Jack-knifing – "cut the crap"

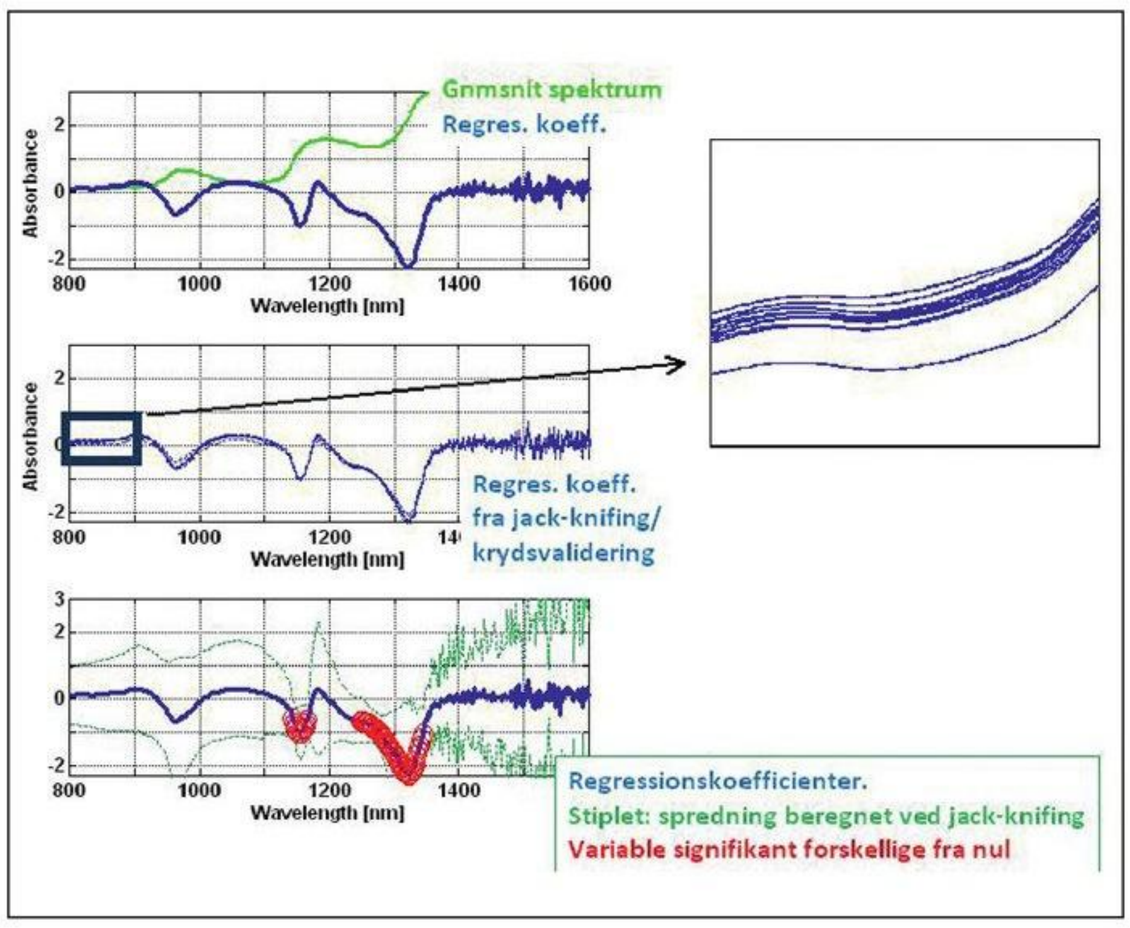
Jack-knifing er en generel statistisk metode, som kan bruges til at beregne usikkerheder. Den kan f.eks. bruges til at fjerne støjfyldte variable

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

I de seneste klummer har vi arbejdet med at udvælge variable. En typisk variabel selektionsmetode fungerer ved at man finder et sæt af variable, som giver gode prædiktioner. I denne klumme beskrives en lidt anderledes metode, som mere ser på, hvordan man kan fjerne variable med store usikkerheder og som derfor ikke bidrager konstruktivt til den multivariate model.

Jack-knifing

Jack-knifing er en generel metode til at beregne usikkerheder af parametre [1]. Navnet stammer fra John Tukey, der ville angive, at metoden var bredt anvendelig uden dog at være prangende god i alle situationer (som en foldekniv, der kan tjene mange formål). Metoden er velegnet, når man ikke har en klassisk statistisk metode til direkte at beregne usikkerheder. Det er netop tilfældet i de fleste multivariate regressionsmodeller såsom PLS. Der findes f.eks. ikke en almen accepteret metode til at beregne usikkerheder på regressionskoefficienterne, men dette kan man gøre ved hjælp af jack-knifing.



Figur 1.
Øverst: Regressionsvektor fra PLS-model (blå) og gennemsnitsspektrum (grøn).
Midt: Jack-knife-beregnede regressionsvektorer fra en ti-segmenteret krydsvalidering (altså 10 forskellige regressionsvektorer).
Nederst: Konfidensintervaller beregnet ved jack-knifing er vist sammen med regressionvektoren. Koefficienter, der er signifikant forskellige fra nul, er markeret med røde cirkler.

