

Her er sammedragene til det 18. Norske kjemometricsymposium, Øyer 20-22 mars 2006.  
De er sortert etter mottatt fil-størrelse i kB!

## **PPLS: Kort oversikt over en vri på vanlig PLS-regresjon**

Ulf Indahl, IKBM, UMB

PPLS er blitt vårt ”in-house” akronym for metoden presentert av foredragsholderen i artikkelen ”*A twist to partial least squares regression*”, *J. Chemometrics* 2005, 19: p. 32–44. Metoden har vanlig PLS1-regresjon som spesialtilfelle, men kan i tillegg søke fram ladningsvektorer som i større grad favoriserer enkeltvariable dersom disse viser seg å bidra til god prediksjonsevne.

Foredraget gir en rask oversikt over de viktigste ideene bak metoden, hvordan den kan benyttes som eksplorativt verktøy og motivasjon av utvidelser til problemer med multi-respons (PLS2) og kategorisk respons (diskriminantanalyse).

### Analysing multisample 2DE proteom data on the level of pixels

E.M. Færgestad<sup>1</sup>, Frans van der Berg<sup>2</sup>, H.Grove<sup>1,3</sup>, F.Westad<sup>1</sup>, Ø.Langsrud<sup>1</sup>, K.Hollung<sup>1</sup>, A.H.Bjerke<sup>1</sup>, H.Martens<sup>1,3</sup>

<sup>1</sup> Matforsk as, Osloveien 1, 1440 Ås, Norway

<sup>2</sup> Quality and Technology, Spectroscopy and Chemometrics group, The Royal Veterinary and Agricultural University, KVL

<sup>3</sup> Departement of chemistry, biotechnology and food science, University of Life Sciences

A novel method for analysing two-dimensional electrophoretic data is presented. The method consists of an automatic aligning procedure developed by Frans van der Berg, KVL, followed by unfolding of the gels, between gel adjustments, data reduction and multivariate analysis on the level of pixels. Proteins found to be significant are subsequently analysed on the level of spot volume. Finally, the results are evaluated using technical replicates of the same sample, and subsequently on biological replicates. The data analysed are from muscles of ten cattle before and after slaughtering where seven animals constitute the calibration data, and three animal constitute the biological test set.

## **Biokjemometri - EMSC forbehandling av kjemiske og fysiske interferenter i FTIR**

Harald Martens(1,2,3), Susanne W. Bruun(4), Achim Kohler(1,5)

(1) Matforsk, Ås

(2) CIGENE-IKBM(UMB), Ås

(3) KVL, Danmark

(4) Biocentrum, DTU, Danmark

(5) U.Nantes, Frankrike

Fourier Transform Infra Rød (FTIR) spektroskopi med biokjemometrisk datamodellering er en informativ, rask og billig måleteknikk for metabolom-målinger.

Men en del systematiske feil gjør dataene mer komplekse enn nødvendig:

Vanndamp og CO<sub>2</sub> i spektrofotometerets lysvei fører til irriterende spektrale forstyrrelser.

Variierende vanninnhold i biologiske prøver gir store – ofte uønskede -effekter på IR spektret.

Variasjoner i prøvenes temperatur fører til store endringer i vannets spektrum.

Saltinnholdet i prøvene påvirker også vannets spektrum.

Vi rapporterer her en måte å kvantifisere og korrigere disse effektene. Dataene er basert på modellforsøk i et FTIR instrument med ATR målecelle. Metoden bruker Extended Multiplicative Signal Correction (EMSC) forbehandling til selektivitets-forbedringen.

### **On-line multisppektral transflektansspektroskopi for analyse av fett i laksefiléter - en praktisk kalibreringsstrategi**

Vegard H. Segtnan, Jens Petter Wold, Bjørg Narum & Frank Lundby

Matforsk AS

Måling av gjennomsnittlig fettinnhold i laksefiléter vha NIR-spektroskopi er en hyppig testet men vanskelig applikasjon. Fordi fettinnhold varierer svært mye innenfor hver enkelt filét, både horisontalt og vertikalt, er klassisk refleksjonsspektroskopi ikke veldig praktisk for denne type applikasjon, og transmisjonsmålinger er vanskelige på grunn av fiskeskinnet. For å kunne plukke opp både horisontal og vertikal informasjon om fettfordeling trenger man transfleksjonsspektra, dvs. at instrumentets lyskilde og detektor sitter på samme side av prøven. Normalt krever dette kontakt mellom prøve og

instrument/probe, hvilket ikke er ønskelig fra et praktisk hygienisk ståsted. Første versjon av en norskprodusert multispektral transfleksjonsskanner er under implementering hos en laksebedrift. I løpet av mindre enn ett sekund bygges det opp en datablokk bestående av omtrent 300x60 spektra i NIR-området, og denne skal brukes til å måle gjennomsnittlig fettinnhold i hver enkelt laksefilét.

En strategi for spektroskopisk sampling og referansesampling brukt i denne applikasjonen vil bli presentert. Fokus blir lagt på å lage en mest mulig robust og relevant regresjonsvektor.

