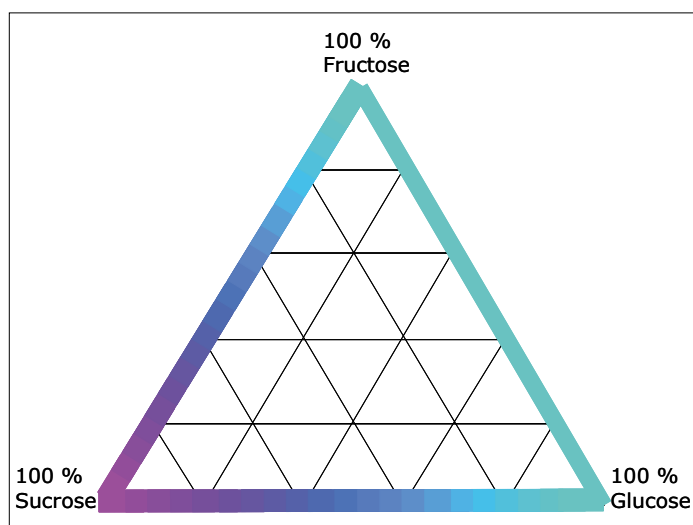


## Principal Component Analysis af nærinfrarøde spektroskopiske data

Spektroskopiske data er generelt kendetegnet ved at være stærkt ko-lineære; dvs. to nabobølglængder er positivt korrelerede med høje korrelationskoefficienter. PCA er skræddersyet til at håndtere den slags data, og det er i analysen af spektroskopiske data, PCA virkelig viser sit værd

Af Lars Norgaard, Søren Balling Engelsen og Rasmus Bro, Københavns Universitet

Anvendelsen af PCA på spektroskopiske data illustreres bedst med et eksempel. Til dette formål har vi målt et tre-komponent blandingsdesign hvor sukrose og dets to monomer komponenter: glukose og fruktose er blandet sammen med hver komponent i 21 niveauer [0%; 100%] (se figur 1). Et sådant fuldt 3-komponent blandings design fører til i alt  $21+20+19+ \dots + 1 = 231$  blandinger, der alle blev målt med nærinfrarød spektroskopi (NIR).



Figur 1. Blandingsdesign (produceret af Hanne Winning, Københavns Universitet).

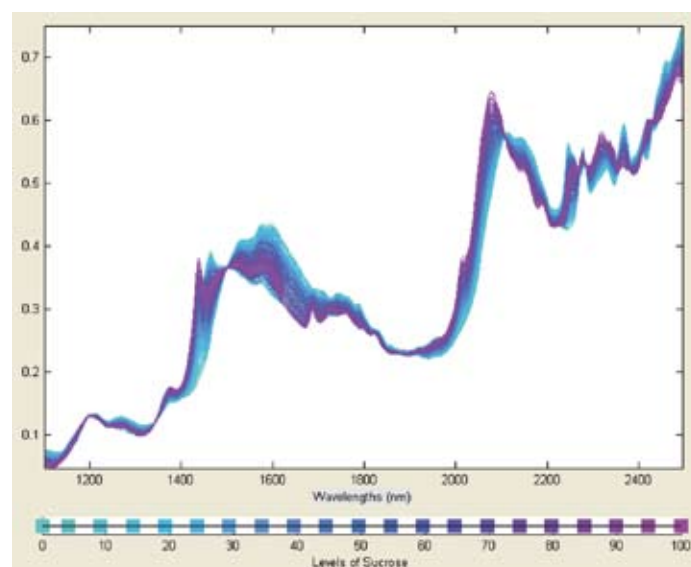
Nærinfrarød spektroskopi har været hoveddrivkraften ved udviklingen af den tidlige kemometri i 80'erne. Nærinfrarød spektroskopi måler overtoner og kombinationstoner af de fundamentale molekulære vibrationer, som ligger i det infrarøde område. Det er specielt de asymmetriske vibrationer, som er intensive i det nært infrarøde område dvs. strækningsvibrationer, der involverer hydrogen (f.eks. C-H, O-H og N-H). Det gør NIR spektroskopi særdeles anvendeligt til at analysere biologiske systemer.