Metoden fungerer som følger. Når man krydsvaliderer en PLS-model, betyder det implicit, at man beregner PLS-modeller på en række forskellige datasæt. Laver man f.eks. fuld krydsvalidering, hvor en prøve udelades ad gangen, så vil man først fjerne prøve et og beregne en PLS-model på de resterende prøver. Dernæst fjerner man prøve to, tilbagelægger prøve et, og beregner en PLS-model og så fremdeles (Dansk Kemi 3, 2009). Hvis man f.eks. har tyve prøver, har man således efter endt krydsvalidering beregnet tyve forskellige PLS-modeller. Normalt anvendes disse kun til at beregne en RMSECV-værdi, hvorefter modellerne "smides ud". Men i jack-knifing anvender man de tyve modeller til at beregne usikkerheder af bl.a. regressionskoefficienterne. I figur 1 kan man se beregnede jack-knifing-resultater for en PLS-model af øldata (præsenteret første gang i Dansk Kemi 8, 2008 og anvendt i de seneste klummer).

Figuren er en smule forvirret, men vi gennemgår den del for del. Øverste plot viser et gennemsnitsspektrum (grønt) og regressionsvektoren, som man får med normal PLS. Gennemsnitsspektret er blot vist for at kunne orientere sig. Normalt antager man, at absolut store regressionskoefficienter angiver vigtige variable, men nu vil vi gerne beregne usikkerheden på hver koefficient ved at se på resultatet af jack-knifing. Det bemærkes, at de laveste og højeste bølgelængder er udeladt af plottet for at fokusere på den vigtige del af spektret (Dansk Kemi 3, 2010).

Det midterste plot viser de forskellige regressionsvektorer, man får, når man laver en ti-segmenteret krydsvalidering. For hvert segment får man et bud på regressionsvektoren, og det er disse ti bud, man anvender i jack-knifing. Ud fra disse kan man beregne standardafvigelsen på PLS-regressionsvektoren [1].

Det nederste plot viser regressionsvektoren baseret på alle prøver og usikkerheden for hver koefficient beregnet vha. jack-knifing. Hvis de grønne kurver for en variabel ligger både over og under nul, så er den pågældende koefficient så usikker, at den formentlig lige så vel kunne være nul. De koefficienter, hvor usikkerheden inkluderer nul, udelades tentativt, for at se om modellen forbedres.

Som det kan ses, er der herefter kun ganske få variable tilbage, og vi har faktisk fået udeladt alle irrelevante (ift. ekstrakt i øl) variable. Anvendes en PLS-model baseret på de valgte (røde) variable fås en prædiktionsfejl på testsættet på 0,20.

Det skal bemærkes, at man udmærket kan anvende jack-knifing iterativt ved at genberegne modellen og eliminere yderligere variable. I det konkrete tilfælde bliver modellen ikke

bedre, men ofte kan man nå frem til den bedste model ved at fjerne dårlige variable lidt ad gangen og holde skyldigt øje med outliere undervejs.

