Klassifikation med Partial Least Squares-Diskriminant Analyse

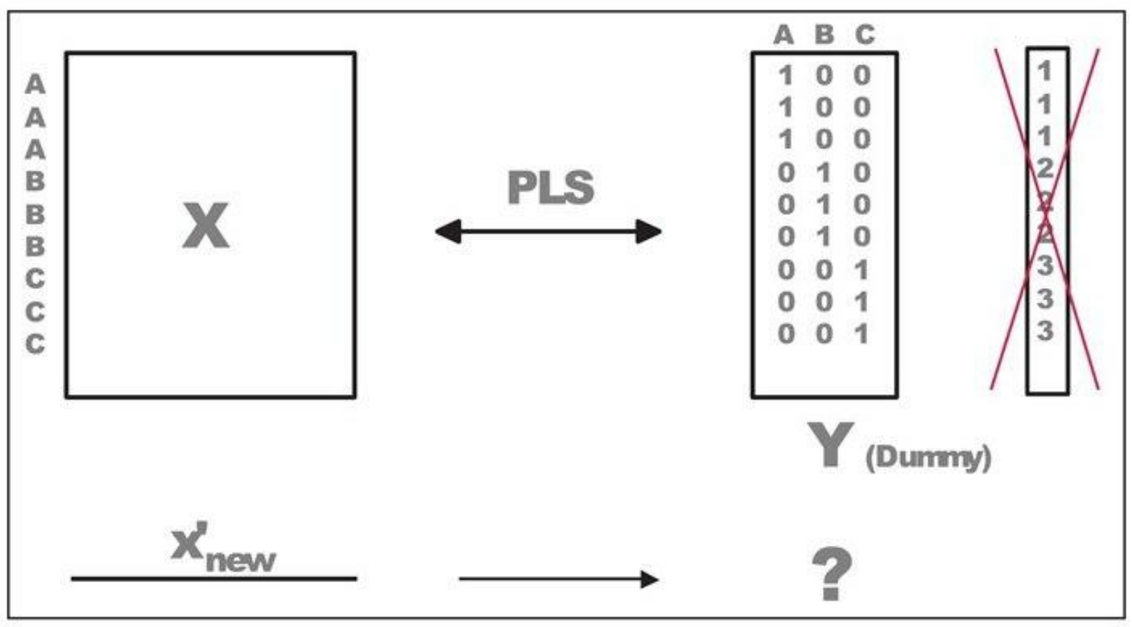
En alternativ klassifikationsmetode til SIMCA er Partial Least Squares-Diskriminant Analyse (PLS-DA). Metoden er udviklet med henblik på at finde variationer i data, der adskiller grupper af prøver

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

SIMCA er en effektiv klassifikationsmetode. Den er baseret på PCA-modellering af hver af de i forvejen kendte grupper efterfulgt af en sammenlignende projektion af ukendte prøver på hver af disse PCA-modeller. I SIMCA beskrives variationen inden for hver enkelt gruppe, uden hensyntagen til om variationen er relevant for adskillelsen af grupperne. Ved at anvende PLS som klassifikationsmotor fokuseres der på variationer, der kan adskille grupperne.

Princip i PLS-DA

Idéen bag PLS-DA [1] er uhyre 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; alle A-prøver har ettaller i første søjle af dummymatricen og nul på de resterende pladser, alle B-prøver har ettaller i anden søjle og nul på de resterende pladser, og så fremdeles. Det er således let at fremstille en dummymatrix, der indeholder information om prøvernes gruppetilhørsforhold.



Figur 1. Illustration af princippet i PLS-DA. Se tekst for mere information.

Der beregnes derefter en PLS-model med de målte data. Den danner baggrund for klassifikationen, som X-matrix og dummymatricen som Y. PLS-DA er således helt analog til en standard PLS-model. Den eneste forskel er, at Y-matricen kun indeholder ettaller og nuller og ikke f.eks. proteinkoncentrationer.

Hvis X-matricen indeholder variation, der er relevant for adskillelsen af grupperne, vil PLS-modellen med stor sandsynlighed finde den. I en kvantitativ PLS-model for protein er formålet at estimere proteinkoncentrationen af en ny prøve baseret på f.eks. et nærinfrarødt spektrum. I PLS-DA er formålet at prædiktere nuller og ettaller, der kan fortælle, om en ny prøve tilhører en given gruppe.

