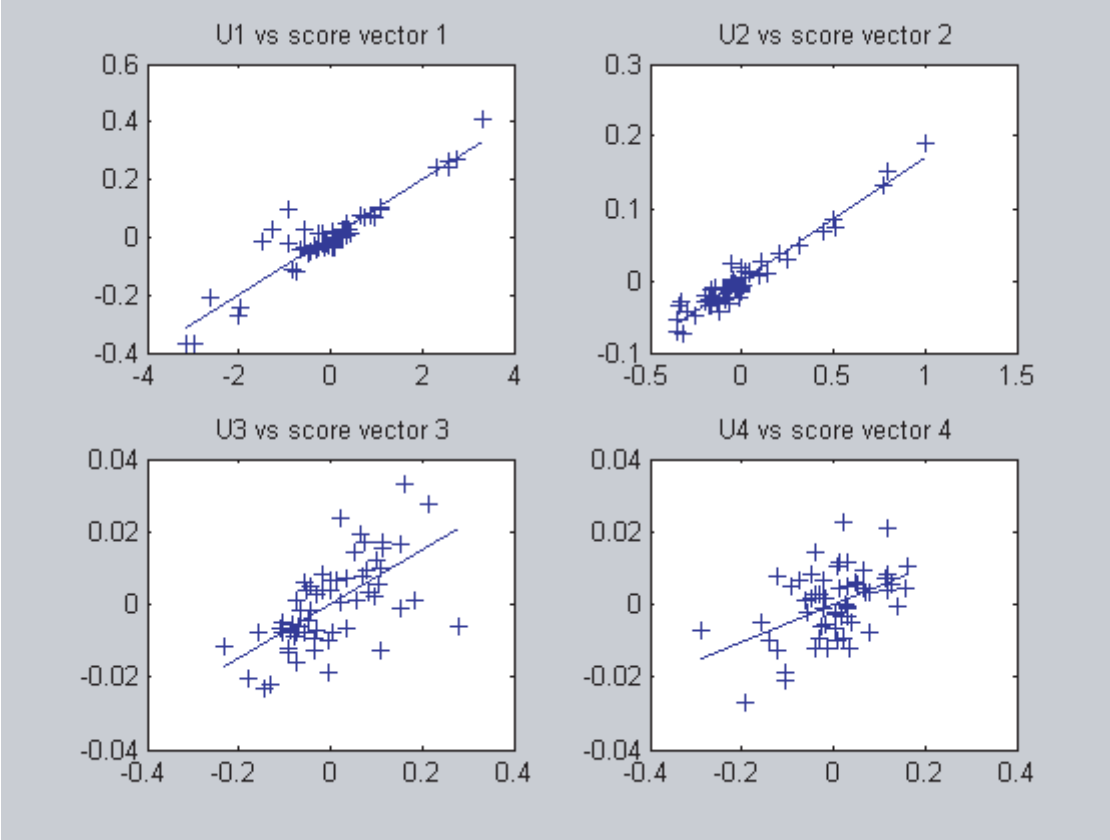
Uddraget variation

Mht. NIR-data er det første, som har interesse, hvor stor variation, der er uddraget på hvert trin. Tabel 1 viser den uddragne variation for de første 10 score-vektorer. Det ses, at den første score-vektor forklarer 96.12% af variationen i **X**. Hvis $\mathbf{X}_1$ er **X**-data, der er reduceret med den første score-vektor, udregnes forklaret variation som 96.12%= $100\text{xtr}(\mathbf{X}_1^T\mathbf{X}_1)/(\text{tr}(\mathbf{X}^T\mathbf{X})$, hvor tr() er sporfunktionen. Den første score-vektor forklarer 94.31% af variationen i kvalitets-parameteren **Y**. De første tre score-vektorer forklarer 98.43% af **X** og 99.42% af **Y**. Score-vektor nummer 4 bidrager kun med 0.13% til variationen i **X** og 0.17% til variationen i **Y**. Man vil således være meget forsigtig med at anvende score-vektor nummer 4 og de senere score-vektorer, fordi de bidrager meget lidt til forklaring af variationen i **X** og **Y**.

