

## Arbejdshesten i multivariat kalibrering: Partial Least Squares

**Partial Least Squares (PLS) regression er et alternativ til Principal Component Regression. Det handler bl.a. om, hvordan man handler bedst i et supermarked**

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

Vi vil her introducere Partial Least Squares (PLS) regression som et alternativ til Principal Component Regression (PCR). Baggrunden for at introducere en ny regressionsmetode er følgende: PCR er en totrins metode, hvor man først beregner scores ( $T$ ) fra en data-tabel  $X$  (f.eks. NIR spektre) og dernæst laver en regressionsmodel til den afhængige variabel ( $y$ , f.eks. kvalitet). Det svarer til, at man går ind i et supermarked ( $X$ ), indkøber varer i forskellige afdelinger som frugt & grønt, kød, desserter, vine etc., og først når varerne er betalt, får man at vide hvilken menu ( $y$ ), man skal lave til middagen. Det er klart, at når man først vælger informationen i  $T$  uden at tænke på, hvad den skal bruges til, så risikerer man, at kalibreringsmodellen, der relaterer  $T$  til  $y$ , bliver unødigt kompliceret.

PLS er opstået som et alternativ til denne måde at lave regressionsanalyse på: i PLS-regression anvendes  $y$  direkte til at finde den relevante information  $T$  i  $X$ ;  $y$  indgår således i første trin af PLS-algoritmen, og ikke først senere. Dette svarer til, at man går ind i supermarkedet med menuen i hånden og derfor har mulighed for at indkøbe præcise varer, man skal bruge. Man skal således ikke sikre sig, at al relevant information i  $X$  er repræsenteret i  $T$ , men kan nøjes med at uddrage den relevante information. Dermed bliver modellen nemmere at fortolke og forstå, og dette aspekt er ofte centralt i forhold til at optimere og udvikle en analyse.

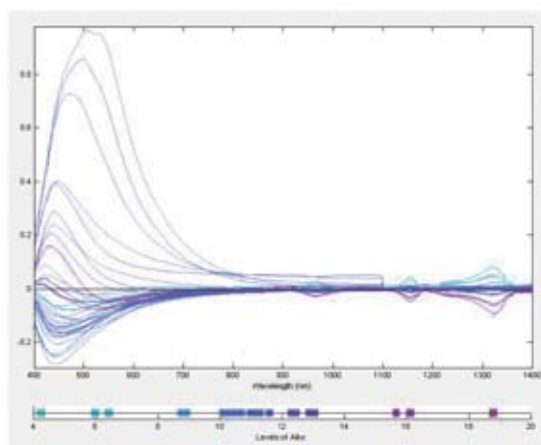
PLS-regression anvendes nu på data fra ølproduktion, og sammenlignes med en PCR-model på de samme data (beskrevet i en tidligere klumme i Dansk Kemi, nr. 8, 2008).

### Øldata

Fyrre prøver bestående af forskellige øl er analyseret for ekstraktindhold i % plato med en laboriemetode. Ekstrakt er en vigtig kvalitetsparameter i bryggeriindustrien og indikerer gærens potentiale til at danne alkohol. Ekstraktprocenten varierer fra 4,2-18,8% plato. De samme fyrre 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 siliciumbaseret detektor i området 400 nm til 1100 nm og blyulfid detektor (PbS) i området 1100 nm til 2500 nm. Figur 1 viser de centrerede spektre for alle prøver farvet efter ekstraktkoncentrationen.

Hvert spektrum er farvekodet efter ekstraktindholdet, og for de centrerede data er det tydeligt, at ekstraktkoncentrationen, som forventet, afspejles bedre i det nærinfrarøde spektrale område (f.eks. omkring 1200 nm) sammenlignet med det synlige (f.eks. omkring 500 nm). Selvom der tydeligvis er mest variation i det synlige område, er informationen i dette område åbenbart ikke relevant for ekstrakt.