Avsløring av årsaken til fotooksidasjon i meieriprodukter ved bruk av fluorescensspektroskopi og multivariat kurveoppløsning

**Jens Petter Wold, Frank Lundby and Annette Veberg**

**Matforsk**

Meieriprodukter er generelt utsatt for fotooksidasjon på grunn av naturlig innhold av lysfølsomme stoffer. Best beskyttelse mot fotooksidasjon og påfølgende kvalitetsforringelse er å skjerme produktene mot UV stråling og synlig lys. Dette er imidlertid i konflikt med forbrukernes ønsker; de vil se hva de kjøper. Utvikling av en gjennomsiktig emballasje som samtidig beskytter produktet er derfor viktig for produsentene. En forutsetning for dette arbeidet er detaljert kjennskap til hvilke aktive lys-sensibilisatorer som finnes i produktene, og deres spektrale egenskaper.

Riboflavin har lenge blitt regnet som den aktive lys-sensibilisatoren i meieriprodukter. Vi har nylig påvist at det i tillegg er naturlige forekomster av ulike porfyriner og klorofyll, svært lysfølsomme stoffer som er kjent for å initiere fotooksidasjon. De spektrale egenskapene til disse molekylene er kjent, men vi vet ikke om alle eller kun noen er aktive i oksidasjonsprosessen i meieriprodukter.

Oppdagelsen av de "nye" lysfølsomme stoffene ble gjort ved bruk av fluorescensspektroskopi (målt direkte på intakt produkt), der alle de aktuelle molekylene har distinkte spektrale profiler.

I dette foredraget vil vi presentere en tilnærming som effektivt kan peke ut de viktigste lys sensibilisatorene i meieriprodukter. Den er basert på fire steg: 1) Designet belysningsforsøk som involverer ulike farger på lyset, variasjon i oksygentilgjengelighet og eksponeringstid, 2) Måling av fluorescens spektra (bl.a. eksitasjons/emisjons landskaper) og sensorisk analyse av prøvene, 3) Separering av rene spektrale komponenter (lys-sensibilisatorene) ved bruk av multivariat kurveoppløsning (PARAFAC og independent component analysis, ICA), og 4) Forholde score verdiene for de ulike komponenter til de sensoriske responser. Resultater for smør vil bli presentert. Forskjellene mellom PARAFAC og ICA blir diskutert.

## ***RESOLUTION OF AN UNKNOWN MIXTURE WITH MULTIVARIATE CURVE RESOLUTION***

Valérie Lengard and Dongsheng Bu\*

CAMO Software AS, Oslo, Norway

\*CAMO Software Inc., Woodbridge, New-Jersey, USA

vl@camo.no, dbu@camo.com

Multivariate Curve Resolution methods intend the recovery of concentration and response profiles of the components in an unresolved mixture using a minimal number of assumptions about the nature and composition of these mixtures. In the present application, fifty-three samples of water, blue dye and yellow dye were prepared following a mixture design. UV/VIS spectra were measured. The data are analysed with MCR-ALS. No initial guess is provided to the algorithm. Constraints of non-negativity and closure are applied. The algorithm detects the presence of two constituents and reconstitutes their spectral profiles. The relative concentrations of the dyes are estimated.

### **Acknowledgement**

The experimental data were provided by Dr. Jean-Paul Wilhelm, Ciba Specialty Chemicals, Switzerland.

### **References**

- 1) Bu D, Brown C, *Appl. Spectrosc.* **54**, 1214 (2000)
- 2) de Juan A, Tauler R, Chemometrics applied to unravel multicomponent processes and mixtures - Revisiting latest trends in multivariate resolution, *Anal. Chim. Acta*, **500**, 195-210 (2003)

**Endo- og exo-LPLS regresjon i et fugleperspektiv.**

Solve Sæbø

Institutt for Kjemi, Bioteknologi og Matvitenskap, UMB

LPLS regresjon (LPLSR) ble presentert av Martens et.al (2005) som en metode for å utforske kovariansstrukturen i tre datamatriser (X, Y og Z) arrangert i et "L-formet" system. Denne formen for LPLSR karakteriseres av en "innad-rettet" regresjon av regressanden Y fra to regressorer X og Z og kan derfor betegnes som endo-LPLSR. Tilsvarende kan en utadrettet regresjon av to regressander X og Z fra én regressor Y betegnes som exo-LPLSR (Martens, 2005). I denne presentasjonen studerer vi endo- og exo-LPLSR i et fugleperspektiv. Vi bruker data fra VinterfuglAtlas prosjektet (i regi av Norsk Ornitologisk Forening) til å sammenlikne endo- og exo-formen av LPLS regresjon og bruker metodene til å besvare spørsmålene alle lurer på mht til økologien til fuglene i vinter-Norge. Her får du f.eks høre at tunge fiskespisende arter helst holder seg langs kysten, og at fjellfugler trives, ja faktisk, i fjellet. Presentasjonen er basert på et arbeid av Sæbø et.al (2006) som vil komme i Handbook of Computational Statistics – PLS and Marketing.

Referanser:

Martens, H., Anderssen, E., Flatberg, A., Gidskehaug, L.H., Høy, M, Westad,F., Thybo, A og Martens, M. (2005), Regression of a data matrix on descriptors of both its rows and of its columns via latent variables: L-PLSR. *Computational Statistics and Data Analysis*, PLS'01 Special Issue, **48** 103-123.

Martens, H. (2005), Domino-PLS: A framework for multi-directional path modelling. In T. Aluja, J. Casanovas, V. Esposito Vinzi, A. Morineau and M. Tenenhaus (eds.) *PLS and related methods*. SPAD. pp 125-132.

