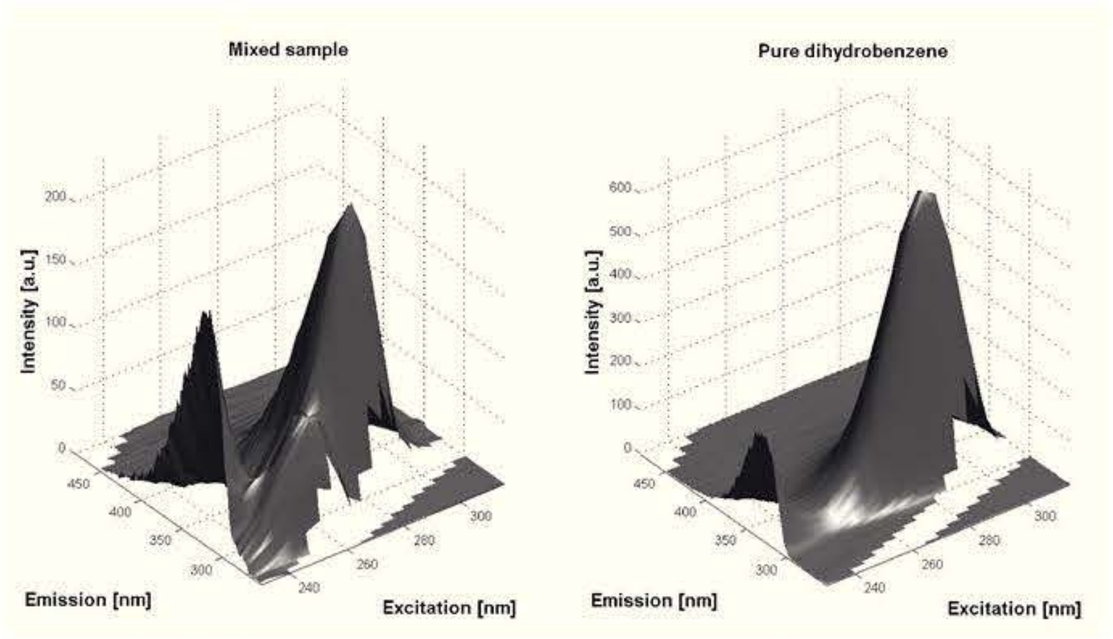
Trevejs-data skal modelleres med trevejs-modeller

Trevejs-data laves ofte om til tovejs-data for at kunne anvende standard PLS-modeller. Denne klumme illustrerer, at det giver langt bedre fortolkning og robusthed at anvende en trevejs PLS-model, når man har trevejs-data.

Af Rasmus Bro, Søren Balling Engelsen, Institut for Fødevarevidenskab, Københavns Universitet og Lars Nørgaard, FOSS

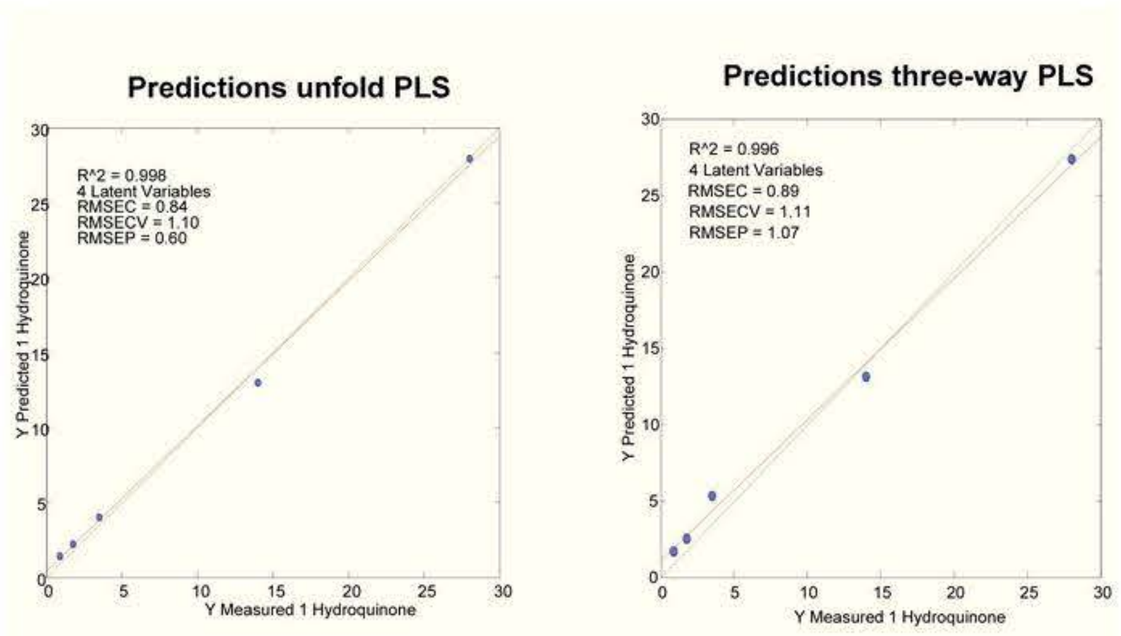
Vi har tidligere beskrevet (Dansk Kemi, 93, 1-2, 2012) et fluorescensdatasæt [1], som består af 23 prøver med varierende koncentration af phenylalanin, 3,4-dihydroxyphenylalanin (L-DOPA), dihydrobenzen og tryptofan. Vi vil nu udvikle en kalibreringsmodel, som kan prædiktere koncentrationen af dihydrobenzen ud fra de målte eksitations-emissions-matricer (EEM'ere).

