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 Fødevarevidenskab, Det Biovidenskabelige Fakultet, Københavns Universitet

En traditionel analytisk-kemisk kalibreringsmodel kunne fremkomme på følgende vis: ti prøver med varierende koncentration af 2-hydroxy-benzaldehyd analyseres ved en absorbansmåling (328 nm), og absorbansen afbildes mod koncentrationen. Ved univariat lineær regression opnås en kalibreringsmodel, der kan benyttes på nye prøver. Ved at måle absorbansen af en ny prøve, kan koncentrationen aflæses eller beregnes ud fra absorbansen.

Ligningen for univariat kalibrering er

$$koncentration_i = b_0 + b_1 \times absorbans_{328\,nm,\,i} + f_i$$

hvor b_0 er skæring (offset) og b_1 er hældning (slope). Indeks angiver den *i*'te prøve og *f* er residualet. Ud fra de ti kendte prøver kan b_0 og b_1 estimeres ved hjælp ad mindste kvadraters metode. Dette kan gøres på næsten enhver lommeregner. Normalt skrives ligningen

$$y_i = b_0 + b_1 \times x_{328\,nm,\,i} + f_i$$

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 regressionskoefficienter.

Fordele og ulemper ved univariat kalibrering

De statistiske forudsætninger for mindste kvadraters metode er særdeles velbeskrevne for univariat kalibrering, hvilket udnyttes til at beregne f.eks. konfidensintervaller for estimater 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 oprenses, så det målte absorbanssignal er selektivt (baselinieseparet). Det vil sige at ingen andre kemiske stoffer i prøven må bidrage til den målte absorbans. Et andet problem ved univariat kalibrering kan være matrix-effekter, som kan have en indirekte 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 fundamental betydning er det imidlertid at uforudsete interferenser ikke kan kompenseres for, når man kun måler absorbansen ved én bølgelængde. Dette er illustreret i figur 1, hvor absorbansen ved 328 nm for prøven med høj absorbans i det høje bølgelængdeområde er fuldt sammenlignelig med de øvrige prøvers absorbans ved 328 nm samtidig med at hele det spektrale 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 én bølgelængde. Den direkte udvidelse af den univariate model til en oligovariat model kan skrives

$$y_i = b_0 + b_1 \times x_{250\,nm,\,i} + b_2 \times x_{300\,nm,\,i} + b_3 \times x_{350\,nm,\,i} + b_4 \times x_{400\,nm,\,i} + b_5 \times x_{450\,nm,\,i} + f_i$$

I ovenstående ligning anvender vi absorbansen ved 5 udvalgte bølgelængder fra 250 til 450 nm. Skrevet på matrix-form skal man finde den regressionsvektor **b**, der minimerer residualet **f** i mindste kvadraters forstand.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{f}$$

hvor **y** (antal prøver × 1) er koncentrationen af standardprøverne, **X** (antal prøver × 5) indeholder spektrene, **b** (5 × 1) er regressions-koefficienterne, og **f** (antal prøver × 1) er residualet (fejlen i koncentrationen) som ønskes minimeret ved kalibreringsmodellen. I ligningen indgår b_0 ikke, da man modellerer på centrede **X** og **y** data (se tidligere klumme om PCA i Dansk Kemi, nr. 2, 2008). Hvis man ikke centrerer, skal b_0 inkluderes i ligningen.