Sæbø, S., Martens, M. og Martens, H. (2006), Three-block data modeling by endo- and exo-LPLS regression. Submitted *Handbook of Computational Statistics – PLS and Marketing*.

***En praktisk tilnærning til on-line spektral analyse***

T.V.Karstang, Rune Mathisen og Frode Brakstad, MUST AS

Ved bruk av spektroskopiske sensorer for å sikre stabil industriell produksjon oppstår gjerne en rekke praktiske utfordringer. Sentralt står metoder for flytting av kalibreringsmodeller, rask gjenkjenning og fjerning av interferenser, samt en god kommunikasjon mellom laboratoriet og kontrollrom. Hvis ikke disse utfordringen løses på en praktisk, er det vår erfaring av gode målepunkter og metoder ikke tas i bruk til tross for en betydelig nytteverdi for prosessens (og produktets) stabilitet.

Foredraget gir et eksempel på hvordan disse utfordringene kan løses.

**Separering av protein-segment fra støy-segment ved DPLSR i 2D-gel elektroforese**

Morten Beck Rye og Bjørn Alsberg (veileder)

Segmentering av er et viktig steg i prosessen for å automatisere analysen av to-dimensjonale proteingeler. Standard segmenteringsrutiner for gel-bilder klarer imidlertid ikke alltid å skille bildesegment forårsaket av proteiner fra segment forårsaket av andre, uønskede forurensninger. Det vil derfor være ønskelig med en klassifisering av segmenter etter at standardsegmenteringen er utført. En vanlig type uønskede artefakter består av såkalt 'hvit støy' og skiller seg visuelt klart fra protein-relaterte bildesegmenter. Vi har laget en deskriptor som beskriver denne visuelle forskjellen, og basert på denne har vi klart å skille segmenter av denne typen fra andre segmenter ved hjelp av DPLSR.

**Design av en selektiv inhibitor for human 8-oxoguanin DNA glykosylase 1 (hOGG1)**

Kristin Tøndel<sup>1</sup>, Jon K. Lærdahl<sup>2</sup>, Torbjørn Rognes<sup>2</sup> og Lars Eide<sup>2</sup>

<sup>1</sup>*Institutt for kreftforskning og molekylær medisin, Det medisinske fakultet, NTNU.*

<sup>2</sup>Gruppe for bioinformatikk, CMBN, Institutt for medisinsk mikrobiologi, Rikshospitalet.

Ytre påvirkninger fra for eksempel stråling og oksygenradikaler kan forårsake skader på DNA. DNA-skader som ikke repareres fører til endringer i genmaterialet (mutasjoner). Cellene har mange ulike metoder for å reparere skader på DNA, og 8-oxoguanin DNA glykosylase 1 (OGG1) er et enzym som er involvert i slik DNA-reparasjon. OGG1 gjenkjenner og fjerner skadede baser fra DNA-molekylet. Genet som koder for OGG1 finnes i ulike varianter, og defekte former av OGG1 er funnet i pasienter med lungekreft og nyrekreft. Imidlertid har nyere studier vist at OGG1 også kan være skadelig i visse nevrologiske sykdommer som Huntingtons sykdom og Cockayne's Syndrome, ved at den starter en reparasjonskaskade som har en verre virkning for nerveceller enn den opprinnelige DNA-skaden. Et medikament som blokkerer funksjonen til OGG1 kan derfor tenkes å beskytte mot disse sykdommene. Vi har brukt Protein Alpha Shape Similarity Analysis (PASSA) kombinert med *de novo* ligand design og ligand docking for å finne en selektiv inhibitor for hOGG1. PASSA sammenlikner strukturen til målproteinene med andre liknende proteiner vha. blant annet Diskriminant PLS, for å finne områder i strukturen hvor målproteinene skiller seg ut fra de andre proteinene. Disse strukturelle områdene kan utnyttes for å oppnå selektiv binding til målproteinene. Vi har brukt resultater fra PASSA sammen med programmet LigBuilder til å lage en database med legemiddelkandidater, som har blitt rangert etter hvor godt de binder til hOGG1. De mest lovende forbindelsene vil senere bli testet eksperimentelt for binding til hOGG1 vha. celle-assay og dyremodeller.

### **Optimering av produksjonskostnader og produktkvalitet når råvarene varierer.**

Ingrid Måge<sup>1,2</sup> og Tormod Næs<sup>1</sup>

<sup>1</sup> Matforsk, Osloveien 1, 1430 Ås

<sup>2</sup> EWOS Innovation AS, 4335 Dirdal

I de fleste produksjonsprosesser er det et stort fokus på å minimere kostnadene, samtidig som produktkvaliteten må være tilfredsstillende. Når råvarene varierer, kan det påvirke både produksjonskostnadene og kvaliteten på sluttproduktet. Produksjonskostnadene kan påvirkes ved at råvarer av ulik kvalitet prises forskjellig, og også ved at ulike råvarer krever ulik prosessering.

Produksjonen kan optimeres ved å minimere kostnadene, med restriksjoner om at kvaliteten på sluttproduktet må oppfylle gitte spesifikasjoner. For å utføre en slik optimering trengs gode modeller for både kostnader og produktkvalitet. Når råvarene varierer må denne variasjonen inkorporeres i begge disse modellene.