Grafiske tegninger

Et vigtigt aspekt af kemometrien er diverse plots, der belyser variationen i data. Blandt de vigtigste kan nævnes

- *Plot af score-vektorer parvis*. Plot af $\mathbf{t}_i$ mod $\mathbf{t}_j$ viser, hvordan den latente struktur ser ud. Opdages grupperinger i den latente struktur, vurderes denne gruppering af data.
- *Observeret mod beregnet Y-værdi*. De **Y**-værdier, der er observeret, tegnes op mod de tilsvarende beregnede værdier $\hat{\mathbf{Y}}$. I den forbindelse ses på en række størrelser med henblik på at vurdere resultaterne.
- ***Y**-værdier mod score-vektorer*. **Y** (eller $\mathbf{u}_i$) tegnes op mod score-vektorerne, $\mathbf{t}_i$. De viser kvaliteten af tilpasningen på de enkelte trin.



*Figur 1. Plot af reducerede **Y**-værdier ($\mathbf{u}_i$-erne) mod de første fire score-vektorer.*

- *Plot af ladningsvektorer parvis*. Ladningsvektorer plottes mod hinanden for at studere den korrelationsstruktur, der er mellem variablene.
- *Plot af kausale vektorer*. De kausale vektorer $\mathbf{r}_i$ transformerer de originale vektorer (variable) til score-vektorer (score-variable), $\mathbf{t}_i=\mathbf{X}\mathbf{r}_i$. Der ses dels på de enkelte kausale vektorer og dels på dem parvis for at studere, hvordan den latente struktur er afledt af de originale variable. Disse plots giver overblik over, hvilke variable der er vigtige, og hvilke der eventuelt kan undværes.

Ved at studere scatter-plots, som måske er forsynet med oplysende tekster ved de enkelte punkter, fås et godt indblik i datavariationen. En detaljeret analyse af disse plots giver godt kendskab til de enkelte prøver, og hvordan de indgår i analysen, samt tillid til at analysen er pålidelig.

Plot af Y-værdier mod score-vektorer

Der skal her vises et af de ovennævnte plots, nemlig scatter-plots af Y-værdier eller mere præcist de reducerede Y-værdier ($\mathbf{u}_i$-erne) mod score-vektorerne.

Den øverste tegning til venstre viser Y-værdierne (værdierne af kvalitetsparameteren) mod den første score-vektor. I den øverste figur til højre er Y-værdierne justeret for den første score-vektor. De første to score-vektorer viser nogenlunde pæn lineær sammenhæng med Y-værdierne. Den nederste figur til højre viser de reducerede Y-værdier på y-aksen og den fjerde score-vektor på x-aksen. Den viser meget svag sammenhæng. Den viser også, at når Y-værdierne (for kvalitetsparameteren) er justeret for de tre første score-vektorer, varierer de mellem –0.03 og +0.03. Det betyder dels, at vi kan forudsige kvalitetsparameteren med hjælp af tre score-vektorer med en usikkerhed af denne størrelsesorden, og dels at vi er nået ned på et niveau svarende til usikkerheden i målinger af kvalitetsparameteren.

Bedømmelse af resultater

Det er vigtigt at vurdere hvor mange score-vektorer, der er nødvendige. Det kan undersøges på forskellig vis. En populær metode er *krydsvalidering,* som kan udføres på flere forskellige måder. Man kan f.eks. udelade 10% af prøverne og benytte dem som en slags testdata. Den beregnede løsning uden de 10% bruges til at beregne kvalitetsparameteren for de 10% af prøverne. Gentages dette, kan det for NIR-dataene konstateres, at kun de tre første score-vektorer bør benyttes. Anvendes flere, forringes forudsigelsesevnen.

