Hvis fit er godt, hvad er overfit så?

Kemometriske og statistiske metoder anklages ofte for at kunne finde sammenhænge, selv hvor der ikke basis for det. Det demonstreres, at dette er sandt for PLS-DA-modeller, men også at det er uhyre nemt at sikre sig imod overfitting vha. enkle valideringsteknikker

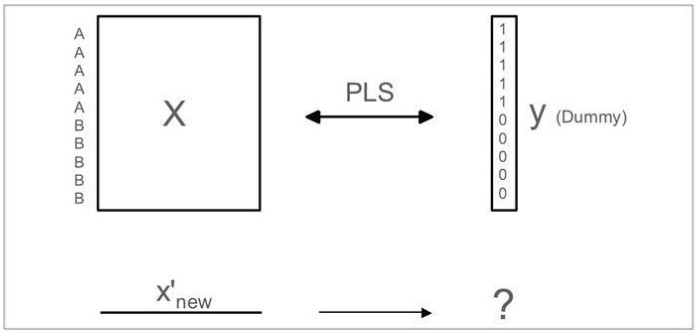
Af Lars Nørgaard, Søren Balling Engelsen og Rasmus Bro, Institut for Fødevarevidenskab, Det Biovidenskabelige Fakultet, Københavns Universitet

I eksplorativ empirisk modellering er formålet at opdage nye sammenhænge i data og efterfølgende opstille hypoteser og teorier baseret på de observerede sammenhænge. For at sikre validiteten af de observerede sammenhænge er det vigtigt, at der er tillid til den anvendte model, og at der er anvendt relevant validering af modellen. Dermed kan man undgå at antage sammenhænge i data, der ikke er basis for, og dermed undgår man at fortolke og danne hypoteser på et forkert grundlag. Man kunne fristes til at indskyde en sætning om klimadebatten her, men det undlader vi.

Som vist i den forrige klumme i nr. 10, 2009, så er PLS-DA et uhyre effektivt klassifikationsredskab. Her demonstreres det, hvor let man kan komme galt af sted med denne metode, hvis man glemmer sin validering.

Princip i PLS-DA

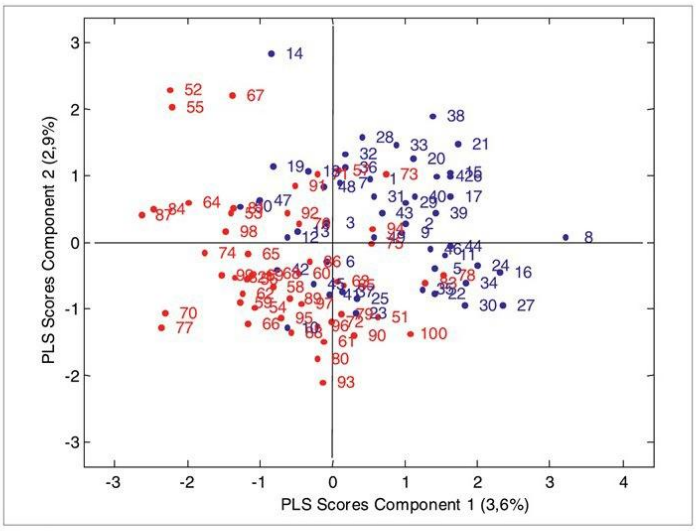
Som tidligere beskrevet er idéen bag PLS-DA [1] enkel: man introducerer en dummymatrix bestående af ettaller og nuller og med antal søjler lig med antal kendte grupper. Antal rækker svarer til antallet af kendte prøver. Et ettal i dummymatricen afspejler, at en given kendt prøve tilhører gruppen. Figur 1 illustrerer princippet for et toklassesystem; alle A-prøver har ettaller i dummyvektoren, alle B-prøver har nuller.

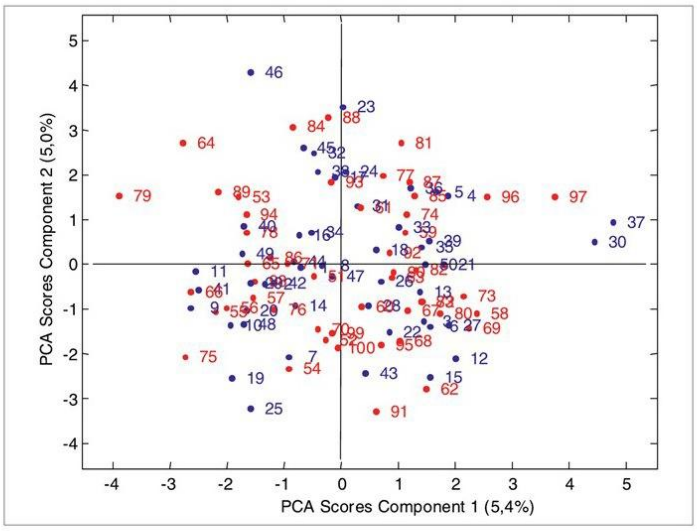


Figur 1. Illustration af princippet i PLS-DA for en togruppemodel.