Jeg vil illustrere en slik optimering ved hjelp av et eksempel fra fiskefôrindustrien. Fiskefôr produseres i en ekstruderingsprosess, og de viktigste råvarene er fiskemel, hvete, mais, soya, og olje. Spesielt en av disse råvarene har en betydelig kvalitetsvariasjon som påvirker fôrets egenskaper. Vi har tidligere vist at denne variasjonen kan måles med NIR spektroskopi, og vi har utviklet modeller som beskriver sluttproduktet som en funksjon av både prosessvariable og NIR-spektra av råvarene [1].

I denne presentasjonen kommer jeg til å snakke om innsamling av data, modellering og optimering. Optimeringens robusthet og sensitivitet med hensyn til kvalitetsspesifikasjoner og råvarepriser vil også bli diskutert.

[1] I. Måge, and T. Næs, Split-plot regression models with both design and spectroscopic variables, Journal of Chemometrics, In press (2006).

***Produksjonsforsøk: Modellering av split-plot og multiple responser***  
Frøydis Bjerke og Øyvind Langsrud – Matforsk, Are Halvor Aastveit – UMB

Sammendrag

I foredraget drøftes to forhold rundt designede forsøk: Restriksjoner på randomiseringen og multiple responser. Typiske randomiseringsrestriksjoner er knyttet til forsøksfaktorer som er vanskelige å endre på, eller til flerfaktor-forsøk med et sekvensielt forløp. En såkalt split-plot-modell beskriver en form for restriksjon på randomisering. Det må tas

hensyn til randomiseringsrestriksjoner både i planleggingen og i den statistiske analysen av forsøkene. Korrelasjon mellom multiple responser må håndteres med statistiske metoder som reduserer risikoen for Type I-feil, altså at tilsynelatende signifikante effekter egentlig er tilfeldig variasjon.

Et eksempel fra et pølseproduksjonsforsøk er beskrevet, hvor en split-plot-modell med multiple, korrelerte responser analyseres, i dette tilfellet GC-MS-data som beskriver mengden av forskjellige flyktige komponenter i pølseprøvene. Målet med forsøket var å undersøke hvordan ulike faktorer (bl.a. råstoff, lagringsbetingelser) påvirker harskning i pølser. I eksempelet benyttes Langsruds 50-50 MANOVA og tilhørende rotasjonstester til den statistiske analysen av forsøksdataene. Foredraget beskriver også de praktiske forholdene rundt planlegging og gjennomføring av forsøket, for å belyse den nære sammenheng mellom praktisk gjennomføring og statistisk analyse.

#### *Referanser*

Langsrud, Ø. (2002) 50-50 Multivariate Analysis of Variance for Collinear Responses, *The Statistician* 51 (3), pp. 305-317.

Langsrud, Ø. (2005) Rotation Tests, *Statistics and Computing* 15 (1), pp. 53-60.

### **Using re-sampling methods for statistical inference in GEMANOVA/PARAFAC models?**

Frans van den Berg

The Royal Veterinary and Agricultural University (KVL), Denmark

Department of Dairy and Food Science, Quality and Technology ([fb@kvl.dk](mailto:fb@kvl.dk))

Traditionally, Design of Experiments (DoE) data is analyzed using additive regression models. Recently multiplicative factor models have been used successfully to get insight into the variance contributions of DoE-data<sup>1-3</sup>. Strong arguments for using multi-way multiplicative factor model analyses of design-data are: (1) The data might simply fit better in a multiplicative representation! (2) The data collected during experimentation can remain in its 'natural shape' (hyper-cubes), and analyzed as such. (3) All the (powerful) graphically-oriented data-analysis tools from chemometrics remain accessible during DoE analysis. (4) And specifically for the GEMANOVA/PARAFAC, the uniqueness of the solution, together with a sound interpretation of the variable-loadings based on the experimenter's domain knowledge, provides for a powerful interpretation of the results. There is however one serious drawback for using these models: statistical inference (e.g. parameter probabilities or significance testing) is not yet 'invented' for complex multiplicative settings, and closed-form/analytical solutions to e.g. standard

errors for factor models will be very hard or even impossible to find. Statistical inference is desirable for two reasons: (1) Even if a design-effect is interpretable for the domain expert based on variable-loadings, analysis should show if it is real/significant, reducing the risk of over interpretation. (2) To get a wider acceptance of factor-based models in the more statistically-oriented community.

To tackle the lack of a closed-form/analytical solution we propose a studentized statistic based on bootstrap re-sampling<sup>4</sup>. We will show how this approach can result in estimates for parameter uncertainty and parameter/design-effect probability values, serving the same role they have in 'classical additive ANOVA'. The concept and theory involved will be illustrated using a large scale study from food packaging and storage as case study.

(1) J. Mandel 'A new analysis of variance model for non-additive data' *Technometrics* 13(1971)1-18; (2) R. Bro and M. Jakobsen 'Exploring complex interactions in designed data using GEMANOVA. Color changes in fresh beef during storage' *Journal of Chemometrics* 6(2002)294-304; (3) L. Nannerup, M. Jakobsen M, F. van den Berg, J. Møller, J. Jensen and G. Bertelsen 'Colour preservation of MA-packed sliced meat products by control of critical packaging parameters' *Meat Science* 68/4(2004)577-585; (4) A. Davison and D. Hinkley 'Bootstrap Methods and their Application' Cambridge University Press (1997)

## **Design and manufacture of Portland cement - application of sensitivity analysis in exploration and optimisation**

### **Part II: Optimisation**

#### ABSTRACTS

##### Purpose

Based on the conclusion of Part I: Exploration, a program, OptPilot, for a model-based optimisation of  $y$  constrained by PLS-components has been developed. The development of the program is based on principles presented by Svinning, Ingerøyen and Dalsveen on SSC6.