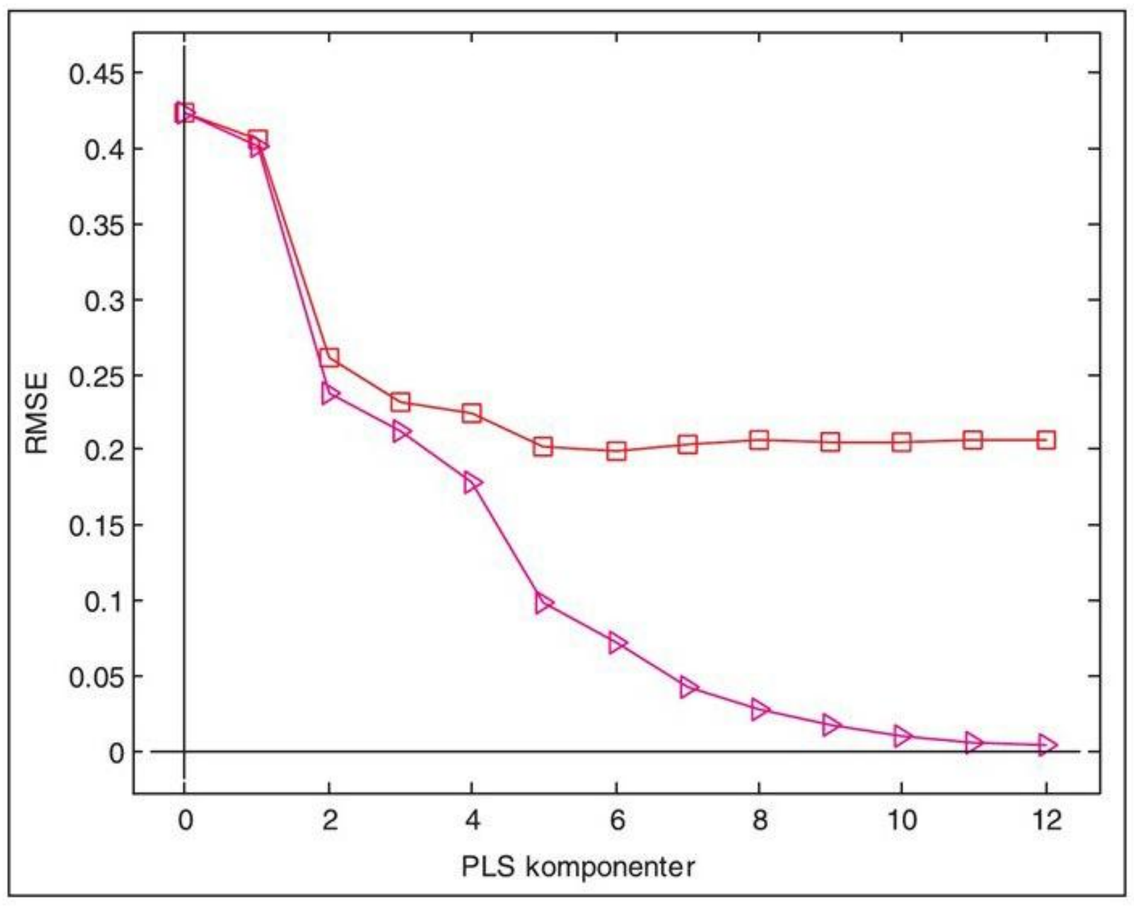
Hvorfor ikke 1, 2, 3?

Umiddelbart kan man fristes til kun at lave én dummyvektor bestående af ettaller for gruppe A, totaller for gruppe B og tretaller for gruppe C (figur 1). Men den går ikke. En sådan dummyvektor antager på forhånd en usandsynlig kvantitativ forskel mellem grupperne; f.eks. at C-prøvernes signaler skulle være tre gange så intense som A-prøvernes.

Eksempel

Vi anvender nu PLS-DA på det samme prøvesæt, som SIMCA blev anvendt på i seneste klumme. Sættet består af 57 prøver fra fire forskellige sukkerfabrikker; der er 14 prøver fra fabrik H, 12 fra fabrik I, 13 fra fabrik K og 13 fra fabrik M; derudover er der fem ukendte (U1-U5) prøver i sættet. Prøveforberedelsen består i at opløse 2,25 g sukker i 15,0 mL ionbyttet vand, hvorefter der måles fluorescens-emissionsspektre ved excitationsbølgelængden 240 nm med et LS50B-instrument fra PerkinElmer.

