Krydsvalidering

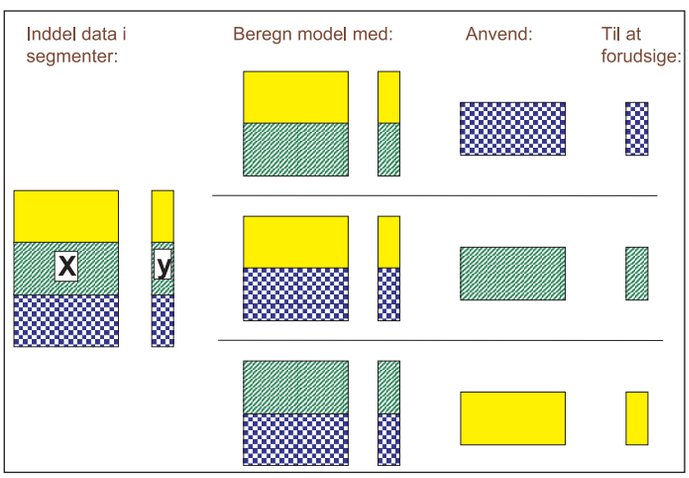
I multivariat regressionsmodellering er det ofte muligt at få vilkårligt gode sammenhænge mellem f.eks. spektroskopiske målinger og referencemålinger. Dette kaldes overtilpasning, og krydsvalidering er en enkel måde at validere en regressionsmodel på, således at overtilpasning til data undgås

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

En anke der ofte høres mod kemometriske metoder er, at "de kan jo få alt til at hænge sammen, bare man fifler nok med data". Udsagnet er helt rigtigt og gælder i princippet al modellering. I kemometrien er man stærkt optaget af at undgå overtilpasninger til data (kaldet "overfit" på engelsk), og der er derfor udviklet forskellige metoder til at validere regressionsmodeller. En meget udbredt metode, som findes implementeret i ethvert program til kemometrisk dataanalyse, er krydsvalidering.

Krydsvalidering

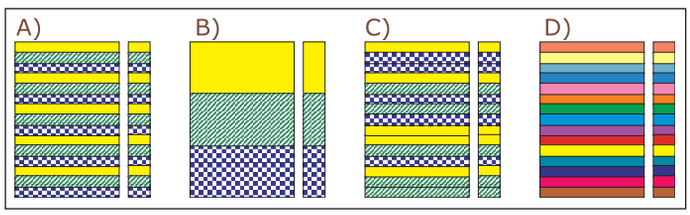
Krydsvalidering er baseret på sekventiel udeladelse af én eller flere prøver, indtil alle prøver har været udeladt netop én gang. Princippet er at estimere modellen baseret på de ikke-udeladte prøver og dernæst anvende den beregnede model på de udeladte data. PLS-prædiktionsfejlen estimeres for de udeladte prøver for 1- til N-komponenter, og ifølge teorien vil en afbildning af prædiktionsfejl mod antal PLS-komponenter have et minimum ved det optimale antal komponenter. Ved dette minimum har man en tilpas god beskrivelse af kendte data, samtidig med at modellen kan anvendes til at estimere nye data på en rimelig måde; sidstnævnte egenskab kaldes også modellens generaliseringsevne.



Figur 1. Princippet i krydsvalidering.
Figuren er gengivet med tilladelse fra Eigenvector Research Inc. (www.eigenvector.com).

Princippet i krydsvalidering er (også skematisk vist i figur 1):

1. Udelad én gruppe af prøver (kan være én prøve)
2. Beregn regressionsmodel med op til N PLS-komponenter baseret på de resterende prøver
3. Prædiktér de udeladte prøver med den beregnede model fra 1- til N-komponenter
4. Udelad den næste gruppe af prøver og gentag punkt 2 og 3.
5. Beregn samlet prædiktionsfejl for 1- til N-komponenter ved at samle alle prædiktioner og sammenligne disse med referenceværdierne



Figur 2. Udvælgelse af krydsvalideringssegmenter.
A) Systematisk udvælgelse af hver n'te prøve også kaldet Venetian blinds (dansk: persienner),
B) sammenhængende blokke, f.eks godt til tidsserier, individuelle batches eller a priori opdeling af prøverne i grupper,
C) tilfældig udvælgelse, gentages ofte flere gange,
D) udelad én prøve ad gangen, anvendes, når der kun er få prøver i datasættet.
Figuren er gengivet med tilladelse fra Eigenvector Research Inc. (www.eigenvector.com).