Det faktum, at man i NIR måler den samme grundlæggende molekulære vibration som et antal forskellige over- og kombinations-toner over praktisk taget hele det nærinfrarøde område, giver stærkt overlappende og nærmest holografiske NIR spektre, der er yderst vanskelige

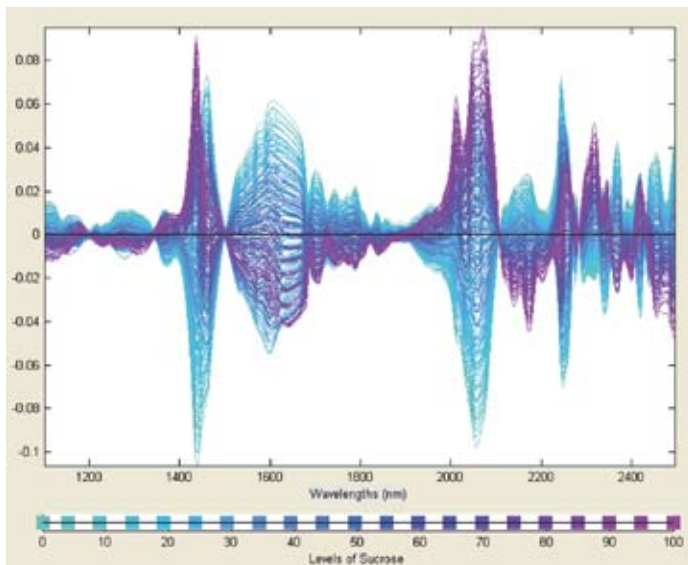
at fortolke på traditionel vis. Mens dette er årsagen til, at NIR i kombination med kemometri er formidabelt informationsrigt, er det paradoksalt nok også årsagen til, at NIR spektroskopien først meget sent blev »stuerent« ved universiteterne (her på KU-LIFE, tidligere KVL, fik vi som et af de første universiteter i Danmark vores første NIR spektrometer i 1992).

Figur 2 viser de 231 NIR reflektansspektre, der er optaget med vores oprindelige dispersive FOSS NIRSystems spektrometer i reflektansmode, dvs. målt mod en hvid keramisk baggrund i området 1100-2500 nm. Enheden på y-aksen er  $\log(1/R)$  hvor R er forholdet mellem intensiteten reflekteret fra prøven og intensiteten reflekteret fra standarden. X-aksen er bølgelængden i nm. Som det tydeligt fremgår af NIR spektrene, er der ingen baselinieseparerede signaler; kun stærkt overlappende peaks.

Data fra dette eksempel (og andre) kan downloades fra [www.models.life.ku.dk/dansk kemi](http://www.models.life.ku.dk/dansk kemi). Programmet LatentiX ([www.latentix.com](http://www.latentix.com)) er anvendt til at plote rådata samt til beregning af PCA modellen.



Figur 2. NIR spektre fra 231 blandinger af sukrose, fruktose og glukose. Spektrene er farvet efter sukrosekoncentration (cyan er 0% og magenta er 100%).



Figur 3. Centrerede NIR spektre farvet efter sukrosekoncentration.

## Centrering af data

Første trin inden PCA modellering er at centrere de spektroskopiske data. Dette gøres for at fokusere på variationerne mellem de enkelte prøver i stedet for det generelle signal niveau. Centrering består simpelthen i at fratrække gennemsnitsreflektansen ved hver bølglængde, således at reflektansen ved hver bølglængde/tal summerer til nul.