The purpose or utility of a model-based optimisation could depend on the different approach to experimental planning prior to the PLS-modelling. Two main routes of approach could be outlined: (a) explorative design based on random and natural sampling and (b) controlled experiments with factorial design followed by new experiments added two the previous with optimisation design. In the case of the latter approach the optimisation could be carried out prior to the modelling by *analysis of variance* (ANOVA). The lack of design may be due to the preliminary nature of the project as a high risk of obtaining a process out of balance or deteriorating the quality of the product

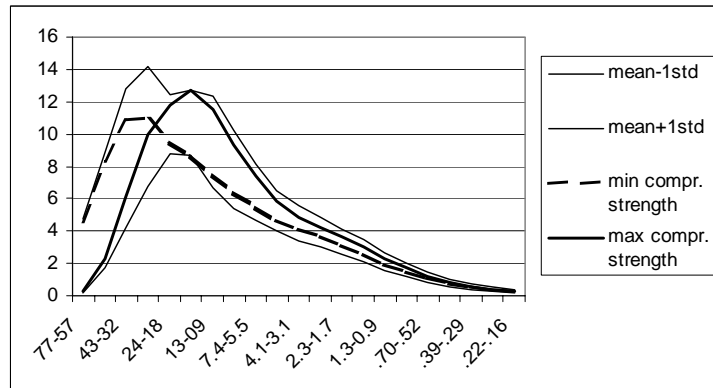
as a result of uncritical variation of a variable. An alternative to application of optimisation design could be a model-based optimisation of the response variable constrained by PLS-components or latent variables.

#### Methods

The optimisation based on a PLS-model could have the form of a linear program, where the constraints describing the influence of one variable on the others are given by one original PLS component or one equal a combination of several. OptPilot contains subprograms in Optimal and Mainpilot. Mainpilot does minimising or maximising along a principal direction of variation in  $\mathbf{X}$  given by a PLS-component or a combination of several. Optimal is searching for the optimal combination of latent variables, which gives max or min  $y$ .

## Results

Figure 1 shows examples of minimising and maximising compressive strength at 28 days as a function of size distribution of particle in cement within the limits of  $\bar{x}_j \pm std(x_j)$ . The optimal combination of latent variable was  $t\mathbf{p}_{1+2} = t(0.22\mathbf{p}_1 + 0.78\mathbf{p}_2)$  giving min and max compressive strength equal to 49.6 and 59.4 Mpa, respectively.



## Conclusion

The method of model-based optimisation constraints by PLS-component or latent variables was very applicable for achieving a realistic results for implementation in the design of Portland cement with respect to performance and quality control during production.

### **Bruk av kjemometri i Prediktors AMO-system**

Ellen Nordgård-Hansen og Gisle Lundeby, Prediktor as, Postboks 296, 1601 Fredrikstad

I løpet av bare ett og et halvt år har Prediktor bygget opp forretningsområdet Avanserte Instrumenter, som i første omgang retter seg mot næringsmiddelindustrien.

I oktober 2005 leverte Prediktor to systemer til Atria Lithells AB i Örebro, som den foreløpig siste leveransen i en rekke som omfatter de norske næringsmiddelprodusentene Espeland, Fatland, Gilde, Tine og en stor norsk fôrprodusent. Leveransene består av systemer som måler innholdet av fett, vann og protein direkte i prosessen mens produksjonen pågår. Hensikten med målingene er å styre produktkvaliteten mer presist i henhold til reseptene, og å kontrollere at råvarene holder seg innenfor de oppgitte spesifikasjonene. Hovedkomponentene i Prediktor-systemet er en måleprobe, et nær infrarødt (NIR) spektrometer og programvaren APIS™ Meat Optimizer (AMO).

Bruken av kjemometri i leveransene er todelt, først offline kalibrering hos Prediktor, og deretter online beregninger som gjøres for hvert scan. Prediktor setter, sammen med kunden, opp en prøveplan for kalibreringen, og deretter analyseres resultatene hos Prediktor. Det tas her 3 uttak for hver batch, og hvert uttak analyseres offline i 3 paralleller. Dermed kan usikkerheten både i uttak og i offline analyse estimeres. NIR-proben sørger samtidig for at det blir lagret et antall NIR-spektre for hver batch. Disse spektrene blir analysert for utligger og deretter midlet, slik at hver batch til slutt representeres ved ett midlere NIR-spekter og en midlere offline-måling. Før kalibreringsmodellen bygges, blir spektrene normalisert, og behovet for andre forbehandlingsmetoder vurderes. For Atria Lithells fikk en f.eks. gode resultater ved bruk av MSC. Deretter modelleres innholdet av en eller flere av komponentene fett, vann og protein ved hjelp av PCR. Modellene valideres vanligvis ved kryssvalidering, dersom kunden ikke ønsker å ta ytterligere prøver for testsettvalidering. Kalibreringen utføres ved bruk av Unscrambler eller Baccos, og kalibreringen dokumenteres så i en kalibreringsrapport som sendes kunden.

Når proben settes i drift, kjøres normalisering, sentrering og regresjon online, der en bruker middelverdiene, loadingmatrisen og regresjonskoeffisientene som fremkom fra kalibreringen.

