Kalibrering i analytisk kemi – Principal Component Regression

Der er en lang og solid tradition for at lave kalibreringsmodeller i analytisk kemi ved hjælp af univariat lineær regression. Denne klumme forklarer hvilke problemer, der kan være ved at anvende sådanne kalibreringsmodeller – og præsenterer et multivariat alternativ kaldet PCR

Af Rasmus Bro, Lars Nørgaard & Søren Balling Engelsen, Institut for Fodevarevidenskab, Det Biovidenskabelige Fakultet, Københavns Universitet


Ligningen for univariat kalibrering er

\[ y_i = b_0 + b_1 \times x_{328 \text{ nm},i} + f_i \]

hvor \( b_0 \) er skærings (offset) og \( b_1 \) er hældning (slope). Indeks angiver den \( i \)te prøve og \( f \) er residualuet. Ud fra de ti kendte prøver kan \( b_0 \) og \( b_1 \)估计es ved hjælp ad mindste kvadratrets metode. Dette kan gøres på næsten enhver lommeregner.

Normalt skrives ligningen

\[ y_i = b_0 + b_1 \times x_{328 \text{ nm},i} + f_i \]

Når \( b_0 \) og \( b_1 \) er bestemt, kan de direkte anvendes til estimering, ofte kaldet prædiktion i kemometri, af koncentrationen i nye prøver. \( b_0 \) og \( b_1 \) kaldes også for regressionskoeficienter.

Fordele og ulemper ved univariat kalibrering
De statistiske forudsætninger for mindste kvadratrets metode er særdeles velbeskrevne for univariat kalibrering, hvilket uddybes til at beregne f.eks. konfidensintervall for estimerer og prædiktionsintervaller for koncentrationen i nye prøver. Dette er en absolut fordel!

En væsentlig ulempe er, at man skal være helt sikker på at nye prøver, som måske er af en mere kompleks beskaffenhed, skal kunne oprettes, så det målte absolveringsignal er selektivt (baselineiseret). Det vil sige at ingen andre kemiske stoffer i prøven må blive til med melde absorbans. Et andet problem ved univariat kalibrering kan være matrix-efekter, som kan have en direkte effekt på absorbansen. Dette kan være af stor betydning ved nogle typer af industrielle målinger, hvor fx. ionstyrke, pH og ikke-signalgivende kemiske stoffer kan variere betydeligt.

Af mere fundamentalt betydning er det imidlertid at uforudsete interferencer ikke kan kompenseres for, når man kun måler absorbansen ved en bølgelængde. Dette er illustreret i figur 1, hvor absorbansen ved 328 nm for prøver med høj absorbans i det høje bølgelængedomære er fuldt sammenligneligt med de øvrige prøvers absorbans ved 328 nm samt med at hele det spækkende mønster afviger. Etablerer man en univariat kalibreringsmodel baseret på bølgelængden ved 328 nm, vil man begå en fejl, da der er en uforudset interferens i prøven som bidrager til absorbansen målt ved 328 nm (figur 2).

Multivariat kalibrering
Univariat kalibrering kan udvides til at anvende mere end blot en bølgelængde. Den direkte udvidelse af den univariat model til en oligovariat model kan skrives

\[ y_i = b_0 + b_1 \times x_{328 \text{ nm},i} + b_2 \times x_{310 \text{ nm},i} + b_3 \times x_{350 \text{ nm},i} + b_4 \times x_{400 \text{ nm},i} + f_i \]

Få info@pumpegruppen.dk

www.pumpegruppen.dk

OBL Proces doseringpumper
Kvalitetsspumper til konkurrencedygtige priser
- Membran- og stempelpumper
- Ev. eller flernovedet løsninger
- Manuels-, elektrisk- eller pneumatisk styring
- API 675, ATEX, FDA
- Op til 6500 l/h, 3-150 bar
- 316L, PP, PVC, PVDF, PTFE

PUMPE
GRUPPEN A/S

Tlf. +45 45 93 71 00
Faks +45 45 93 47 55
I ovenstående ligning anvender vi absorbanse på 5 udvalgte bølgelængder fra 250 til 450 nm. Skrepet på matrix-form skal man finde den regressionsvektor \( \mathbf{b} \), der minimerer residalet \( \mathbf{f} \) i mindste kvadraters forstand.

\[
y = \mathbf{Xb} + \mathbf{f}
\]

hvor \( y \) (antal prøver \( x 1 \)) er koncentrationen af standardprøverne, \( \mathbf{X} \) (antal prøver \( x 5 \)) indeholder spektrene, \( \mathbf{b} \) (5 \( \times \) 1) er regressions-koefficienterne, og \( \mathbf{f} \) (antal prøver \( x 1 \)) er residualet (fejlen i koncentrationen) som ønskes minimeret ved kalibreringsmodellen. I ligningen indgår \( b_i \) ikke, da man modellerer på centred \( \mathbf{X} \) og \( y \) data (se tidligere klummer om PCA i Dansk Kemi, nr. 2, 2008). Hvis man ikke centrerer, skal \( b_i \) inkluderes i ligningen.

