Ekstrakt i øl bestemt med Principal Component Regression

I seneste klumme gennemgik vi teorien for Principal Component Regression, og i denne klumme vil vi demonstrere anvendelsen af metoden på et datasæt med det formål at bestemme koncentrationen af ekstrakt i øl ud fra visuel- og nærinfrarød spektroskopi

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

Principal Component Regression (PCR) er en metode til multivariat regression. Formålet er at fortolke og forstå data samt at forudsige koncentrationen af en given kemisk komponent eller funktionel egenskab for det analyserede produkt ud fra f.eks. hurtigt målte spektroskopiske data.

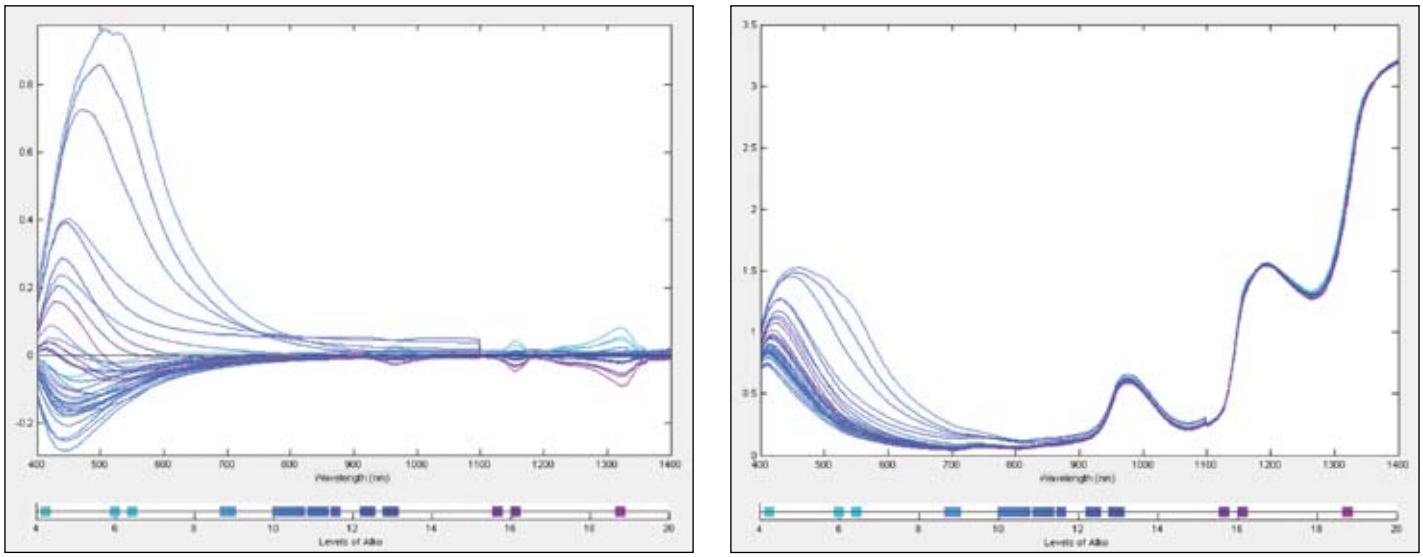
Øl data

40 prøver bestående af forskellige øl er analyseret for ekstraktindhold i % plato med en laboratoriemetode. Ekstrakt er en vigtig kvalitetsparameter i bryggeriindustrien som indikerer gærens potentiale til at danne alkohol. Ekstraktprocenten varierer fra 4,23-18,76 % plato. De samme 40 prøver er ligeledes målt med visuel- og nærinfrarød spektroskopi i området 400 nm til 1400 nm med to nanometers interval, dvs. 501 spektrale variable. Prøverne er afgasset inden måling på et NIRSystems 6500 spektrofotometer i en 30 mm kuvette. Spektrofotometeret anvender et delt detektorsystem med silicium-baseret detektor i området 400 nm til 1100 nm og bly-sulfid detektor (PbS) i området 1100 nm til 2500 nm. Figur 1A viser spektrene for alle prøver farvet efter ekstraktkoncentrationen, og figur 1B viser de tilsvarende centrede spektre.