## **Modellreduksjon**

Rolf Ergon

Høgskolen i Telemark

PCR- og PLS -modeller basert på latente variabler har som oftest flere komponenter enn strengt tatt nødvendig, og av den grunn kan modellreduksjon være hensiktsmessig. Jeg vil her sammenligne to prinsipielt ulike metoder:

1. Ortogonal-signal-korreksjon (OSC) søker å fjerne  $Y$ -ortogonal informasjon fra  $X$  før regresjonen, og en populær algoritme for dette er OPLS [1]. Siktemålet med OSC var opprinnelig å oppnå bedre prediksjoner, men det har en ikke fått til. Prediksjonene fra OPLS er for eksempel helt identiske med de fra vanlig PLS-regresjon, og det er vist at en kan oppnå eksakt samme modellreduksjon med en similaritetstransformasjon etter PLS-regresjonen [2]. Denne transformasjonen medfører at opprinnelige ladninger blir projisert på kolonnerommet til  $\hat{Y}$ .
2. Som et alternativ kan en velge å projisere opprinnelige score på kolonnerommet til  $\hat{B}$ , og det ble først foreslått allerede på 80-tallet [3]. En kan her velge å utvide kolonnerommet til  $\hat{B}$  med valgte deler av kolonnerommet til  $P$  (PCR) eller  $W$  (PLSR) [4,5,6].

Begge metodene skiller ut det en kalle en  $Y$ -relevant del av  $X$ , og la oss her kalle disse for  $X_Y^L$  (L for ladningsprojeksjon som i metode 1) og  $X_Y^S$  (S for scoreprojeksjon som i metode 2). En sammenlikning viser at  $X_Y^S$  er en delmengde av  $X_Y^L$ , dvs. at  $X_Y^S$  er den minste delmengden av  $X$  (målt med for eksempel Frobenius-normen), og den er slik sett å foretrekke. En annen fordel er at  $X_Y^S$  kan brukes direkte for å finne  $\hat{Y}$ , noe som ikke gjelder for  $X_Y^L$ .

- [1] Trygg J, Wold S. *Orthogonal projections to latent structures, O-PLS*. Journal of Chemometrics 2002; **16**: 119-128.
- [2] Ergon, R. *PLS post-processing by similarity transformation (PLS+ST): A simple alternative to OPLS*, Journal of Chemometrics 2005; **19**: 1-4
- [3] Kvalheim OM, Karstang T. *Interpretation of Latent-Variable Regression Models*, Chemometrics and Intelligent Laboratory Systems 1989; **7**: 39-51.
- [4] Ergon, R. *Compression into two-component PLS factorizations*, Journal of Chemometrics 2003; **17**: 303-312
- [5] Ergon, R. *Informative PLS score-loading plots for process understanding and monitoring*, Journal of Process Control 2004; **14**: 889-897

- [6] Ergon, R. *Reduced PCR/PLSR models by subspace projections*, in press, Chemometrics and Intelligent Laboratory Systems 2006

## **Kvalitetskontroll av microarray bilder**

Øystein Gjerstad og Ulf Indahl IKBM, UMB

Støy og spotkvalitet er kjente problemer ved microarray-eksperimenter. Det er normalt at enkelte spotter er såpass misvisende at de må sensureres fra datasettet. Det er ønskelig med en mest mulig automatisert deteksjon av slike spotter.

Det er 2 årsaker som er vanlige ved slik sensurering av spotter:

1. Avvik i spotstørrelse, -form og plassering i forhold til grid-mønster i arrayet.
2. Mangel på likhet med andre spotter som representerer replikater for det samme genet. (Spotten i seg selv behøver ikke avvike fra et normalt utseende.)

Vi ser på muligheten for å trekke ut egnede egenskapsmålinger for en spot/"spotboks", og utvikle en klassifikasjons modell som basert på erfaringsdata (manuelle sensureringer) er i stand til å utføre en god automatisk sensur.

### **HySpex - high resolution, high speed hyperspectral cameras for laboratory, industrial and airborne applications.**

Ivar Baarstad, Trond Løke, Peter Kaspersen

Norsk Elektro Optikk AS, Solheimvegen 62A, N-1471 Lørenskog, Norway

Norsk Elektro Optikk AS has over the last ten years developed a series of compact high performance imaging spectrometer systems, or hyperspectral cameras (HySpex), partly funded by space and defense projects. The unique hyperspectral camera concept has also demonstrated significant potential for use in civilian airborne, laboratory and industrial applications of imaging spectrometry. Four different versions of the instrument have been realized so far, with the following main specifications:

<b>Module</b>	<b>VNIR-640</b>	<b>VNIR-1600</b>	<b>SWIR-320i</b>	<b>SWIR-320m</b>
Detector	CCD 640*480	CCD 1600*1200	InGaAs 320*256	CdHgTe 320*256
Spectral range	0.4-1 $\mu$ m	0.4-1 $\mu$ m	0.9-1.7 $\mu$ m	0.8-2.5 $\mu$ m
Spatial pixels	640	1600	320	320
Total FOV (across track)	18.4°	17°	14°	14°
Pixel FOV	0.5mrad	0.2mrad	0.75mrad	0.75mrad
Spectral sampling	5nm or 10nm	3.7nm	5nm	5nm
# spectral bands	128/64	160	160	256
Max frame rate to HD	500 or 850fps	120fps	350fps	100fps

The instrument design is flexible, and the specifications can be tailored to individual users and applications. The unique mirror based fore optics minimizes spherical and chromatic aberrations. A slit defines the instantaneous field of view, and a polarization independent transmission grating disperses the light spectrally before it is focused by a six-element lens system onto the focal plane array detector. The lens system has been optimized for minimization and equalization of the point spread function across the FOV and spectral range, as well as for minimization of distortions such as spectral keystone and smile effect. The high performance demonstrated in the optical simulations has been verified experimentally. All instruments are calibrated spectrally and radiometrically, using narrow band light sources and a calibrated integrating sphere in order to produce absolute radiance spectra (in  $W/m^2 \text{ nm sr}$ ) for each pixel in the image.