Løsningen til ligningen er givet ved

\[
\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y
\]

hvor \( ^{-1} \) betyder den inverse matrix. Så længe de spektrale

---

**Pipettecenteret**
Kalibrering og service af alle fabrikater pipetter.
Vi kalibrerer både ved indsendelse eller på kundens adresse.
Salg af pipetter og laboratorie varer.

**Svingmølle MM 400**
Allround mølle til de mindre prøver, bl.a. DNA/RNA og XRF analyser.
Tilberedet op til 20 prøver samtidigt.

**Laboratories Scandinavia**
Komplet tilbehør til laboratoriet including: Glucose, Fructose, Sucrose, Ethanol, L-Malic and D-Lactic acid.

**Thermo Scientific**

---

**Det Kemometriske Rum**

---

**FIGUR 1.** Absorptionsspektre for fem prøver i det spektrale område 250-450 nm målt med 2 nm’s interval; dvs. i alt 101 spektrale variable er registreret. Maksimumsintensiteten er ved 328 nm, som anvendes ved univariat kalibrering. En prøve har en afvigende form, som kun kan afsløres ved at måle ved flere bølgelængder.

**FIGUR 2.** Den afvigende prøve er sammensat af bidrag fra to kemiske komponenter. Den interfererende komponent (blå) bidrager til absorbanse målt ved 328 nm, og dermed opnås et fejlagtigt estimat i den univariante kalibrering.

X-variable ikke er stærkt korrelede, og antallet af variable er mindre end eller lig med antallet af prøver, giver ovennævnte løsning mening. Metoden kaldes Multiple Linear Regression (MLR).

Hvis man nu ønsker at inkludere alle målte bølgelængder ud fra den forudsætning, at man ikke a priori ønsker at eliminere variable for sin regressionsmodel, så fås følgende ligning

\[
y_i = b_0 + b_1 x_{250\text{nm},i} + b_2 x_{252\text{nm},i} + b_3 x_{254\text{nm},i} + \ldots + b_100 x_{448\text{nm},i} + b_101 x_{450\text{nm},i} + f_i
\]

Her har vi som eksempel målt absorbanse ved 101 bølgelængder fra 250 til 450 nm. Skrepet på matrix-form skal man igen finde den regressionsvektor \( \mathbf{b} \), der minimerer residalet \( \mathbf{f} \) i mindste kvadraters forstand.

\[
y = \mathbf{Xb} + \mathbf{f}
\]
hvor \( y \) (antal prøver \( \times 1 \)) er koncentrationen af standard-prøverne, \( X \) (antal prøver \( \times 101 \)) indeholder spektrene, \( b \) (101 \( \times 1 \)) er regressions-koefficienterne, og \( f \) (antal prøver \( \times 1 \)) er residualet (fejlen i koncentrationen) som ønskes minimeret ved kalibreringsmodellen.

Mindstekvadraters løsning til ligningen indebærer, at man skal finde den inverse til en \( X \) matrix med f.eks. dimensionen 10 prøver \( \times 101 \) spektrale variable. Det vil sige 10 ligninger med 101 ubekendte, og da dette som bekendt ikke kan løses umiddelbart, må man gå alternative veje for at finde en løsning.

**PCA som redning**

I stedet for at arbejde direkte på \( X \) matricen kan man anvende principal komponent analyse (PCA), til at komprimere \( X \) matricen ifølge ligningen

\[
X = T_a P'_a + E_a
\]

hvor \( X \) er de centrerede spektre, og indeks \( a \) angiver antal principale komponenter, der er beregnet i modellen.

Hvis man nu lader \( T_a \) (antal prøver \( \times a \)), vi har tidligere kaldt dem scores, repræsenterer de kvantitative variationer i \( X \), kan man i stedet løse ligningen

\[
y = T'_a b^* + f
\]

hvor \( b^* \) (\( a \times 1 \)) angiver, at vi arbejder med score-matricen. Denne ligning kan løses med mindste kvadrater, da søjle-vektorerne i \( T_a \) er ortogonale og antal søjler er mindre end eller lig med antal prøver (vi har allsidå igen opnået et fordelaigt forhold mellem ligninger og ubekendte). Den matematiske løsning til at finde regressionsvektoren ser således ud

\[
b^* = (T'_a T_a)^{-1} T'_a y
\]

Vi ønsker nu at finde en regressions-vektor som kan ganges direkte på et målt absorbans spektrum, og denne kan estimeres som følger

\[
b = Pb^*
\]

Metoden, der er udledt her, hedder Principal Component Regression (PCR) og er en fundamental regressionsmetode i kemometrien.

**Outro**

Det var måske en rimelig hård omgang, men nu er banen kridlet op til at anvende PCR på virkelige data, hvilket vi vil gøre i næste klumme.

Den opmærksomme læser vil måske have overvejet om ikke de præsenterede data ser konstruerede ud, og vi må gå til bekendelse og indrømme dette. Meget mod kemometriens væsen er der anvendt simulerede data til at illustrere principperne i denne klumme; dette er således undtagelsen, der bekæftet reglen om, at rigtige kemometriker analyserer rigtige data.

**ALSIDENT® udsugningssystemer**

– gør arbejdsmiljøet renere!

www.alsident.com

Finlandsvej 10 · DK-8450 Hammel · Tlf. +45 86 96 50 00