Usikkerheder på beregningsresultater

Ved disse beregninger har det en vis interesse at kende de statistiske usikkerheder, der er på en række størrelser. Løsningsvektoren for NIR-dataene har 100 værdier, hvor det kan være interessant at kende usikkerheden på de enkelte værdier. Det ordnes ved Jacknifing eller Bootstrapping, hvor beregningerne gentages mange gange. Derved opnås pålidelige konfidensintervaller for de størrelse, der udregnes ved analysen.

Sammenligning mellem metoder

Under standardforudsætninger har løsningsvektoren en kovariansmatrix, der er lig med $\mathrm{Var}(\mathbf{B})=\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Når metoder sammenlignes, benyttes ofte dette udtryk fra de forskellige metoder. Sammenligning kan baseres på $\sigma^2(\mathbf{X}\mathbf{T}^T\mathbf{X})^{-1}$, hvor $\mathbf{X}$ er den del af $\mathbf{X}$-data, der er brugt til at bestemme løsningsvektoren. Mht. industrielle data, viser det sig, at den $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, der fås ved de her nævnte metoder, bliver væsentlig mindre end ved brug af andre metoder. Årsagen er, at man ved at optimere på hvert trin sikrer

maksimal størrelse og minimalt antal af de score-vektorer, der skal anvendes.

Fortolkning af matematiske modeller

De teoretiske videnskaber stiller ofte forventninger om fortolkning, som vi kender fra f.eks. tryk, temperatur og volumen i termodynamisk ligevægt. En tilsvarende præcis fortolkning af løsningsværdier er ofte vanskelig at opnå for industrielle data. Når der anvendes kemometriske metoder giver man afkald på detaljeret fortolkning af løsningsvektoren. Man ser kun på størrelsesforholdene for værdierne i løsningsvektoren. Ved analyse af industrielle data er det ikke en væsentlig mangel, da kendskabet til de nærmere fysisk/tekniske forhold er så dårligt. De kemometrisk metoder er særlig velegnede til analyse af data fra det moderne avancerede måleudstyr, der automatisk giver meget store datamængder. Hvis antal søjler i **X** er f.eks. 1050, vil løsningsvektoren bestå af 1050 værdier. En præcis fortolkning af disse værdier er normalt umulig.

Forretningsmæssig succes

Disse metoder har givet en række virksomheder forretningsmæssig succes. Som eksempel kan nævnes McMaster Advanced Control Consortium ved McMaster universitetet i Toronto, Canada [5]. Her har 20 virksomheder etableret et center med henblik på at bruge teknikkerne i den daglige drift. Implementeringen af teknikkerne har givet dem væsentlige konkurrencemæssige fordele. Endvidere har det haft positiv effekt på personalet (operatørerne), at det er blevet trænet i at aflæse score-plot og andre grafer med henblik på at vurdere om de driftsmæssige forhold er normale. I Danmark kan nævnes Foss Electric, der har implementeret de kemometriske betragtninger i deres måleinstrumenter, der giver et salg på over 1000 mio. kr. på årsbasis. I efteråret 2000 fik Foss Electric en guldmedalje i Frankrig for deres måleapparater til måling af kvaliteten af vin.

Kemometrien i Danmark

Kemometrien betragtes som særlig stærk i Danmark. Det videnskabelige institut, der er størst i verden målt ved antal medarbejdere, der arbejder med kemometrien til daglig, er Levnedsmiddelinstituttet på Landbohøjskolen. Se [6] for en introduktion til de emner, de arbejder med. Fra 1. juli i år er Rasmus Bro udnævnt til forskningsprofessor i kemometri ved Levnedsmiddelinstituttet. Ligeledes er Kim Esbensen blevet udnævnt til professor i kemometri ved Ålborg Universitet, Esbjerg. Dansk Kemometrisk Selskab holder regelmæssigt møder om kemometriske emner, som ofte tiltrækker mange deltagere (bl.a. et PMP møde i IDA i foråret). En række institutioner/virksomheder er begyndt at markere sig centralt ved anvendelse af kemometrien, her kan nævnes Bioteknologisk Institut i Kolding [7]. Den syvende nordiske kemometrikonference afholdes i august i København. Ved de nordiske kemometrikonferencer er der uddelt tre »Herman Wold guldmedaljer«. Guldmedaljen gives for banebrydende udvikling inden for kemometrien. To af guldmedaljemodtagerne, Harald Martens og undertegnede, bor i København.