Løsningen til ligningen er givet ved

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

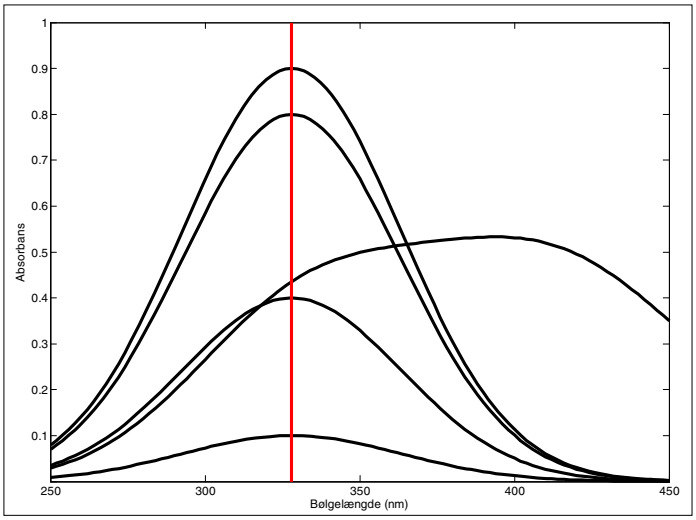
hvor $^{-1}$ betyder den inverse matrix. Så længe de spektrale X-variable ikke er stærkt korrelerede, og antallet af variable er mindre end eller lig med antallet af prøver, giver ovennævnte løsning mening. Metoden kaldes Multiple Linear Regression (MLR).

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

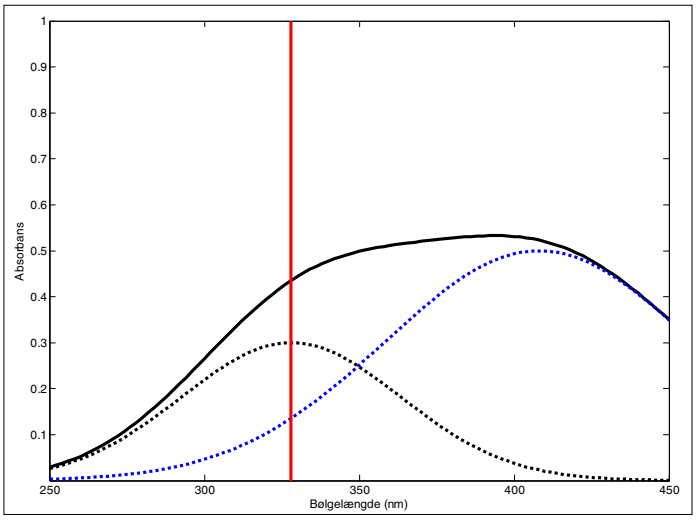
$$y_i = b_0 + b_1 \times x_{250\,nm,\,i} + b_2 \times x_{252\,nm,\,i} + b_3 \times x_{254\,nm,\,i} + \ldots + b_{100} \times x_{448\,nm,\,i} + b_{101} \times x_{450\,nm,\,i} + f_i$$

Her har vi som eksempel målt absorbansen ved 101 bølgelængder fra 250 til 450 nm. Skrevet på matrix-form skal man igen finde den regressionsvektor **b**, der minimerer residualet **f** i mindste kvadraters forstand.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{f}$$



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. Én 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 absorbansen målt ved 328 nm, og dermed opnås et fejlagtigt estimat i den univariate kalibrering.

hvor **y** (antal prøver × 1) er koncentrationen af standard-prøverne, **X** (antal prøver × 101) indeholder spektrene, **b** (101 × 1) er regressions-koefficienterne, og **f** (antal prøver × 1) er residualet (fejlen i koncentrationen) som ønskes minimeret ved kalibreringsmodellen.

Mindste kvadraters løsning til ligningen indebærer, at man skal finde den inverse til en **X** matrix med f.eks. dimensionen 10 prøver × 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

$$\mathbf{X} = \mathbf{T}_a\mathbf{P}'_a + \mathbf{E}_a$$

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

Hvis man nu lader $\mathbf{T}_a$ (antal prøver × *a*), vi har tidligere kaldt dem scores, repræsentere de kvantitative variationer i **X**, kan man i stedet løse ligningen

$$\mathbf{y} = \mathbf{T}_a\mathbf{b}^* + \mathbf{f}$$

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

$$\mathbf{b}^* = (\mathbf{T}_a'\mathbf{T}_a)^{-1}\mathbf{T}_a'\mathbf{y}$$

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

$$\mathbf{b} = \mathbf{P}\mathbf{b}^*$$

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 kridtet 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 bekræfter reglen om, at rigtige kemometrikere analyserer rigtige data.

lan@life.ku.dk