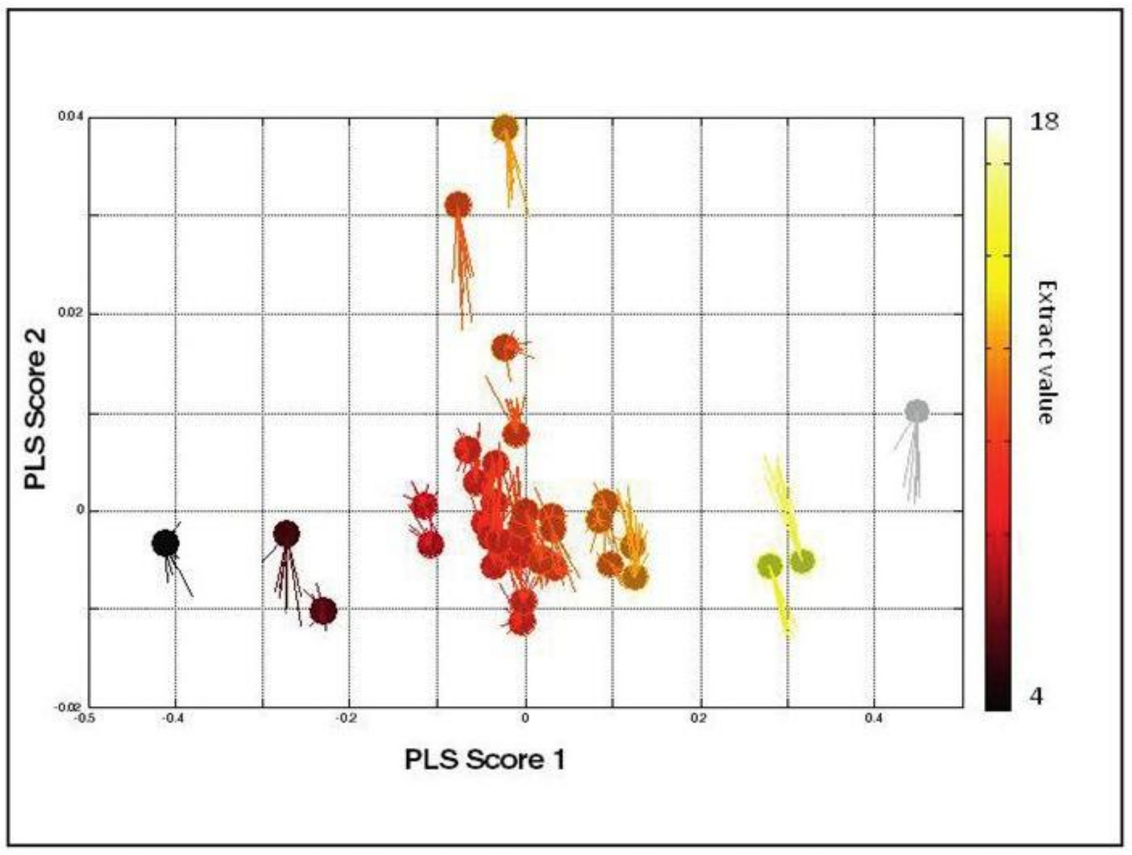
Jack-knifing adskiller sig på to markante områder fra mange andre former for variabelselektion. For det første er de valgte variable baseret på en model af hele datasættet. Jack-knifing kan derfor kun anvendes i en situation, hvor man kan lave en valid model på hele datasættet. Viser krydsvalidering, at modellen på hele datasættet slet ikke kan prædiktere, så betyder det direkte, at de beregnede parametre såsom regressionskoefficienter ikke giver mening, og dermed at jack-knifing heller ikke giver mening. Et andet vigtigt aspekt ved jack-knifing er at man ikke direkte søger efter variable, der prædikterer godt. Man fjerner simpelthen blot variable, som muligvis har en regressionskoefficient på nul. Dvs. at man fjerner variable, der har koefficienter tæt på nul eller har så stor usikkerhed, at det ikke kan udelukkes, at de er nul. Ideen bag dette, er at regressionskoefficienter, der er nul, ikke bidrager til prædiktionen. Uanset, hvad den pågældende variabel er, så bliver den multipliceret med nul og indgår således ikke. At fjerne sådanne variable gør, at risikoen for overfit er langt mindre end ved f.eks. forward selection (Dansk Kemi 6/7, 2010).

Den eksplorative del

Jack-knifing kan bruges til andet og mere end at beregne usikkerheder. Resultatet fra jack-knifing kan visualiseres på et hav af måder, som giver mulighed for at forstå data og detektere outliere [2,3].

Når vi i ovenstående eksempel krydsvaliderede vores PLS-model i ti segmenter, så fik vi, ud over de almindelige scores for hver prøve, også scores for enhver prøve i de ni tilfælde, hvor prøven ikke er med i det segment, der udelades. Disse scores kan plottes sammen med de "rigtige" scores, for at se om nogle prøver påvirker modellen markant.

I det konkrete tilfælde (figur 2) er der ikke nogle markante variationer, bortset fra at de prøver, der er forholdsvis ekstreme, forventeligt har større usikkerhed end de, der er mere normale.



Figur 2. Scoreplot fra PLS-model. Scores er farvet efter ekstrakt-værdi. For hver prøve er der også vist scores som fundet under krydsvalideringen.

Outro

Vi garanterer, at der ikke kommer mere om variabelselektion foreløbig. Den eneste ting, vi ønsker at nævne her, er, at der findes en relateret metode, der kaldes bootstrapping. Det er et interessant, men mere beregningstungt alternativ til jack-knifing [1], som vi dog vil gemme til en anden god gang.

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

Referencer
1. B. Efron & G. Gong. A leisurely look at the Bootstrap, the Jackknife, and the cross-validation. American Statistician, the 37:36-48, 1983.
2. H. Martens & M. Martens. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). Food Quality and Preference 11 (1-2):5-16, 2000.
3. J. Riu & R. Bro. Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. Chemometrics and Intelligent Laboratory Systems. 65 (1):35-49, 2003.

Løsning til kemikryds 6/7

Det hemmelige ord som Carsten Krogh havde gemt i kemikryds nr. 6/7 var:
Sommer og sol

2010-6/7	1		2	3		4	5	6	7	8
1	A	9\2	C	S	10\3	G	A	R	Y	↓
4	L	B	O	11\5	O	H	P	A	C	G
6	B	I	2	S	3	12\7	M	O	2	R
8	E	S	13\9	T	14\10	N	Ø	D	11	A
12	R	P	O	E	M	15\13	L	16	17\14	N
15	T	H	18\16	E	A	P	L	A	S	T
17	→	E	I	N	S	T	E	I	N	19
	20\18	N	S	21\19	L	22\20	R	S	C	N
21	B	O	N	M	O	T	23	24\22	L	I
23	U	L	I	25\24	W	O	C	L	4	26
25	L	I	N	B	27\26	F	28\27	I	29\28	C
29	B	A	G	E	P	U	L	V	E	R