I figur 1 (venstre) ses et typisk fluorescens-landskab af en prøve, der indeholder alle fire komponenter og til højre et landskab af en prøve, der kun indeholder dihydrobenzen. Hvert landskab består af en matrix, der angiver fluorescens-intensitet målt ved 116 emissionsbølgelængder og ved 18 eksitationsbølgelængder.



Figur 1. Eksitations-emissions-matrix (EEM) af blandingsprøve (venstre) og ren dihydrobenzen (højre).

Fem prøver gemmes som et testsæt og kalibreringssættet består altså af en 19 (prøver) ×116 (emissionsbølgelængder) ×18 (eksitationsbølgelængder) trevejs EEM-struktur og en 19×1 vektor med koncentrationer. Ud fra dette datasæt kan der enten laves en trevejs PLS-model, eller datasættet kan foldes ud til en tovejs-matrix med 116 gange 18 (=2088) variable. Med denne udfoldede 19×2088 matrix kan man så lave almindelig tovejs-PLS. Dette kaldes også *unfold-PLS*, fordi den laves på udfoldede data.



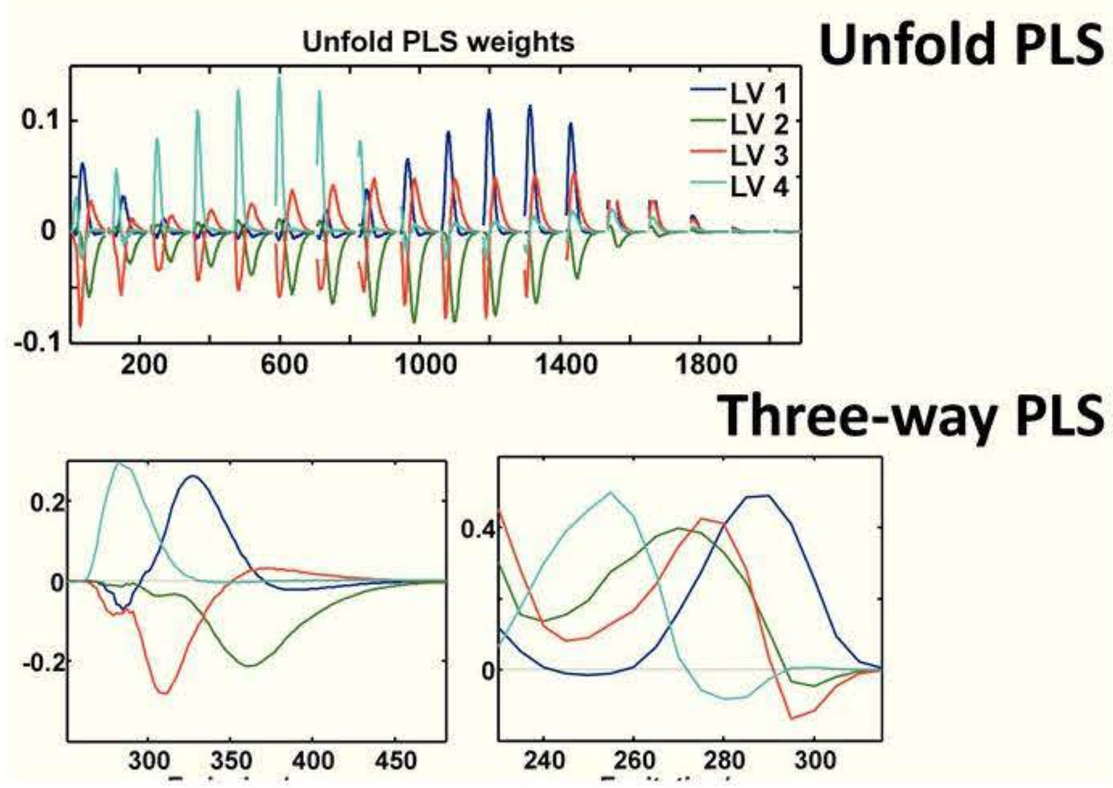
Figur 2. Prædiktioner af de fem prøver i testsæt vha. unfold-PLS (venstre) og trevejs-PLS (højre).