## PCA på NIR data

Præcis som for eksemplet med McDonalds data (Dansk Kemi, januar 2008) opstilles en PCA model

$$\mathbf{X} = \mathbf{TP}$$

For at være lidt mere præcis kan modellen skrives

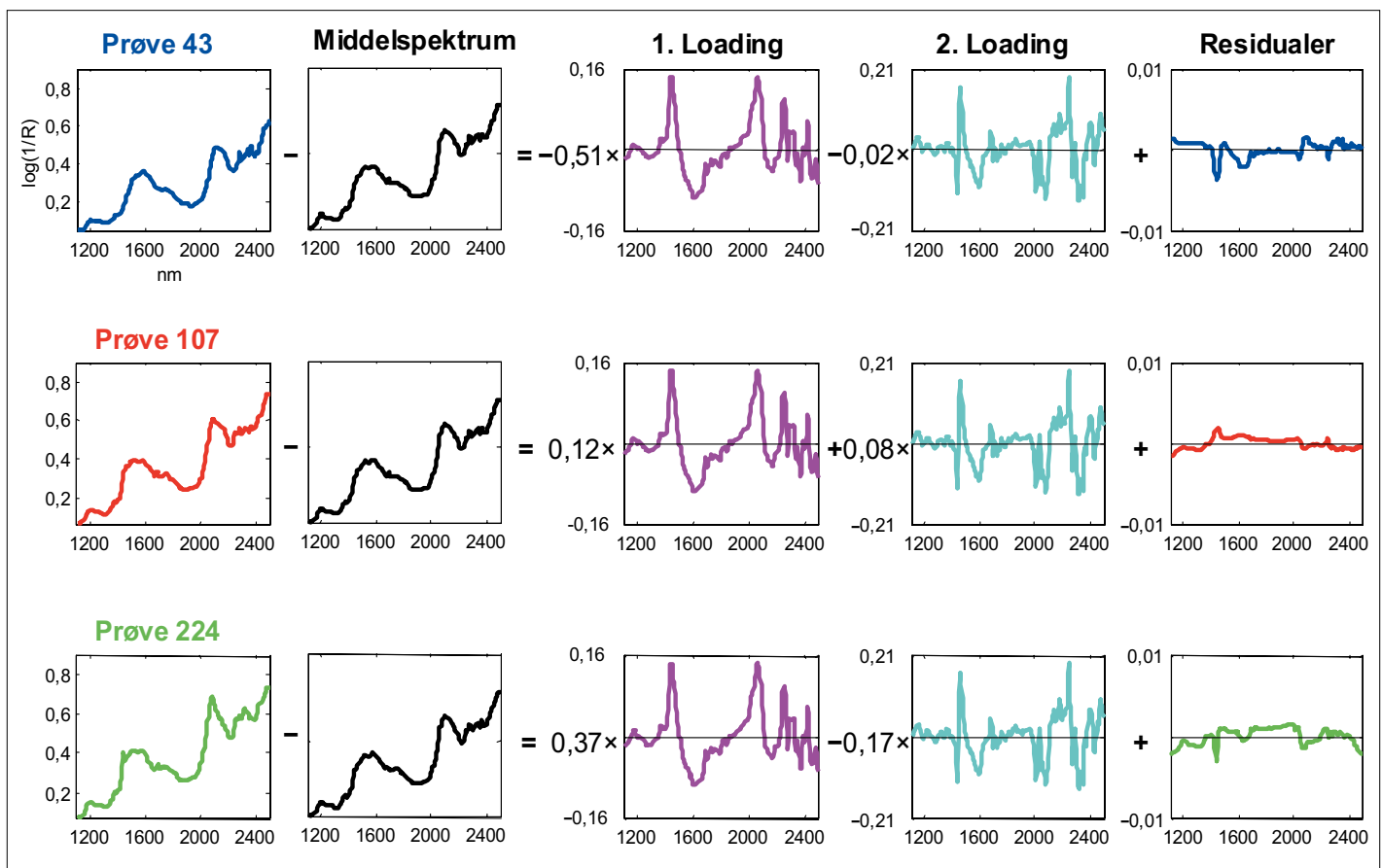
$$\mathbf{X}_c = \mathbf{T}_a \mathbf{P}'_a + \mathbf{E}_a$$

Her betyder  $\mathbf{X}_c$  de centrerede spektrale data med samme dimensioner, som den oprindelige  $\mathbf{X}$  matrix, altså en tabel med 231 objekter og 350 variable. Indeks  $a$  angiver antal *principale komponenter*, der er beregnet i modellen. I dette eksempel vil vi nøjes med at inspicere de første to principale komponenter, hvilket giver god mening i forhold til antal kemiske variationskilder i prøverne: tre kemiske komponenter i et blandingsdesign (sum er 100%) giver ideelt anledning til to uafhængige variationskilder. Vi gemmer til en senere klumme, hvorledes det optimale antal komponenter i en PCA model kan bestemmes matematisk.