Udvælgelsen af prøver kan ske efter forskellige principper som angivet i figur 2. I de fleste kemometri-programmer er det muligt at vælge alle de viste skemaer som krydsvalideringsmetode. Hvis målingerne eksempelvis stammer fra forskellige batche, kan man ved at vælge krydsvalideringsmetode udelade prøverne batchvis, og således verificere at modellen kan prædiktere prøver fra nye batche. Bemærk at replikater altid bør holdes ude i samme segment (kun undtaget i helt specielle situationer). Forudsætningen for at anvende krydsvalidering er, at prøverne kan betragtes som uafhængige. Det er f.eks. ikke tilfældet for naboprøver i en tidsserie, og derfor anvendes metoden, hvor sammenhængende blokke udelades.

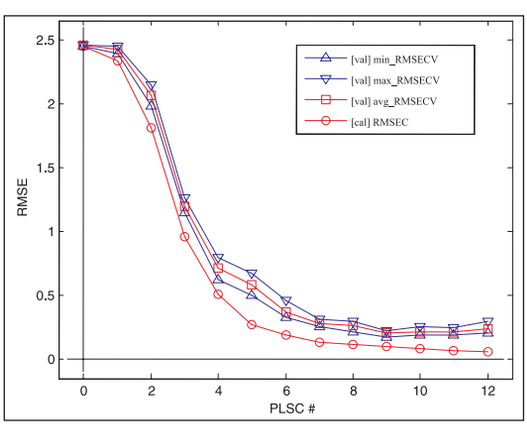
Øldata

Fyrre prøver bestående af forskellige øl er analyseret for ekstraktindhold i % Plato med en laboratoriemetode. Ekstrakt er en vigtig kvalitetsparameter i bryggeriindustrien og indikerer gærens potentiale til at danne alkohol. Ekstraktprocenten varierer fra 4,2-18,8% Plato. De samme fyrre prøver er ligeledes målt med visuel- og nærinfrarød spektroskopi i området 400 nm til 1400 nm med to nanometers interval, dvs. 501 spektrale variable. Prøverne er afgasset inden transmissionsmåling på et NIRSystems 6500 spektrofotometer i en 30 mm kuvette. Spektrofotometeret anvender et delt detektorsystem med siliciumbaseret detektor i området 400 nm til 1100 nm og blysulfid detektor (PbS) i området 1100 nm til 2500 nm.

PLS med krydsvalidering

En PLS-model med NIR-spektrene som **X** og ekstrakt som **y** beregnes. Der vælges krydsvalidering med tilfældigt udvalgte prøver i 10 segmenter; dvs. 10% af prøverne udelades ad gangen. Hele proceduren gentages 20 gange for at få et bud på usikkerheden på den gennemsnitlige prædiktionsfejl, også kaldet RMSE (Root Mean Square Error). Beregningerne er foretaget i LatentiX 2.00 (www.latentix.com).

I figur 3 ses RMSE for krydsvalideringen, RMSECV, med maksimums- og minimumsværdier angivet (for hver af de 20 gange proceduren gentages) som funktion af antal komponenter. Derudover er tilpasningen til data også vist (RMSEC for kalibrering); denne vil altid falde med stigende antal komponenter og kan derfor ikke anvendes til at bestemme det optimale antal komponenter. Af figur 3 ses, at ni PLS-komponenter er det optimale for de analyserede data; en tommelfingerregel er, at man bør have mindst fem rimeligt repræsentative prøver pr. komponent, man anvender i modellen. I dette tilfælde er dette rimeligt opfyldt og ni komponenter anvendes i den færdige model. Af figuren ser vi at en PLS-model med ni komponenter kan forventes at give en fejl i prædiktionen på cirka 0,2% Plato.



Figur 3. Prædiktionsfejl, udtrykt som RMSE (Root Mean Square Error), som funktion af antal PLS-komponenter. Gennemsnitlig prædiktionsfejl (rød med firkanter) med minimum ved ni PLS-komponenter, maksimum prædiktionsfejl i de tyve runder (blå med trekant med spids nedad), minimum prædiktionsfejl i de tyve runder (blå med trekant med spids opad), og modeltilpasningsfejl (rød med cirkler). Sidstnævnte vil fortsætte med at falde med stigende antal komponenter.

Outro

Det er afgørende at validere regressionsmodeller, og det første spørgsmål man bør stille, når man præsenteres for en sådan er: "hvordan er modellen valideret?". Er svaret, at det er den ikke endnu, kan man stoppe samtalen der og vente på, at den bliver det. Krydsvalidering har en svag tendens til at overtilpasse (dvs. at foreslå et lidt højere antal komponenter end det reelt optimale), men det er ikke noget, der kan ødelægge den generelle validitet af krydsvalidering som metode. Skal modellen anvendes i f.eks. kommercielle sammenhænge skal den valideres yderligere ved at undersøge modellens performance på nye ukendte data: et uafhængigt testsæt.

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