Forskellen på tovejs- og trevejs-PLS

En almindelig tovejs PLS-kalibreringsmodel på de udfoldede data giver ved krysdvalidering en model, som prædikterer glimrende med fire PLS-komponenter. Fire PLS-komponenter er også, hvad man ville vente, der var optimalt, når der er fire stoffer, der fluorescerer og varierer uafhængigt i prøverne.

Til sammenligning er også vist prædiktionerne fra en helt tilsvarende trevejs PLS-model. Som det kan ses af RMSEC- og RMSECV-værdierne er modellerne meget lig hinanden. Dog har trevejs-modellen noget højere prædiktionsfejl på testsættet (RMSEP). Det skal dog nok tages med et gran salt, givet de få prøver der er i dette sæt.

Unfold-PLS-modellen har også et sæt loadings, og som man kan se i figur 3, er det ganske vanskeligt at tyde disse loadings, fordi man har 2088 forskellige variable at "undersøge" for hver komponent. Det er faktisk et af hovedproblemerne, når man folder sine data ud. Prædiktionerne fra en model baseret på udfoldede data er ofte meget lig de prædiktioner, man får med en trevejs-model. Men selve fortolkningen af modellen er meget mere vanskelig. Og fortolkning er væsentlig for at kunne nå frem til gode modeller og for at kunne præsentere og diskutere disse.

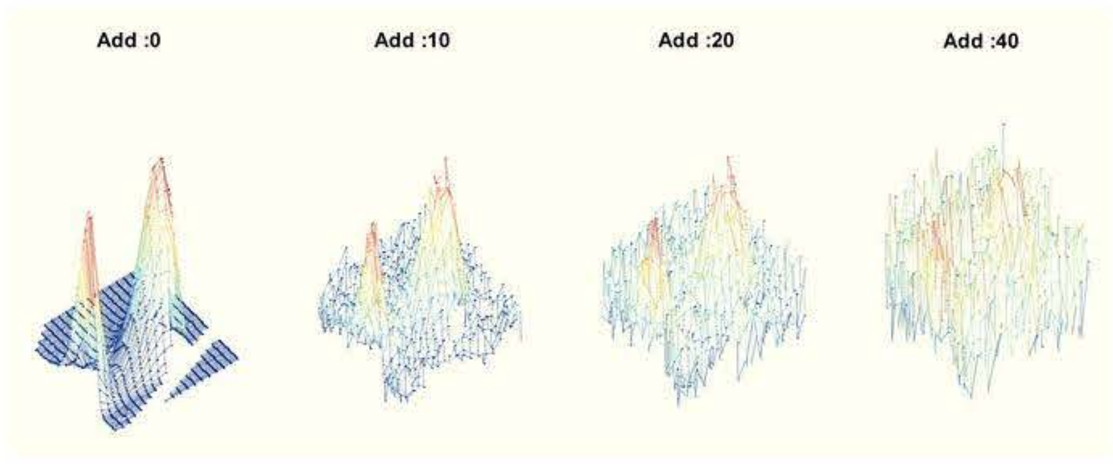


Figur 3. PLS-loading-vægte. Øverst for unfold-PLS og nederst for trevejs-PLS, hvor der ét sæt for emission og ét sæt loading-vægte for eksitation.

Nederst i figur 3 er vist, hvordan trevejs-PLS giver loadings, som passer langt bedre med strukturen af data og som derfor langt nemmere kan visualiseres og forstås. Det er f.eks. let at se, at første trevejs PLS-komponent, som forklarer hovedparten af respons-koncentrationen, har høje emissions- og eksitations-loadings i samme områder som dihydrobenzen (figur 1). Den første komponent er den mørkeblå.