$\mathbf{T}_a$  og  $\mathbf{P}'_a$  indeholder henholdsvis scores og loadings for  $a$ -komponent modellen, og  $\mathbf{E}_a$  indeholder *residualerne*; dvs. den del af data, der ikke er beskrevet af modellen.

## PCA - en lineær & additiv model

I figur 4 er princippet i PCA illustreret for tre udvalgte prøver; bemærk at PCA modellen er beregnet på alle 231 prøver. Til venstre i figuren ses de rå spektre for prøve 43 (blå), prøve 107



Figur 4. Illustration af PCA på NIR data. Se tekst for detaljeret beskrivelse.

(rød) og prøve 224 (grøn), der kommer direkte fra spektrometret. Søjle to viser gennemsnitspektret, der subtraheret hvert enkelt prøvespektrum svarende til centreringen af data. Gennemsnitspektret er det samme for alle prøver og derfor vist i samme farve (sort).

Den første loading vektor (magenta) er den spektrale struktur, der bedst beskriver variationen i de centrerede data (figur 3). **Ingen** anden underliggende struktur kan forklare mere af variationen i data end denne. Første loading er fælles for alle prøverne; det, der gør prøverne forskellige fra hinanden, er indholdet (eller 'koncentrationen') af denne struktur i deres spektrum: dette kaldes prøvens score-værdi. Prøve 43 har score-værdien  $-0,51$  for 1. loading og de 230 andre prøver i datasættet har andre scores. Tager man loading vektoren gange  $-0,51$  så er det den bedst mulige beskrivelse man kan få af prøve 43, når loading vektoren også skal beskrive de øvrige prøver.

Anden loading (cyan) er den struktur, der beskriver næstmest af variationen i datasættet, og vektoren har desuden den egen-skab, at den er orthogonal (vinkelret) på den første loading. Igen fremgår prøvernes forskellighed af score-værdien, som er  $-0,02$  for prøve 43.

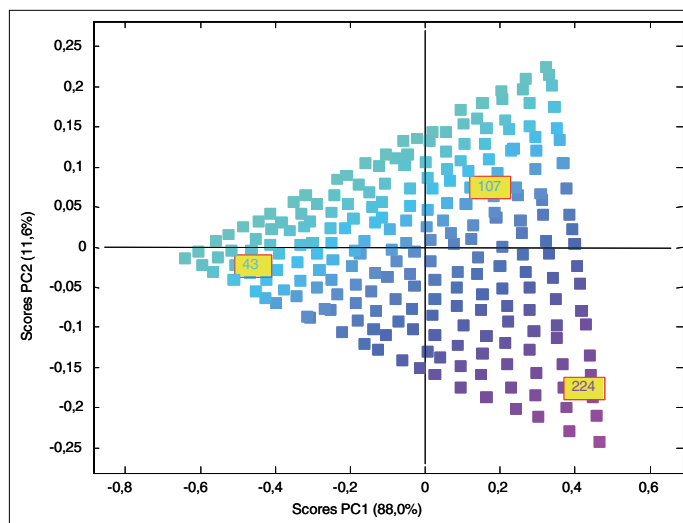
## Residualer og varians forklaret

Den del af variationen i datasættet, der ikke er beskrevet af de to første loading vektorer fremgår af residualerne yderst til højre i figur 4. Residualerne er specifikke for hver prøve og kan bl.a. anvendes til detektion af afvigende mønstre i enkelt-prøver. Bemærk y-aksen for residualerne: den numeriske værdi svinger indenfor  $\pm 0,002$ . Disse værdier kan sammenlignes direkte med variationen i de centrerede spektrale data (figur 3), som varierer mellem  $-0,1$  og  $+0,09$ .