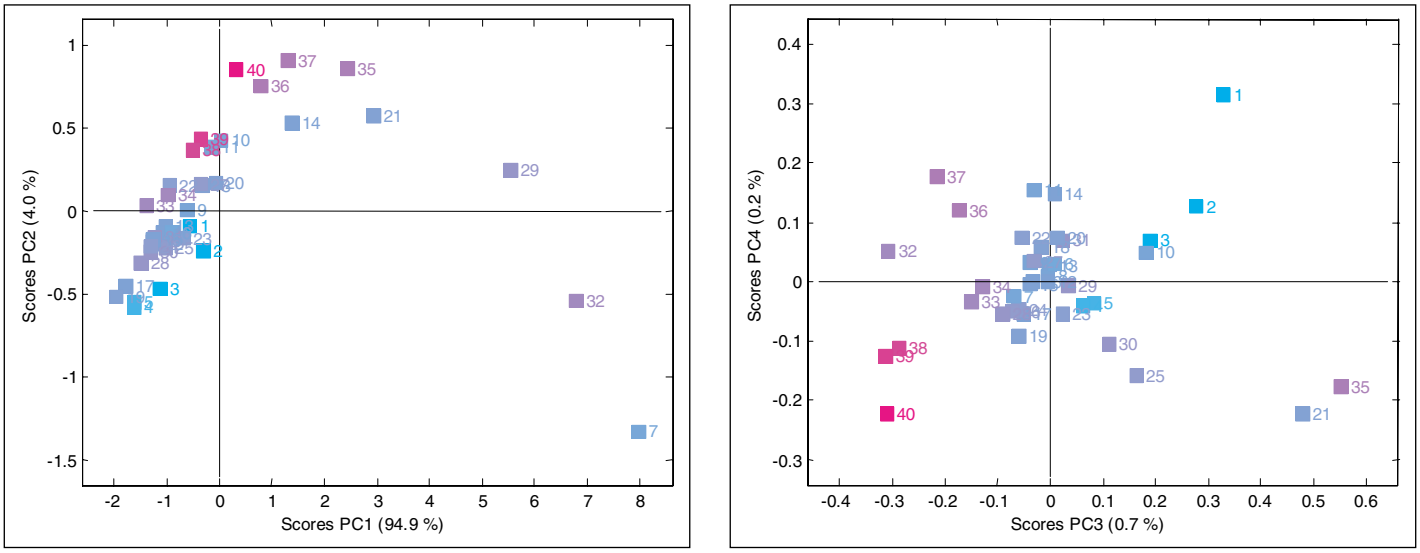
Det fremgår, at variationen i det synlige bølgelængdeområde fra 400 nm til 800 nm er væsentlig mere udtalt end i det nærinfrarøde område (figur 1A). For de centrerede data (figur 1B) er det tydeligt, at ekstrakt-koncentrationen, som forventet, afspejles bedre i det nærinfrarøde spektrale område sammenlignet med det synlige.

PCA på de spektrale data

Først beregnes en PCA ($\mathbf{X}= \mathbf{T}_a\mathbf{P'}_a + \mathbf{E}_a$) på de spektrale data, og prøverne i score-plottet (baseret på $\mathbf{T}$) farves efter den aktuelle ekstrakt-koncentration. Den forklarede varians i de spektrale data for de principale komponenter eet, to, tre og fire er henholdsvis 94,9; 4,0; 0,7 og 0,2 %. Som forventeligt for spektrale data, er den forklarede varians for den første komponent langt større end de øvrige, fordi den forklarer den overordnede form alle spektrene har. Det betyder dog ikke at de følgende komponenter er uvæsentlige. Der er fx. ikke umiddelbart systematisk variation i forhold til ekstrakt-koncentrationen (farvekode) for principal komponent eet mod to i score plottet (figur 2A). Hvis man i stedet inspicerer komponent tre mod fire (figur 2B) ses en ekstrakt gradient fra lave koncentrationer øverst til højre til høje koncentrationer nederst til venstre. Vi vil således forvente, at de første fire komponenter, og primært de sidste af disse, kan relateres til ekstrakt-koncentrationen i en regressionsmodel.



Figur 1. A) Absorptionsspektre for fyrre øl-prøver i det spektrale område 400-1400 nm målt med 2 nm's interval; dvs. i alt 501 spektrale variable er registreret. Spektrene er farvet efter prøvens ekstraktkoncentration. B) Centrerede absorptionsspektre for de samme prøver. Området fra 1100 nm til 1375 nm ses at afspejle ekstraktkoncentrationen bedre end det synlige område.



Figur 2. A) Scoreplot for PC1 mod PC2 for de spektrale data. B) Scoreplot for PC3 mod PC4. Prøverne er farvet efter ekstraktkoncentrationen.