The instruments are capable of acquisition speeds compatible with airborne and industrial hyperspectral imaging applications. A tripod mountable rotation stage has also been designed, providing synchronous operation of the spectrometer with the scanning platform. This setup can be used to acquire lab or field measurements of stationary scenes.

Instrument design, results and applications will be presented.

### **Multivariate Curve Resolution Applied to Hyperspectral Confocal Images of Live Cells\***

David M. Haaland,<sup>1</sup> Howland D. T. Jones,<sup>1</sup> Jerilyn A. Timlin,<sup>1</sup> Michael B. Sinclair,  
Edward V. Thomas<sup>1</sup>, Linda Nieman, and David K. Melgaard<sup>1</sup>  
Sawsan Hamad,<sup>2</sup> and Willem F. J. Vermaas<sup>2</sup>

<sup>1</sup>Sandia National Laboratories, Albuquerque, New Mexico, 87185-0886, USA

<sup>2</sup>School of Life Sciences and Center for the Study of Early Events in Photosynthesis,  
Arizona State University, Tempe, AZ 85287-4501

We have developed a 3D hyperspectral confocal fluorescence microscope that can optically section live cells at diffraction-limited spatial resolutions. The design and operation of the microscope will be discussed along with its advantages over current commercial confocal microscopes. When coupled with multivariate curve resolution (MCR), the new microscope can resolve multiple spatially and spectrally overlapped emission components in the cells. However, because of the high degree of overlap, recovering pure emission spectra is a challenge. Some approaches to extracting the pure spectra using MCR will be presented along with a discussion of procedures used to

properly weight the data to compensate for the presence of heteroscedastic noise. The influence of initial guesses for the starting spectra on the MCR results and pixel selection based on genetic algorithms will be presented. These methods will be demonstrated with hyperspectral images of live *Synechocystis* cells. *Synechocystis* is a cyanobacterium, and is a member of a class of organisms responsible for a large fraction of carbon sequestration from the atmosphere. In these experiments, cells from wild type and from mutants lacking a photosystem or a step in chlorophyll biosynthesis were imaged to monitor the relative concentrations and spatial distributions of photosynthetic pigments in these bacteria, providing a way to directly localize specific pigments in a cell with sub- $\mu\text{m}$  spatial resolution. Studies as a function of incident laser power and photobleaching aid in the resolution of the pure emission spectra. The new microscope and associated multivariate analyses constitute an enabling new technology for cell imaging and for understanding a variety of molecular and physical processes occurring in live cells.

\*Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000. This work was funded in part by the US Dep't of Energy's Genomics: GTL program ([www.doegenomestolife.org](http://www.doegenomestolife.org)) under project, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling," ([www.genomes-to-life.org](http://www.genomes-to-life.org)).

## **PPLS2: Multiresponsvariant av PPLS**

Alf Synstad og Ulf Indahl, IKBM, UMB

I laboratorieforsøk måles som regel mer enn en egenskap for de prøvene man har tatt. Og gjør man en referanseanalyse måles det ofte flere egenskaper man kunne tenke seg å modellere. I slike tilfeller har man derfor mer enn en respons, og da kan det være av interesse å gjøre modellering med en prediksjonsmetode som kan håndtere mer enn en respons.

Power PLS2 (PPLS2) er en utvidelse av PPLS1 slik at situasjoner som har flere responser også kan håndteres, tilsvarende det man gjør med bruk av tradisjonell PLS2. Utfordringen i denne utvidelsen består i å finne et relasjonsmål som erstatter unirespons korrelasjon/kovarians og kan ta hensyn til flere responser.

Det klassiske valget for antall komponenter i unirespons tilfellet er det antall som gir minimum RMSEP (Root Mean Square Error of Prediction). Hvordan velge ut antall komponenter i en multirespons problemstilling har man derimot ingen ubestridt tommelfingerregel for. I denne presentasjonen vil vi blant annet diskutere en mulighet for utvelgelse av et felles antall komponenter for multirespons regresjonsproblemer.

Illustrasjon av PPLS2 komponentutvelgelse demonstreres på et datasett med 5 responser som tar for seg marine fettsyrer i svinekjøtt hvor fett er analysert gjennom FT-IR biomodulen.

## Status on the NTNU hyperspectral microarray scanner project

Bjørn K. Alsberg, NTNU

A commonly used microarray technology is based on labelling two samples, a control and test sample with fluorophores which are detected at two main emission wavelengths. However the intensities of the two fluorophores are often influenced by factors such as spectral overlap between the fluorophores and contributions from fluorescing impurities. Thus, using only two wavelengths is not sufficient to obtain the best fluorophore concentration and consequently also gene expression estimates. Instead a hyperspectral approach can be used where a whole spectrum is recorded in each array pixel. This enables the use of chemometric multicomponent methods to obtain improved estimates of the various fluorophore and impurity contributions. This type of microarray scanner was originally constructed by David Haaland and his group at Sandia Labs in the US where they demonstrated that much better signal to noise ratios could be obtained. As this scanner technology is very promising for the microarray field the Chemometrics and Bioinformatics Group (CBG) at NTNU decided two years ago to build a similar instrument. The current presentation will describe the status of the project and some of the plans for the near future.