Ved at sammenligne residualernes størrelse med de centrerede datas variation kan man beregne *variens forklaret* for hver enkelt *principal komponent*. I dette tilfælde forklarer første komponent 88,0% af den total variation, anden komponent 11,6% af variationen, og samlet forklarer de to komponenter 99,6% af variationen i data.

## Scores plot

Ved at plote alle 231 score-værdier for den første principal komponent mod de tilsvarende værdier for anden komponent fås et score plot (figur 5). Det bemærkes at hvert punkt i dette scoreplot repræsenterer et NIR spektrum med oprindeligt 350 variable. I det givne tilfælde kan man se at prøve 43 placeres i koordinatsystemet med koordinaterne  $(-0,51; -0,02)$ , prøve 107 ved  $(0,12; 0,08)$  og så fremdeles for de resterende 229 prøver.



Figur 5. Score plot fra en PCA model på NIR data. Blandingsdesignet ses tydeligt. Objekterne er farvet efter sukrosekoncentrationen (cyan er 0% og magenta er 100%).

## Model med centrering

Som det fremgår af figur 4 kan man flytte middelspektret om på denne anden side af lighedstegnet og derved opnå følgende beskrivelse af PCA modellen *inklusive* centrerings-trinet

$$\mathbf{X} = \mathbf{1}\mathbf{x}'_{\text{snit}} + \mathbf{T}_a\mathbf{P}'_a + \mathbf{E}_a$$

Her er  $\mathbf{X}$  ucentrerede rådata,  $\mathbf{1}$  er en søjlevektor bestående af 1-taller,  $\mathbf{x}'_{\text{snit}}$  er en rækkevektor, som er gennemsnittet over alle objekter (=gennemsnitsspektrum) og  $\mathbf{T}_a$ ,  $\mathbf{P}_a$  og  $\mathbf{E}_a$  er beskrevet ovenfor.

## Outro

PCA er overlegen til at håndtere stærkt ko-lineære data som ofte ses i spektroskopien. Som det fremgår af eksemplet er PCA et godt værktøj til eksplorativ dataanalyse: man kan se enkeltprøvers opførsel og karakteristik samt studere hvilke bølgelængdeområder, der har betydning for ligheden/forskellen mellem prøver.

PCA kan opfattes som en »omvendt« Lambert-Beer model: modellen estimerer latente spektre (loadings) og bestemmer koncentrationen af disse i prøverne (scores) ud fra de målte spektre.

I eksemplet har vi arbejdet med data korrigeret for lysspredning; dette er udført ved hjælp af metoden Multiplicative Scatter/Signal Correction (MSC), som vi vil beskrive i en senere klumme.

## KU-forsker på Wireds top-10

Professor Henrik Clausen fra Det Sundhedsvidenskabelige Fakultet ved Københavns Universitet, indtager en fornem fjerdeplads på magasinet Wireds top-10 over videnskabelige gennembrud i 2007.

Gennem studier af komplekse sukkerstoffer har han opdaget, hvordan vi kan omdanne blodtyperne A, B og AB til den neutrale type O, som alle kan tåle at få ved en transfusion og specielle behandlinger af kræft, leukæmi, og blodmangelsygdomme. Samtidig kan den afhjælpe kriser, som da det amerikanske National Institute of Health i efteråret 2007 løb ind i en kritisk mangel på blodtype O.

Denne alkimistiske blodtype-forvandling er mulig, fordi Henrik Clausen studerer glykobiologi, dvs. de for organismen livsvigtige sukkerstoffer - glykaner. De udgør et livets tredje sprog og spiller en rolle for f.eks. blodtyper, immunforsvarets funktion, og så alsidige sygdomme som turistmave, influenza, malaria og kræft. Kilde: KU