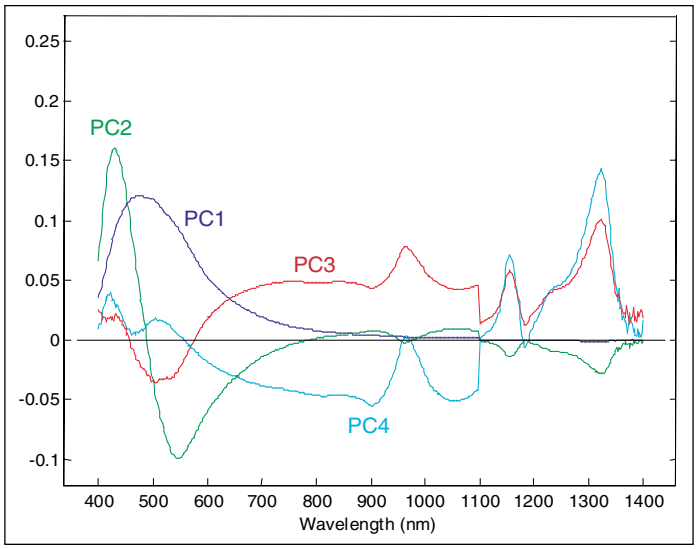
PCR

Vi vil nu løse ligningen $\mathbf{y} = \mathbf{T}_a\mathbf{b}^* + \mathbf{f}$, hvor $\mathbf{b}^*$ angiver, at vi arbejder med den lav-dimensionale score-matrice (måske 4-5 variable) i stedet for den oprindelige **X** matrix (501 variable). Den matematiske løsning til at finde regressionsvektoren ser således ud: $\mathbf{b}^* = (\mathbf{T}_a'\mathbf{T}_a)^{-1}\mathbf{T}_a'\mathbf{y}$, og vi vælger at se på de første fem komponenter. Det er muligt at beregne, hvor meget af variationen i **y**, hver enkelt komponent forklarer. For de første fem komponenter er forklaringsgraden af henholdsvis **X** og **y** således:

PC nummer	Forklaret %-varians i X	Forklaret %-varians i y
1	94,9	1,8
2	4,0	24,8
3	0,7	36,3
4	0,2	31,6
5	0,1	0,2

*Tabel 1. Forklaret %-varians for **X** og **y** i en PCR model af øl-data.*

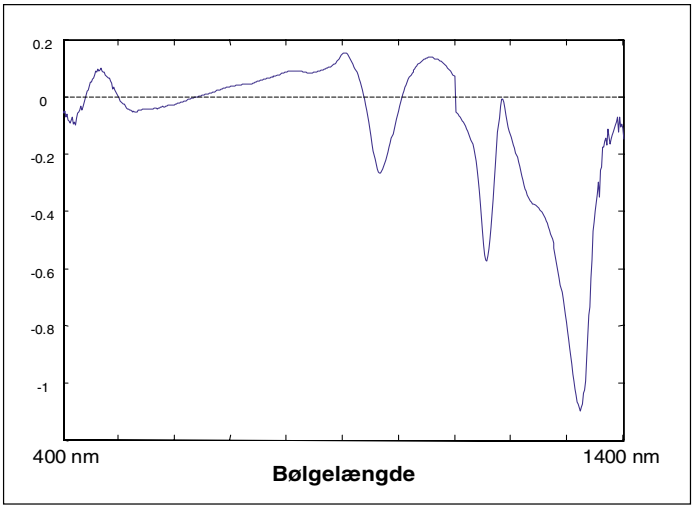
Bemærk at vi primært er interesserede i forklaringen af **y** (ekstrakt). Det ses, at den første komponent ikke er specielt relevant for forklaringen af **y**, mens komponent to til fire bidrager væsentligt. Dermed er disse komponenter de vigtigste i modellen. Komponent fem og op forklarer samlet under 6 % af den samlede variation i **y**. Loadingvektorerne (**P**) for komponent eet til fire er vist i figur 3, hvor det fremgår, at loadings for kom-



Figur 3. Loadings for de første fire PC'er. PC1 ses at have lave værdier i området fra 900 nm til 1400 nm.

ponent to til fire har høje værdier i området 1100 til 1400 nm i forhold til komponent eet. Helt overordnet er loading-værdierne i det synlige område høje sammenlignet med det nærinfrarøde område, hvilket stemmer overens med variationen i de centrerede data (figur 1B).

