Kalibrering med multivejsdata - teorien bag

Ligesom vi kan lave PLS-regression med matrix-data, kan vi også lave PLS med multivejsdata. I denne klumme beskriver vi, hvordan en trevejs PLS er opbygget. Det gør vi med afsæt i den velkendte tovejs-matrix PLS.

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

Vi har tidligere beskrevet PLS-regression i detaljer; både teoretisk og mht. anvendelser. Ligesom i PCA, så får man scores og loadings i en PLS-model. Modellen af **X** data kan skrives:

$$\mathbf{X} = \mathbf{T}\mathbf{W}^{T} + \mathbf{E}$$

Hvor **T** er en score-matrix og **W** et tilsvarende sæt loadings. Af historiske årsager er der en del forskellige måder at beregne PLS-modellen på. Alle giver samme prædiktioner, men beregningerne foregår på lidt forskellig måde, og specielt loadings kan være lidt forskellige. Vi vil ikke gå i detaljer med det her, men blot nævne at den type PLS-model, som kan udvikles til multivejsdata, svarer til det man i litteraturen kender som Martens-versionen af PLS [1]. Dette er ikke den metode, som man normalt forbinder med den traditionelle NIPALS-algoritme [2], hvor der indgår et ekstra sæt loading-vektorer.

Teorien bag tovejs PLS-regression

Konceptet i en almindelig PLS-model med én afhængig y-variabel er at finde en score-matrix, som har følgende egenskaber startende fra komponent ét. Man finder en loading-vektor $\mathbf{w}_1$, som giver en score-vektor på vanlig vis – dvs. at 'mængden' af loading-vektor giver score-vektoren ved:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$$

På den måde sikres det, at score-vektoren er en del af **X** eller rettere, at **t** ligger i det rum, som kolonnerne i **X** udspænder. Det er vigtigt, fordi vi ønsker at scores for en ny prøve (hvor vi kun kender X-data og ikke y), skal kunne bruges til at prædiktere med.

Den score-vektor vi finder, skal være den score-vektor, som giver maksimal kovarians med **y**. Dvs., at vi ikke ville kunne vælge en anden $\mathbf{w}_1$ og få en **t**-vektor med højere kovarians med **y**. Grunden, til at man ønsker at maksimere kovariansen, er trefoldig. Kovariansen kan beskrives som korrelationen mellem **t** og **y** ganget med spredningen på hver af disse. Man ønsker, at dette produkt er så stort som muligt, og det betyder, at alle tre dele skal være (absolut) høje. Er en enkelt f.eks. nul, så vil produktet også være det. Rationalet er, at vi, ved at maksimere dette produkt, sikrer at:

> *Reel (stor) information i **X** (stor spredning på **t**) skal være lineært relateret (høj korrelation mellem **t** og **y**) til den vigtige (store) information i **y** (høj spredning af **y**, men det giver sig selv).*

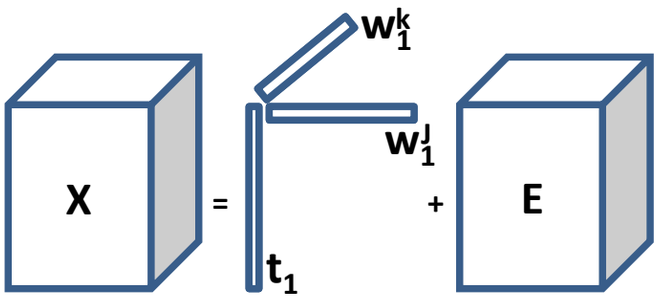
Det kan lyde lidt kryptisk, men denne del af PLS er den helt centrale grund til, at PLS-regression er et effektivt redskab i typiske kemometriske problemer. Vi sikrer, at valid information er beskrivende for vigtig information i **y** på lineær vis.

Når vi har fundet den første score-vektor, kan vi beregne et estimat af **y** ud fra den fundne information i **X** som:

$$\hat{\mathbf{y}} = \mathbf{t}_1 b$$

og dernæst kan man trække den beskrevne del af **X** ($\mathbf{t}_1\mathbf{w}_1^{T}$) og den beskrevne del af y ($\mathbf{t}_1 b$) fra henholdsvis **X** og **y**. Dette giver en residual for henholdsvis **X** og **y**.