### **PPLS-DA: Egenskapsuttrekking med klassifikasjonsdata for bedret prediksjon og tolkbarhet**

Kristian H. Liland og Ulf Indahl IKBM, UMB

Presentasjonen tar utgangspunkt i PPLS metoden, og viser hvordan ideen i denne kan tilpasses klassifikasjonsproblemer kalt PPLS-DA. Utvidelsen til multirespons regresjonstilfellet (PPLS2) og diskriminantanalyse (PPLS-DA), kan kombineres videre for utnyttelse av dummy-kodede klassifikasjonsdata kalt PPLS2-DA.

Relasjonskriteriene som henter ut vektorer til potensering og bestemmer potensene tilpasses multirespons og kategorisk respons. Blant mulige kandidater i det kategoriske tilfellet finner vi mellom- og innen-gruppe kovarians og optimalisering med hensyn på diskriminantmetoder.

I unirespons regresjon kan et kriterium med en passende  $\chi^2$ -fordeling som referanse benyttes for å finne enkleste modell som ikke er signifikant dårligere enn den beste modellen (identifisert ved kryssvalidering). Med kategorisk respons er det naturlig å innføre en binomisk referansefordeling til bruk i modellutvelgelsen. Det viser seg hensiktsmessig å ta LDA/QDA tilbake til sin opprinnelige form, uten forenkling med logaritmer, for å få direkte tilgang til estimerte aposteriori-sannsynligheter (sannsynligheten for tilhørighet til de forskjellige gruppene for hver enkelt observasjon). Vi ser på gjennomsnitt og produkt av aposteriori sannsynligheter og vurderer disse i mulige kriterier for utvelgelse av antall komponenter i en modell.

# Utvidelser til kryssmodellvalideringen med applikasjon på mikromatrisedata

Lars Gidskehaug, Inst. Kjemi, NTNU

Mikromatriseforsøk kjennetegnes ved at aktiviteten til veldig mange gener måles over

forskjellige forsøksbetingelser. Identifisering av aktive gener over de gitte betingelsene kan

være viktig for diagnose, prognose eller for forståelsen av biologiske prosesser.

Fra et

dataanalytisk perspektiv ønsker man å finne gener som gir en godt validert, prediktiv modell.

Kryssmodellvalidering sikrer at utvelgelsen av gener inkluderes i valideringen av modellen.

Dette er viktig for å unngå overoptimistiske resultater grunnet variabelseleksjonen.

Ofte gjør dataanalytikeren en del valg under analysen som ikke blir validert. Man kan

prøve mange forskjellige modellstørrelser eller algoritmer (som i maskinlæring) og velge den

som ser ut til å passe best. For lineære modeller kan det være fordelaktig å fjerne gener

etappevis for gradvis å nærme seg den beste modellen. I et slikt tilfelle må man bruke

alternative metoder for å si hvorvidt et gen er signifikant uttrykt, da pverdier og lignende ikke

finnes direkte. Det kan også tenkes at man ikke ønsker å ta stilling til signifikansnivåer i det

hele tatt, men derimot ønsker å finne det antall gener som gir den beste prediktive modellen.

En utvidet kryssmodellvalidering som også tester for forskjellige modellstørrelser, presenteres. Eksempelet viser hvordan en slik utvidelse kan brukes for å lette valideringen av

modeller der flere parametre varieres og få prøver er tilgjengelige for testing.

## Bruk av grafteori i kjemometrien

Arnar Flatberg

Innenfor multivariat analyse av bioinformatikkdata kreves det ofte integrering av heterogene

datatyper. Svært mange datatyper kan enkelt representeres i en grafstruktur hvor et

dataelement utgjør en node og en relasjon mellom to dataelement er en kant.

Slike tilfeller

oppstår ofte i bioinformatikkfagfeltet, f.eks. i søk av relasjonelle databaser, sekvenslikhet, kjemiske reaksjonsnettverk og litteratur.

En graf er ett sett av punkter koblet sammen av linjer, punktene er ofte kalt noder og linjene som kobler disse er kanter. Her ser vi på *enkle* grafer, definert som grafer hvor kantene ikke har en retning og det er ikke tillatt med mer enn en kant mellom hver node (uvektet kant), og det er ikke tillatt med kanter fra en node og til seg selv. Disse grafene blir kalt *bevisgrafer*.

I første øyeblikk ser det ut som en slik graf kun inneholder lokal informasjon i form av naboskap, men ved å undersøke prosesser på grafen som diffusjon og tilfeldig bevegelse kan vi beskrive globale egenskaper mellom nodene i grafen.

I dette foredraget vil jeg gi en oversikt over arbeid som er gjort på å beskrive informasjonen som ligger i en graf på vektorform. Det fokuseres på representasjoner med opphav fra lineære projeksjoner av prosesser på grafen. Anvendelsen av grafstrukturer er gitt ved analyse av mikromatrise data, hvor noder er gener og ulike bakgrunnskunnskap mellom gener er gitt ved kanter i grafen.