Figur 1



Figur 1. Centrerede absorptionsspektre for fyrrø 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, og området fra 1100 nm til 1375 nm ses at afspejle ekstraktkoncentrationen bedre end det synlige område.

### PLS versus PCR

Der beregnes nu en PLS-model på de givne data med NIR-spektrene som  $X$  og ekstrakt som  $y$ . Helt analogt til PCR beregnes et antal PLS-komponenter; disse kaldes helt specifikt PLS-komponenter for at understrege, at de *ikke* er lig med de principale komponenter.

Forskellene mellem PLS og PCR kan illustreres ved inspektion af de forklarede varianser i  $X$  og  $y$  for fem komponenter for både PLS og PCR; disse kan ses i tabel 1, side 56. For både PCR og PLS ses, at første komponent ikke forklarer en særlig stor del af ekstraktvariationen. Dette er egentlig lidt i modstrid med, hvad man ville forvente, specielt for PLS. Første komponent forventes at være den vigtigste komponent, da PLS netop i første komponent leder efter den variation i spektrene som er mest relateret til ekstrakt. Årsagen til den lidt besynderlige første komponent vil vi vende tilbage til i en senere klumme omkring variabel-selektion.

Bevæger vi os videre til anden komponent, så ser vi at anden PLS-komponent forklarer mere af ekstraktvariationen end PCR, og dette er netop fordelene ved PLS. De komponenter man finder i PLS er mere relevante for  $y$  end tilfældet er for PCR.

For fem komponenter beskriver PLS-modellen 98,7%, mens PCR beskriver 94,7%. For  $X$  forholder det sig omvendt: forklaret  $X$ -varians i PCR-modellen vil altid være højere end for

Komponent	PCR		PLS	
	Forklaret %-varians i X	Forklaret %-varians i y	Forklaret %-varians i X	Forklaret %-varians i y
1	94,9	1,8	92,6	5,5
2	4,0	24,8	6,1	37,9
3	0,7	36,3	0,8	40,7
4	0,2	31,6	0,3	11,5
5	0,1	0,2	0,0	3,2
Sum fem komponenter	99,9	94,7	99,8	98,7

Tabel 1. Forklaret %-varians for X og y i en PLS og PCR-model af øldata.

PLS-modellen, fordi PCA netop finder præcis de komponenter, der bedst muligt beskriver - alt i - X.

I figur 2 ses hvorledes ekstrakt bliver estimeret ud fra henholdsvis en fire- og fem-komponent PLS-model. Korrelationskoefficienten mellem estimeret og målt er hhv. 0,96 og 0,99 for de to modeller. Valget af antal komponenter er vigtigt, for selvom prædiktionerne fra fem-komponent-modellen ser bedst ud, så er vi interesseret i at modellen er bedst på nye prøver. Og det er ikke sikkert, at det er fem-komponent-modellen, der vil være bedst på nye prøver, blot fordi den er god til at prædiktere de prøver, der blev brugt til at lave modellen. Vi vil senere se på teknikker til objektivt at afgøre, hvor mange komponenter man skal bruge.

En PLS-model giver de samme diagnostiske redskaber som PCR. Man får scores, loadings, regressionskoefficienter og residualer. Disse kan bruges til fortolkning, til validering og til at finde mulige outliers.

Forskellene mellem en PLS-model og en PCR-model er sjældent dramatiske, mht. hvor godt man prædikterer, men oftest anvender man et lavere antal komponenter i PLS-modellen. PLS kan især have en fordel, hvis det er små variationer i X, der er relevante for y, mens PCR kan have en fordel, hvis y er meget støjfyldt, da y anvendes to gange i PLS-algoritmen. Der findes ydermere en variant af PLS kaldet PLS2, som kan bruges, når man har mange forskellige y-variable. Det kunne f.eks. være,