Trevejs PLS-regression er robust

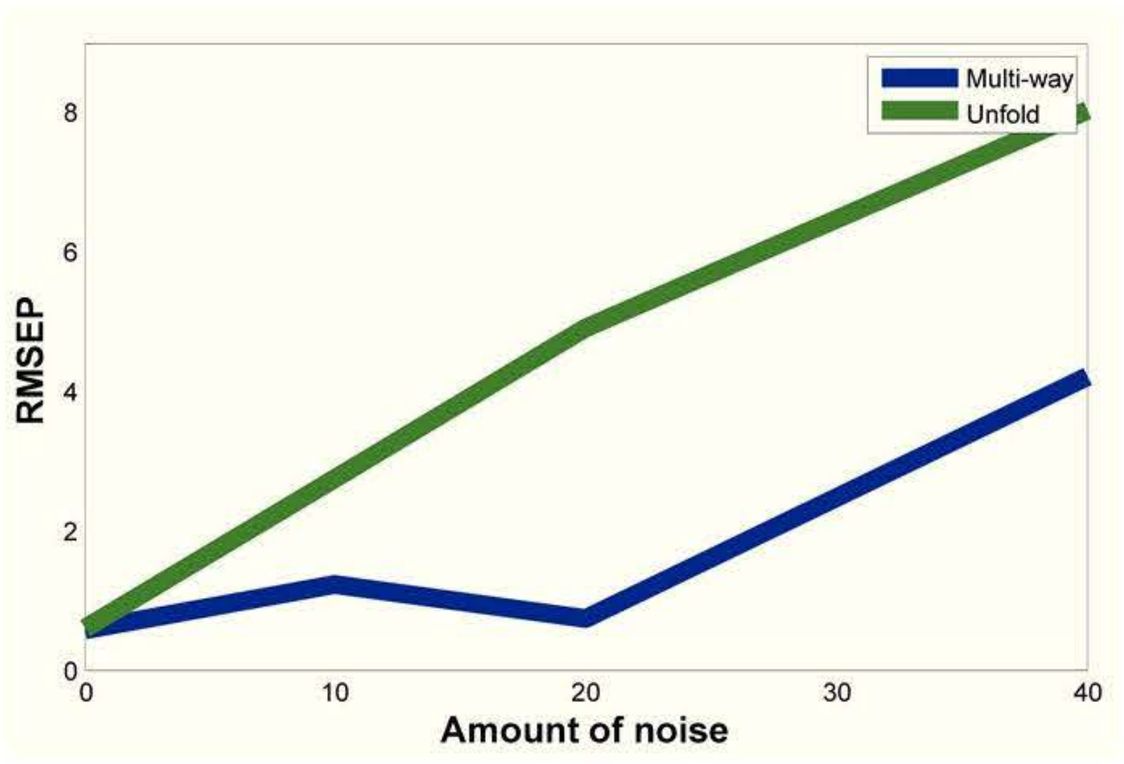
I et forsøg på at se hvordan trevejs- og tovejs-PLS håndterer støj, er ovenstående model blevet gentaget med mere og mere støj lagt på fluorescens-landskaberne. For hvert scenarie laves en tovejs og en trevejs kalibreringsmodel og testsættet prædikteres. Testsættet er ikke det samme som foregående model, men ellers er metoden til at lave prediktionsmodel den samme som ovenfor.



Figur 4. EEM-landskab for første kalibreringsprøve med forskellige mængder adderet støj (i fraktion af oprindelig støj).

I figur 4 kan man se fluorescens-landskabet for den første kalibreringsprøve i hver af kørslerne. Først uden adderet støj og dernæst med mere og mere støj lagt på. Som man kan se, bliver det til sidst næsten umuligt at se strukturen i fluorescens-landskabet.

Som det fremgår det af figur 5, kan trevejs-PLS prædiktere anstændigt selv ved meget store mængder støj, mens tovejs-PLS har større problemer. Dette peger på den anden væsentlige fordel ved trevejs-PLS. Da modellen indeholder langt færre parametre end tovejs-PLS, så er modellen mere robust overfor støj. Som eksempel så har én tovejs PLS-komponent en loading-vektor, der indeholder 2088 parametre. En enkelt komponent fra en trevejs PLS-model har til sammenligning bare 134 (116+18) parametre. Ikke nok med at de 2088 parametre er vanskelige at visualisere, de fører også til overfit (modellering af støj), når der ikke er yderligere systematisk variation at gå efter.



Figur 5. RMSEP-værdier for testsæt som funktion af mængden af adderet støj for både tovejs- og trevejs-PLS.

Outro

Det er demonstreret, at trevejs-data bør modelleres med trevejs-modeller. Det giver to væsentlige fordele: modellerne bliver simplere at fortolke og mere robuste overfor støj i data.

E-mail
Rasmus Bro: rb@life.dk.
Søren Balling Engelsen: se@life.ku.dk
Lars Nørgaard: lno@foss.dk

Referencer
1. D. Baunsgaard, Factors affecting 3-way modelling (PARAFAC) of fluorescence landscapes, Intern KVL rapport, August 1999, www.models.life.ku.dk/sites/default/files/dorrit.pdf