

## Hvorfor multivariat dataanalyse

Med stor jubel har vi taget imod invitationen til at lave en klumme i Dansk Kemi om kemometri. Den kemometriske tilgang til problemanalyse i kemien og mange andre forskningsområder er eskaleret i de seneste 10 år; en søgning alene på ordet 'chemometrics' i litteraturlæsebasen Web of Science viser, at antallet af hits stiger fra 1256 i 1997 til 4900 i 2007. Potentialet inden for mange forskningsområder er stadig stort, og vi vil i de kommende klummer vise hvorledes kemometriske metoder kan bidrage til kompleks problemløsning og problemkarakterisering

Af Lars Norgaard, Søren Balling Engelsen og Rasmus Bro

Der findes mange forskellige definitioner af kemometri; en af de mest udbredte er følgende fra omslaget på *Chemometrics & Intelligent Laboratory Systems*, det ene af de to primære kemometriske videnskabelige tidsskrifter: "Chemometrics is the chemical discipline that uses mathematical and statistical methods to design or select optimal procedures and experiments, and to provide maximum chemical information by analysing chemical data". Denne definition fremstår mildt sagt en smule kedelig!

Der er mere energi i den kemometriske angrebsvinkel end definitionen antyder; det handler ikke bare om "Anvendt matematik" eller "Anvendt statistik". Vi vil gerne vise at kemometri skelner mellem *anvendt* matematik og *anvendelig* matematik, og at det er den dynamiske dialog mellem problem/applikation og dataanalyse, der bidrager til ny innovativ indsigt i komplekse problemer.

### Hvorfor multivariat dataanalyse?

Den generelle antagelse i kemometri er at flere forskellige målinger på et sæt prøver er korrelerede; det er usandsynligt at målinger af protein, pH, fedt, kulhydrat, vand, salt, mættet fedt,

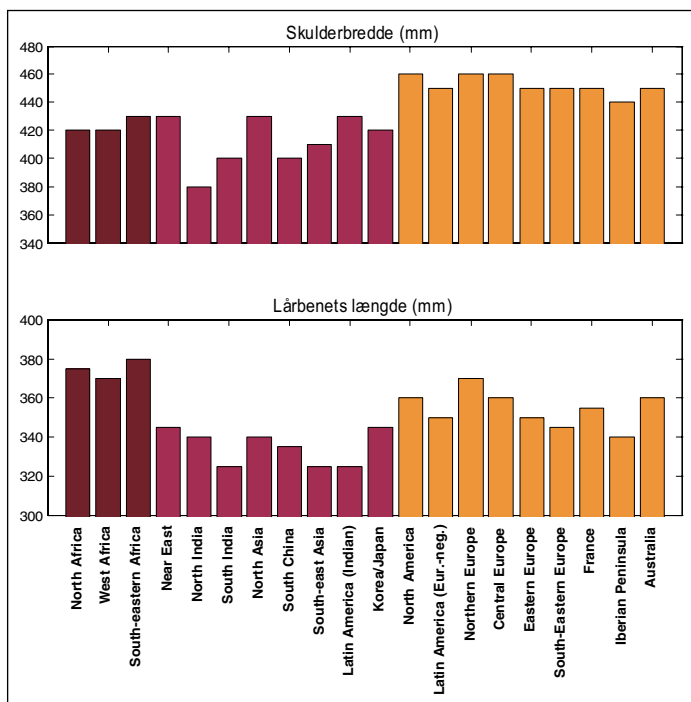
sensorik, vandbinding, m.m. i 100 forskellige kød-prøver ikke hænger sammen på en eller anden måde.

Kemometrien udnytter, at der er information i *sammenhængen mellem målingerne* som illustreret ved følgende eksempel. I antropologi er man bl.a. interesseret i antropometriske (studiet af menneskets fysiske dimensioner) målinger med henblik på at gruppere og karakterisere forskellige etniske grupper [1]. Som vist på figur 1 har man for tyve forskellige områder på jordkloden målt lårbenets gennemsnitlige længde (balde til knæ) og den gennemsnitlige skulderbredde (begge i millimeter) for mænd.

I studiet er man interesseret i om mennesker af afrikansk, asiatisk og kaukasiske oprindelse kan adskilles alene baseret på de antropometriske målinger. Værdien af lårbenets gennemsnitlige længde giver ikke mulighed for at gruppere i forhold til race; der er dog en klar tendens til at afrikanere har længere lårben og tilsvarende at asiater generelt korte lårben. Minimums-maksimumsværdierne for de tre racer er 370-380 mm, 325-345 mm og 340-370 mm for henholdsvis afrikansk, asiatisk og kaukasiske race; det ses, at der overlap mellem intervallerne parvis, og det er således ikke mu-

	Energi (kJ/g)	Protein (%)	Kulhydrat (%)	Fedt (%)	Mættet fedt (%)
Apple Pie	11.5	2.7	32.2	15.0	4.3
Big Mac	9.5	12.4	19.6	11.0	4.0
Cheeseburger	10.5	13.7	26.0	10.2	4.4
Filet-O-Fish	11.8	10.8	26.4	14.8	2.8
Grilled Chicken	8.1	12.6	14.7	9.4	0.0
Hamburger	10.0	12.6	28.8	8.0	2.5
McChicken	9.2	11.3	19.3	10.8	1.7
McFeast	8.6	11.6	14.6	11.3	4.5
Pommes Frites	12.2	5.0	37.1	13.5	2.0
Quarter Pounder m/ost	10.5	15.6	17.9	12.9	6.4
Sundae Chokolade	7.8	4.3	28.4	6.2	4.4
Sundae Jordbær	6.8	3.6	28.3	3.6	2.4
Sundae Karamel	7.8	4.0	31.7	4.7	3.1