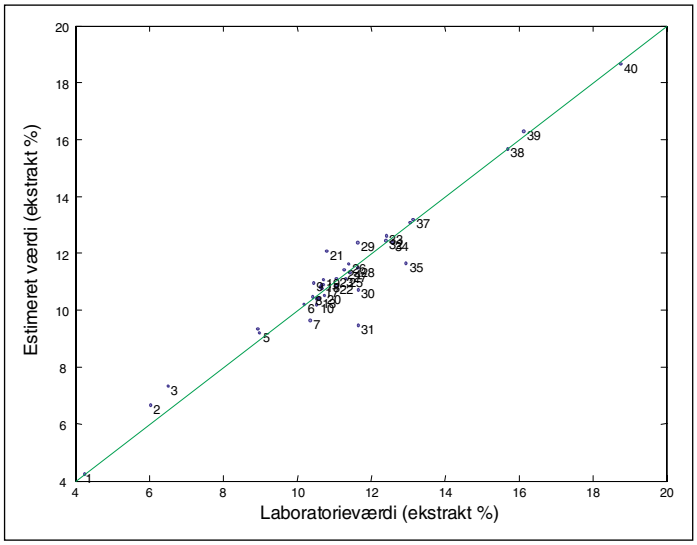
Regressions-vektoren, som skal bruges til at estimere ekstrakt-procenten i en ny prøve ved at gange den direkte på det centrerede målte absorbans spektrum, estimeres fra ligningen: $\mathbf{b} = \mathbf{Pb}^*$. Regressions-vektoren for en model baseret på fire principale komponenter er vist i figur 4.



Figur 4. Regressionskoefficenter for en PCR model baseret på fire principale komponenter. Området fra 900 nm og opefter dominerer med høje værdier.

I området fra 900 nm til 1400 nm ses tre områder med høje absolutte værdier af regressionskoefficienterne. Et område omkring 960 nm som skyldes en relation til vands (O-H strækning) anden overtone, et område ved 1150 nm som skyldes anden overtone af C-H strækninger samt et område omkring 1320 nm som ligger på yderflanken af vands første overtone (O-H strækning). Mens det er noget nær umuligt at tilordne de holografiske nærinfrarøde spektre, så er det dog indlysende at alle tre områder direkte eller indirekte kan relateres til ekstrakt koncentrationen i prøverne, hvilket underbygger validiteten af PCR modellen. Ligningen, der anvendes til estimering af ekstrakt-procenten i en ny prøve, er:

$$y = b_0 + b_1 \times x_{400\,nm} + b_2 \times x_{402\,nm} + b_3 \times x_{404\,nm} + \ldots + b_{500} \times x_{1398\,nm} + b_{501} \times x_{1400\,nm}$$



Figur 5. Estimerede ekstrakt % værdier baseret på nærinfrarøde spektroskopi i en fire komponent PCR model sammenlignet med de målte laboratorie-værdier.

hvor x-værdierne er centrerede, og b-værdierne kan aflæses i figur 4; f.eks. er $b_1 = -0{,}051$. Estimeres koncentrationen for de fyrre prøver ud fra denne ligning kan de estimerede koncentrationer sammenlignes med de målte koncentrationer. I figur 5 er illustreret hvorledes det ser ud for de aktuelle data. Den kvadrerede korrelationskoefficient er 0,95. Hvis den udviklede model giver acceptable prædiktioner, kan man forestille sig en on-line metode baseret på PCR og nærinfrarød spektroskopi. Det vil senere vise sig, at modellen kan forbedres dramatisk ved hjælp af variabel selektion.

Outro

PCR er en meget anvendt og anvendelig(!) metode til multivariat regressionsanalyse. I nærværende eksempel så vi, at de beregnede komponenter ikke nødvendigvis er de mest relevante for y-variablen, hvilket kan være en ulempe i PCR. En alternativ multivariat regressionsmetode, som delvist løser denne problemstilling, er Partial Least Squares (PLS) Regression, som vi vil se nærmere på i en efterfølgende klumme.

I ovenstående eksempel har vi ikke gjort et stort nummer ud af at *validere* modellen; dvs. estimere antal signifikante komponenter i modellen samt evaluere dens prædiktionsusikkerhed for kommende prøver. Der findes specielle metoder til dette, som vi også efterfølgende vil komme ind på.