I dette eksempel vil vi fokusere på at adskille fabrik I-prøverne fra resten af prøverne; dvs. vi kan nøjes med en Y-vektor i ▶



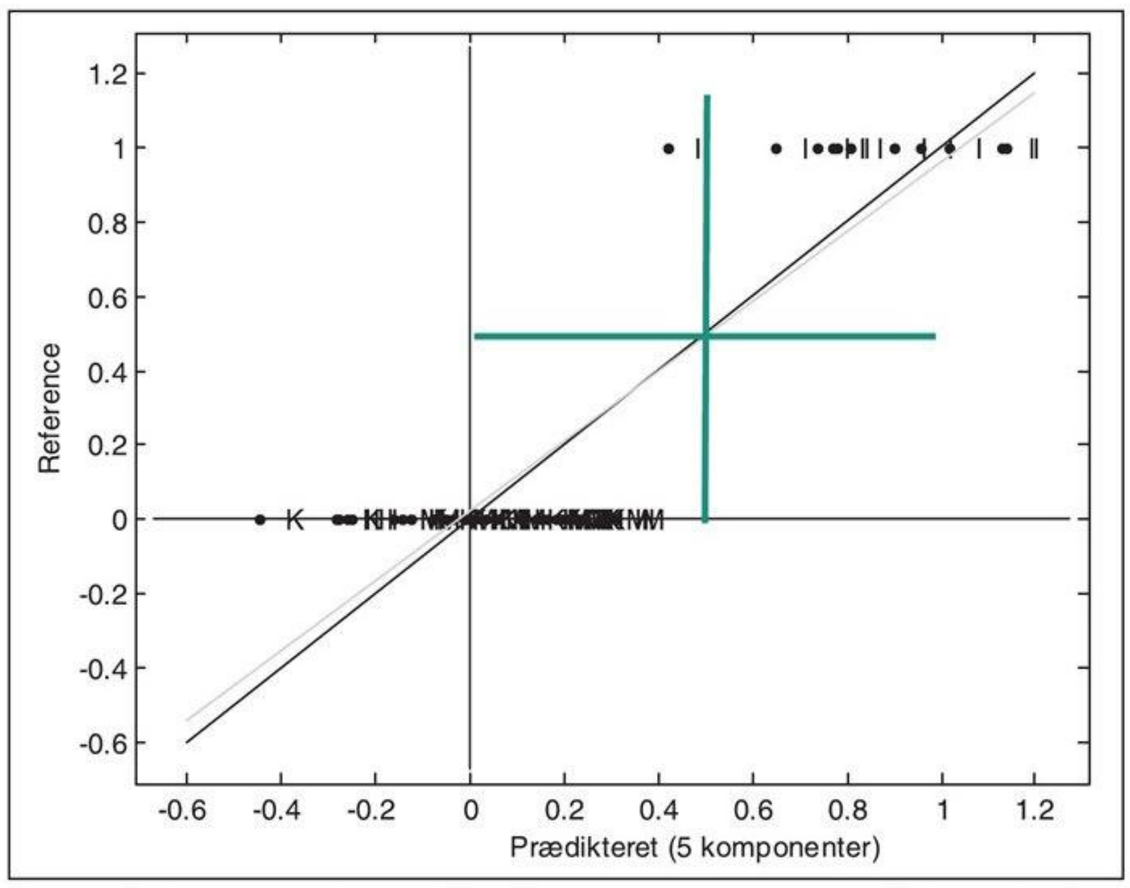
Figur 2. Kalibreret (trekant) og krydsvalideret (firkant) Root Mean Square Error for Y-vektoren (der indeholder nuller og ettaller). Det vurderes, at det er fornuftigt at anvende fem komponenter.

Det Kemometriske Rum

stedet for en Y-matrix. Vektoren indeholder et ettal, hvis prøven er fra fabrik I og ellers nul.