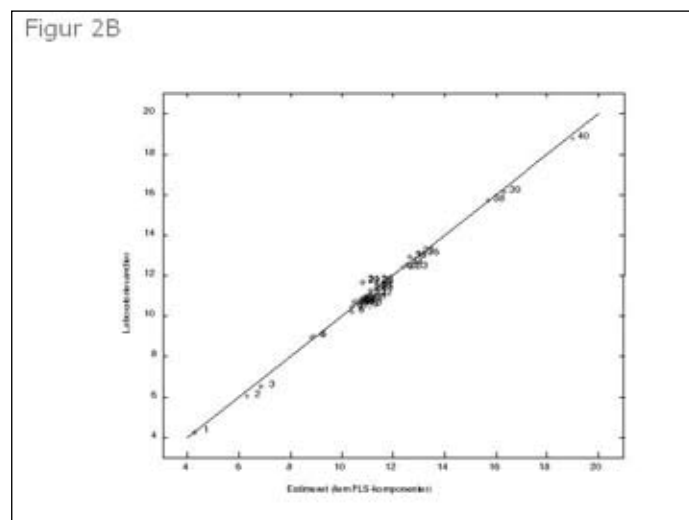
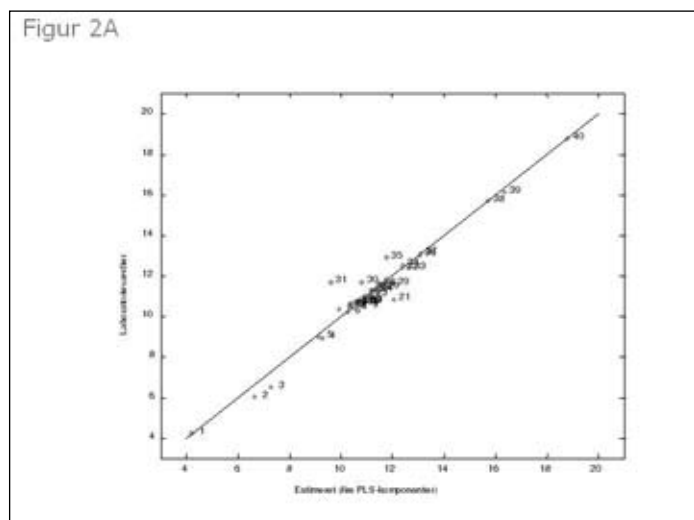
at man ville prædiktere udbytte, spild og energiforbrug. Ved hjælp af PLS2 kan man lave disse tre forskellige modeller på én gang og således få mulighed for direkte at forstå, hvorledes de tre forskellige kvalitetsparametre spiller sammen. Dermed kan man f.eks. lettere finde det rette kompromis, som kun maksimerer udbyttet i den grad, det ikke går dramatisk ud over energiforbrug.

En dansk indføring i PLS og multivariabel kalibrering kan findes i reference [1].

## PLS algoritme

En enkel version af PLS-algoritmen ses nedenfor. Der findes andre mere optimale versioner, men nedenstående viser hvordan y involveres i modelleringen straks fra start:

1. Centrér eller autoskalér X og y
2. Løs  $\mathbf{X} = \mathbf{y}\mathbf{w}' + \mathbf{E}_1$  mht.  $\mathbf{w}$   
Løsning: " $\mathbf{w} = \mathbf{X}/\mathbf{y}$ " eller  $\mathbf{w} = \mathbf{X}'\mathbf{y}(\mathbf{y}'\mathbf{y})^{-1}$   
Normalisér  $\mathbf{w}$  til længde én
3. Løs  $\mathbf{X} = \mathbf{t}\mathbf{w}' + \mathbf{E}_2$  mht.  $\mathbf{t}$   
Løsning: " $\mathbf{t} = \mathbf{X}/\mathbf{w}$ " eller  $\mathbf{t} = \mathbf{X}\mathbf{w}(\mathbf{w}'\mathbf{w})^{-1}$   
dvs.  $\mathbf{t} = \mathbf{X}\mathbf{w}$  da  $|\mathbf{w}|=1$



Figur 2. Estimerede ekstrakt % værdier baseret på nærinfrarød spektroskopi og PLS sammenlignet med de målte laboratorieværdier. A) Fire-komponent PLS-model. B) Fem-komponent PLS-model.