Hvis beskrivelsen af **y** ikke er tilstrækkelig god, kan man beregne endnu en komponent ved at gentage hele proceduren, men nu med udgangspunktet i residualerne.



Figur 1. Første komponent i en PLS-model af x-data.

Trevejs PLS-regression

Ud fra ovenstående beskrivelse af tovejs PLS kan vi udvikle en trevejs PLS-regressionsmodel med tilsvarende egenskaber. Den eneste lille detalje, der adskiller de to, er, at i en trevejs PLS, er der ikke én, men to, loading-vektorer.

For hver af de to variabel-retninger finder man en loading som vist i figur 1. Ligesom tovejs PLS-modellen af **X** har samme (algebraiske) form som PCA, så har trevejsmodellen samme form som PARAFAC. Som i tovejs PLS-modellen, så er PLS-komponenten givet ved at vægtene $\mathbf{w}^J$ og $\mathbf{w}^K$ giver en score-vektor **t,** som har maksimal kovarians med **y**.

Det er vigtigt at understrege, at selvom PLS-modellen ligner PARAFAC, så er der ikke nogen unikke løsninger som i PARAFAC. Man får ikke matematisk kromatografi, men i stedet en løsning med egenskaber, der ligner en almindelig PLS-models egenskaber, men nu blot med to loadings i hver komponent.

Outro

Man kan som alternativ til trevejs PLS, folde sine data ud og lave almindelig tovejs PLS, men vi vil se, at det sjældent er en fordel. Trevejs PLS på trevejs-data vil for det meste give den bedste model ift. fortolkning og prædiktioner. Det bliver illustreret i næste klumme.

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

Referencer
1. H. Martens, T. Næs. Multivariate calibration, Chichester:Wiley & Sons, 1989.
2. A.Höskuldsson. PLS regression methods. J.Chemom. 2:211-228, 1988.

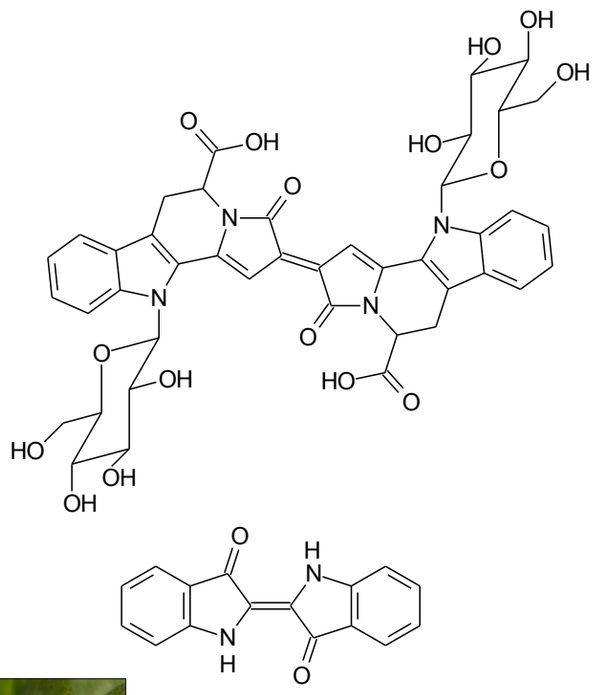
Nyt om …

… Ei blåt til lyst

Blå farver er sjældne i naturen. Fugle og sommerfugles blå farver skyldes ikke blå farvestoffer, men interferensfænomener i vinger og fjer. Endnu færre repræsentanter finder man, hvis man vil anvende de blå farver i levnedsmidler. Man finder blå farver blandt anthocyaninerne i f.eks. blåbær og hyldebær; men de er ikke anvendelige i levnedsmidler pga. farvens pH-afhængighed. Indigo er uopløseligt i vand. Det er nu lykkedes at finde en kandidat, glycosyleret trichotemin fra japanske Kusagi bær fra planten *Clerodendron trichotomum*. Man kan se af formlen, at den centrale del af molekylet har en vis lighed med indigo.

Carl Th.

Bringing blue to a plate near you *Chemical & Engineering News*, 10. Sept. 2012, side 30.



INDIGO