Der beregnes en PLS-model med krydsvalidering (10 segmenter), og RMSE-plottet inspiceres (figur 2). Det fremgår, at 5-6 komponenter er passende. Dernæst ser vi på reference (0/1) mod prædikteret plot (figur 3) for fem komponenter. Det kan vise os kvaliteten af modellen, nemlig antallet af fejlklassificerede prøver. Hvis den prædikterede værdi er over 0,5 antages prøven at tilhøre fabrik I. Denne lidt grove klassifikationsbestemmelse kan forfines på forskellig vis. F.eks. er det muligt, at modellen *klassificerer* bedre ved et andet antal komponenter, end der hvor nul og et prædikteres optimalt. Derfor bør man også teste andre antal komponenter. Det er vigtigt at understrege, at der også skal outlier-testes, når der prædikteres med PLS-DA-modeller. F.eks. vil en prædikteret værdi på 4000 ikke være et sandsynligt resultat fra en prøve, der ligner kalibreringsprøverne.

Konklusionen baseret på figur 3 er, at modellen ser ud til at adskille de to grupper: fabrik I eller ikke-fabrik I. Kun en enkelt I-prøve fejlklassificeres. Og faktisk kan man rykke beslutningsgrænsen til lidt under 0,5. Derved opnås perfekt klassifikation.



Figur 3. Reference mod Prædikteret for en femkomponent PLS-DA-model for fabrik I. Krydset angiver værdien 0,5 på både abscisse- og ordinatakserne.

Prædiktion af ukendte prøver

Den udviklede PLS-model anvendes til prædiktion af de fem ukendte prøver. De prædikterede Y-værdier baseret på en femkomponent-model er:

U1: 0,14
U2: 0,67
U3: -0,17
U4: 0,91
U5: 0,01

De krydsvaliderede prædiktioner (af kalibreringsprøverne) svinger for I-prøver mellem cirka 0,4 og 1,15. Heraf fremgår det, at prøverne U2 og U4 må antages at være I-prøver.

Outro

Antallet af komponenter i PLS-DA-modeller skal estimeres ud fra f.eks. krydsvalidering eller testsæt-validering, præcis som det gælder for standard PLS-modeller. Men man bør fokusere på klassifikation frem for prædiktion, når man beslutter antal komponenter. Man bør altid vurdere, om modellen rent faktisk kan adskille eller delvist adskille grupperne; dvs. estimere værdien nul eller et i dummymatricen.

I eksemplet har vi fokuseret på fabrik I eller ikke-fabrik I. Man kan tilsvarende lave modeller for H mod resten, K mod resten, og M mod resten. Eller der kan laves en samlet model for alle fire grupper med en dummymatrix bestående af fire søjler; det optimale antal komponenter kan være forskelligt for hver Y-søjle. Endelig kan man vælge at lave parvise modeller; f.eks. fabrik I mod fabrik K, fabrik I mod fabrik M etc.

PLS-DA fokuserer på enhver variation i data, som kan klassificere. Derfor er det nødvendigt at være ekstra opmærksom på validering, når man arbejder med PLS-DA-modeller. Dette emne vil blive taget op i en senere klumme.

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.

Løsning til kemikryds 9

Det hemmelige ord som Carsten Krogh havde gemt i kemikryds nr. 9 var: september

2009-9	1	2	3	4	5	6	7	8	9	10	11	12
1	V	Æ	G	T	E	N	→	Ø	G	E	S	↓
2	C	O	3	I	13\4	E	A	5	S	P	A	M
	14\6	N	15	16\7	A	O	F	17	8	O	Z	E
9	N	E	W	T	O	N	S	↓	10	X	18\11	D
12	O	R	F	E	R	R	I	T	13	Y	P	↓
14	F	19\15	6	20\16	T	E	17	R	18	L	E	O
	21\19	4	22\20	I	A	K	23\21	E	24\22	A	R	P
23	M	25\24	S	B	26\25	L	O	D	S	K	U	D
26	E	U	M	E	L	A	N	I	N	27	28\27	R
28	D	29	2	N	30	M	D	E	29\31	B	M	I
32	I	30\33	O	H	31\34	E	32\35	↓	N	A	33\36	F
37	C	H	3	O	H	34\38	K	L	O	R	A	T
39	I	2	40	L	35\41	H	2	O	2	36\42	S	E
43	N	O	44	T	A	T	O	V	Ø	R	45	N