- Løs  $\mathbf{X} = \mathbf{t}\mathbf{p}' + \mathbf{E}_2$  mht.  $\mathbf{p}$   
Løsning: " $\mathbf{p} = \mathbf{X}'\mathbf{t}$ " eller  $\mathbf{p} = \mathbf{X}'\mathbf{t}(\mathbf{t}'\mathbf{t})^{-1}$
- Løs  $\mathbf{y} = \mathbf{t}\mathbf{b}_{score1} + \mathbf{e}_y$  mht.  $b_{score1}$   
Løsning: " $b_{score1} = \mathbf{y}'\mathbf{t}$ " eller  $b_{score1} = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{y}$
- $\mathbf{X}_{ny} = \mathbf{X} - \mathbf{t}\mathbf{p}'$  ( $= \mathbf{E}_x$ )  
 $\mathbf{y}_{ny} = \mathbf{y} - \mathbf{t}\mathbf{b}_{score1}$  ( $= \mathbf{e}_y$ )  
 $\mathbf{b}_{PLScores} = [b_{score1}, b_{score2} \dots b_{scoreA}]$
- Start fra trin 2 med  $\mathbf{X}_{ny}$  og  $\mathbf{y}_{ny}$  for at beregne næste PLS-komponent  
(op til det søgte antal komponenter)

### Outro

PLS anvendes dagligt i rigtig mange industrielle korn- og mejeriapplikationer baseret på NIR og IR, men metoden kan anvendes på alle typer af multivariate data, hvor man ønsker at sammenkoble og fortolke information fra to matricer.

### E-mail-adresser

Lars Nørgaard: lan@life.ku.dk

Rasmus Bro: rb@life.ku.dk

Søren Balling Engelsen: se@life.ku.dk

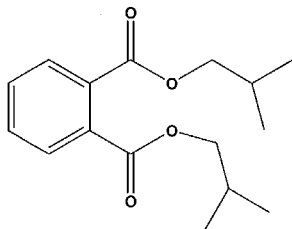
### Referencer

1. R. Bro, Håndbog i multivariabel kalibrering, Jordbrugsforlaget, 1996 (ISBN: 8774324586).

## Nyt om...

### .... Blødgørere i pizzaer

I Italien indeholder mange take away-pizzaer DIBP. Det ender i pizzaen, fordi den leveres i kartonæsker, som indeholder genbrugspapir. Italien har forbud mod brug af genbrugspapir til fødevarer, men alligevel finder analyser, at luften i den varme pizzaboks indeholder phthalatet.



Diisobutylphthalat - DIBP

DIBP er under mistanke for hormonforstyrrende virkning, men det er usikkert om de mængder, der er fundet indebærer nogen helbredsrisiko.

*Carsten Christophersen*

Monica Bononi og Fernando Tateo *Packing Technology and Science*. Offentliggjort online 19. november 2007.

# SCANLAF

## MARS

### Klasse 2 sikkerheds kabinetter

- bedste operatør komfort med et lydniveau under 54 dB(A)
- laveste energiforbrug og CO<sub>2</sub> - emission på under 0.9 A
- tillader installation i lav loftede rum med arbejdsbord højder op til 100 cm
- stort program for alle applikationer
- dansk produceret



# SCANLAF

Experts in Laminar Flow, Cooling and Vacuum Technology

ScanLaf A/S

Nøglegaardsvej 20, Vassingerød, 3540 Lyngby, Denmark

Tel: +45 3940 2566, Fax: +45 4498 1741

Mail: scanlaf@scanlaf.dk, www.scanlaf.dk