Kemometri og teoretisk statistik

Kemometri og klassiske statistiske metoder adskiller sig grundlæggende fra hinanden. Kemometrien er meget databaseret. Udviklingen drives i væsentlig grad af anvendelserne og de nye måleinstrumenter. De klassiske statistiske metoder er implementeret i avancerede statistiske programpakker. Men der er problemer med at bruge de statistiske programpakker på industrielle data. Det skyldes bl.a., at signifikanstestning med (de ortogonale) score-vektorer normalt baseres på størrelser, som er ækvivalente med korrelationskoefficienten. Hvis det anvendes på score-vektorer i tabel 1, kan det vises, at alle korrelationskoefficienter mellem de 10 score-vektorer og (den reducerede) Y-variabel er signifikante. En statistisk analyse vil således vise, at alle 10 score-vektorer er signifikante, men vurderes de ud fra forudsigelser, skal kun de tre første anvendes. Det skyldes at signifikanstestning i programpakkerne ikke tager hensyn til størrelsen af de anvendte score-vektorer. Der er også problemer med programpakkerne mht. udpegning af signifikante variable. Det er nærliggende at udvælge variable trinvis. Udvælges variable for NIR-dataene, vil den første variabel uddrage 1% (der er 100 variable og de er autoskalerede) af variationen i **X**, mens den første score-vektor i tabel 1 uddrager 96.12%. Endvidere er score-vektorerne mere robuste end de enkelte variable og derfor bedre at anvende i industrielle omgivelser.

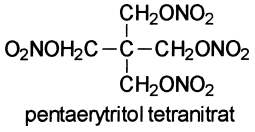
Referencer

1. NIR forum til fremme af NIR indenfor Fødevarer og Landbrug. Nils Bo Büchmann, DLG.
2. Bioteknologisk Institut står for det 3. brugerseminar om Egenkontrol med NIR, 9.-10. maj 2001
3. Höskuldsson, A.: PLS regression methods. *J. of Chemometrics*, 2 (1988) 211-228
4. Höskuldsson, A.: *Prediction Methods in Science and Technology. Vol 1. Basic Teory*. 1996. Thor Publishing. København. ISBN 87-985941-0-9.
5. P. Nomikos & J.F. MacGregor: Multivariate SPC charts for Monitoring Batch Processes. *Technometrics*, 37, 41-59, 1995.
6. Munck L., L. Nørgaard, S. B. Engelsen, R. Bro, C. A. Andersson: Chemometrics in food science – a demonstration of a highly exploratory, inductive evaluation strategy of fundamental scientific importance. *Chemometrics and Intelligent Laboratory Systems*, 44 (1998) 31-60.
7. Pedersen, J.G.: Avanceret dataanalyse viser vej til bedre processtyring. *Dansk Kemi*, nr 3, 82 (2001), 15-17

Nyt om tobak

Noget godt kan man sige om tobak. Det har længe været kendt, at nogle planter og bakterier kan nedbryde affaldsprodukter fra sprængstoffremstillingen. Sprængstoffer er ofte nitroforbindelser eller nitrater, og visse planter kan anvende nitrogenet heri som gødning. De nedbryder desværre kun stofferne langsomt. Forskere i Cambridge har imidlertid nu genetisk modificeret tobaksplanter med et gen fra en bakterie *Enterobacter cloacae*, som producerer store mængder af et enzym pentaerytritoltertranitratreductase, som er kendt for at kunne nedbryde sprængstoffer.

Pentaerytritoltetranitrat (PETN) - se figur - er et meget anvendt sprængstof.



pentaerytritol tetranitrat

Affald fra sprængstoffremstillingen kan anvendes til at gøde tobaksmarker.

Carl Th.

2000, Tobaksplanta oskadeliggör TNT, *Kemivärlden* **3**: 17