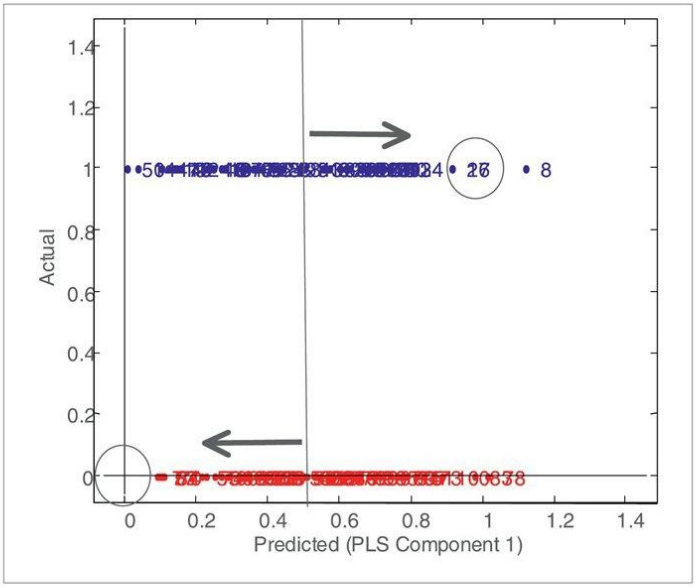
Eksempel

Vi beregner en PLS-DA på et prøvesæt, hvor **X** er tilfældige data: vi anvender en matrix med lutter tilfældige tal med dimensionen 100 prøver × 50 variable. Vi laver en tilsvarende y-vektor, der indeholder ettaller på de første 50 pladser og nuller på de resterende 50 pladser. Det betyder, at de første 50 prøver antages at komme fra samme klasse (A), og de næste 50 fra en anden klasse (B). Vi ved, at dette ikke er tilfældet, da data i **X** er tilfældige, men vi vil alligevel beregne en PLS-DA-model.





Figur 2.
a) PLS-DA-scoreplot baseret på tilfældige X-data. Der ses en klar tendens til gruppering (rød og blå), som let kan lede til overfortolkning.
b) PCA-scoreplot baseret på tilfældige X-data. Der ses ingen tendens til gruppering.



Figur 3. Actual versus Predicted i en krydsvalideret én komponent PLS-model. I en model med prædiktive egenskaber skal de blå prøver prædikteres til værdier højere end 0,5, og tilsvarende skal prædiktionsværdierne for de røde prøver være lavere end 0,5. Det fremgår, at modellen er uhyre dårlig til at forudsige prøvernes tilhørsforhold.

Tolkning af model

Et scoreplot fra PLS-DA-modellen er vist i figur 2a. Det ses, at der er en klar tendens til adskillelse i to grupper, om end der også observeres et overlap mellem grupperne. Dette er et normalt forekommende resultat for klassifikation af f.eks. biologiske prøver. Har man stået i laboratoriet i to år for at udvikle en spektroskopisk baseret målemetode (**X**) til at diagnosticere osteoporose (**y**) ud fra en blodprøve, er man meget opsat på at finde forskelle mellem spektrene for de raske og for patienterne, og man vil sikkert blive umådeligt glad for at se scoreplottet i figur 2a.

Scoreplottet viser, at der er fundet en retning i data, som kan adskille prøverne en smule. Men vi ved, at en sådan retning ikke giver mening. Årsagen er, at PLS aktivt søger efter en retning i **X**, der stemmer med **y**. Når en sådan ikke findes, vil PLS, i mangel af bedre, lede efter små tilfældige korrelationer i data, som der altid vil være. Normalt overskygges disse langt af de reelle informationer, men i dette tilfælde findes ingen sådanne. At det er små tilfældige variationer, PLS har brugt, kan også indirekte ses af, hvor lidt de to komponenter beskriver (se akserne i plottet som angiver forklaret varians i **X** for hver komponent).

En PLS-DA-model forsøger, at finde forskelle i **X** mellem de angivne grupper, men denne forskel gælder kun de aktuelle data og ikke et nyt datasæt. Dette kaldes også overtilpasning eller overfit. Modellen har altså ikke prædiktive egenskaber, og det er netop, hvad en model til diagnosticering skal have.

Validering

For ikke at lande i ovenstående misfortolkning skal der anvendes validering: man skal altid vurdere modellens prædiktive egenskaber ved f.eks. krydsvalidering, inden man begynder at fortolke baseret på den udviklede model. Har modellen ikke prædiktive egenskaber, må man ganske simpelt ikke fortolke på den. På ovenstående data beregnes en PLS-DA-model med segmenteret krydsvalidering, og som det første inspiceres *Actual versus Predicted*-plottet for én komponent (figur 3). Det fremgår, at det ikke er muligt at prædiktere om en given prøve tilhører gruppe A eller ej; helt som forventet. Prædiktionerne varierer fra 0 til knap 1,2 og ikke kun inden for det angivne grå område, som ville være ideelt eller blot mindre end 0,5, der er den pragmatiske grænse for en acceptabel klassifikationsmodel. Derfor kan man heller ikke anvende modellen til fortolkning, og enhver videre anvendelse af modellen stopper her.

I figur 2b ses et PCA-scoreplot af de samme tilfældige data, og det er klart, at der her ikke ses de samme tendenser som i PLS-scoreplottet. I PCA er der ikke nogen y-vektor som **X** styres over mod, og derfor er risikoen for overfit ikke til stede på samme vis. Til gengæld er det sværere at bruge PCA til at fange mindre kausale og dermed reelle forskelle i **X**, som rent faktisk vil kunne gruppere et givet datasæt. Alting har en pris.

Outro

Antallet af komponenter i PLS-DA-modeller skal estimeres ud fra f.eks. krydsvalidering eller testsæt-validering, præcis som det gælder for andre typer af PLS-modeller. Man bør altid vurdere, om modellen rent faktisk kan adskille eller delvist adskille grupperne; dvs. estimere værdien nul eller ét i dummymatricen. Hvis dette ikke er tilfældet, kan man ikke bruge modellen til klassifikation og ej heller til fortolkning.

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

Referencer

1. Ståhle L, Wold S. *Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study.* Journal of Chemometrics 1: 185-196, 1987.