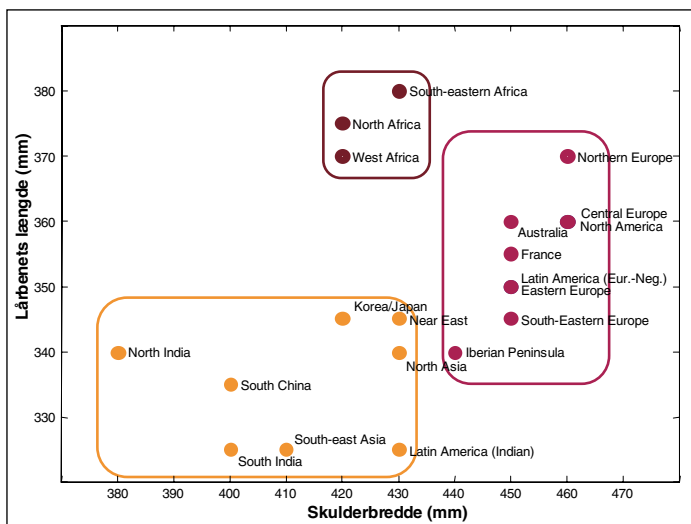
Næringsstoffer i udvalgte McDonalds produkter.



Figur 1. Lårbenets gennemsnitlige længde og tilsvarende skulderbredde (begge i millimeter) for mænd fra tyve forskellige lokaliteter og tre racer.

ligt at skelne perfekt mellem racerne baseret på lårbenets længde alene.

Betrager vi skulderbredden på samme måde er min-max værdierne 420-430 mm, 380-430 mm og 440-460 mm for henholdsvis afrikansk, asiatisk og kaukasiske race. Kaukasierne er generelt mere bredskuldrede, mens der er interval overlap for afrikanere og asiater. Heller ikke denne måling giver *alene* mulighed for at gruppere racerne.



Figur 2. Lårbenets længde afbildet mod skulderbredde for data vist i figur 1.

Ved at plote lårbenets længde mod skulderbredden fås figur 2. I denne er de tre racer adskilte baseret på information fra *begge* målinger. Det er således kombinationen af to målinger, der muliggør denne adskillelse. Der opstår ny information, som ikke er tilgængelig ved at betragte hver variabel for sig. Det er altså utilstrækkeligt og uvidenskabeligt kun at analysere én variabel ad gangen!

Sandsynligheden for at værdifuld information forbliver

updaget er ikke bare teoretisk. Denne sidste konklusion har selv den amerikanske Food and Drug Administration definitivt erkendt ved i deres nylige Process Analytical Technology (PAT) vejledning [2] for den farmaceutiske industri at skrive ”Traditional one-factor-at-a-time experiments do not effectively address interactions between products and process variables”. PAT vil blive emnet for en fremtidig artikel i Det Kemometriske Rum.

## Latente variable metoder

Ovennævnte eksempel viser, at der findes information latent (skjult) i multivariable data og mange af de metoder, der anvendes i kemometrien, kaldes derfor også *latente variable metoder*. Vi har i eksemplet betragtet to variable (lårbenslængde og skulderbredde) målt på tyve prøver (mennesker). Grundlæggende set handler kemometri om *2D datateknologi* dvs. analyse af multiple målinger på et sæt af sammenlignelige objekter. Men hvis vi nu i stedet måler 1000 variable per prøve, f.eks. et nærinfrarødt spektrum, er det ikke optimalt at plote variabel 1 mod 2, variabel 1 mod 3, variabel 1 mod 4 osv. Det vil give en uoverstigelig mængde figurer og ikke bidrage til at give et overblik over sammenhængen i data. Til det formål er metoden Principal Component Analyse (PCA) velegnet og i næste klumme vil vi beskrive princippet bag PCA.

## Udfordring

En udfordring til læserne er følgende: i tabel 1 er angivet indholdet af næringsstoffer samt energi i forskellige McDonalds produkter (data fra 2003). Hvorledes kan man (f.eks. grafisk) placere eller karakterisere de forskellige produkter i forhold til hinanden baseret på indholdet af både fedt, protein, kulhydrat, mættet fedt og energi?

1. Lee S. and Bro R., Regional Differences in World Human Body Dimensions: The Multi-Way Analysis Approach, Theoretical Issues in Ergonomics Science, in press.

2. Guidance for Industry. PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. U.S. Department of Health and Human Services, Food and Drug Administration, September 2004 (<http://www.fda.gov/Cder/guidance/6419fn1.htm>).

### Klumeskriverne:

- Lars Nørgaard (LN) er ph.d. og lektor i eksplorativ dataanalyse & kemometri, Formand for Dansk Selskab for Kemometri
- Søren Balling Engelsen (SBE) er ph.d., professor og leder af faggruppen for Kvalitet & Teknologi
- Rasmus Bro (RB) er ph.d. og professor i kemometri og proces analytisk teknologi

Klumeskriverne er alle uddannet som civilingeniør i kemi fra Danmarks Tekniske Universitet og er ansat på Det Biovidenskabelige Fakultet, Københavns Universitet. De har mere end 10 års undervisnings erfaring i kemometri og kvantitativ spektroskopi med over 2000 studerende/deltagere. Deres forskningsområde er kemometri, kvantitativ spektroskopi, proces analytisk teknologi, metabolomics, kvalitetskontrol, molekylær funktionalitet og sundhedseffekter af plantebaserede fødevarer. Dette har ledt til en omfattende videnskabelig produktion, herunder to bøger om kemometri, udvikling af kemometrisk software og tæt industriel kontakt. Se mere på [www.models.life.ku.dk](http://www.models.life.ku.dk) og [www.odin.life.ku.dk](http://www.odin.life.ku.dk).