Klassifikation af spektrale data med Extended Canonical Variates Analysis

Den statistiske metode til klassifikation, Canonical Variates Analysis, kræver uafhængige variable og kan derfor ikke anvendes direkte på f.eks. multivariate spektroskopiske data. Men en enkel modifikation af CVA-metoden gør dette muligt

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

I vores introduktion af klassifikationsmetoden Canonical Variates Analysis (CVA) (Dansk Kemi 11, 2010) blev det beskrevet, at CVA ikke kan håndtere datasæt med flere variable end prøver og ej heller variable, der er stærk korrelerede. Vi vil her beskrive en ændring af CVA-metoden, så den kan håndtere "ægte" multivariate data. Metoden kalder vi Extended Canonical Variates Analysis (ECVA).

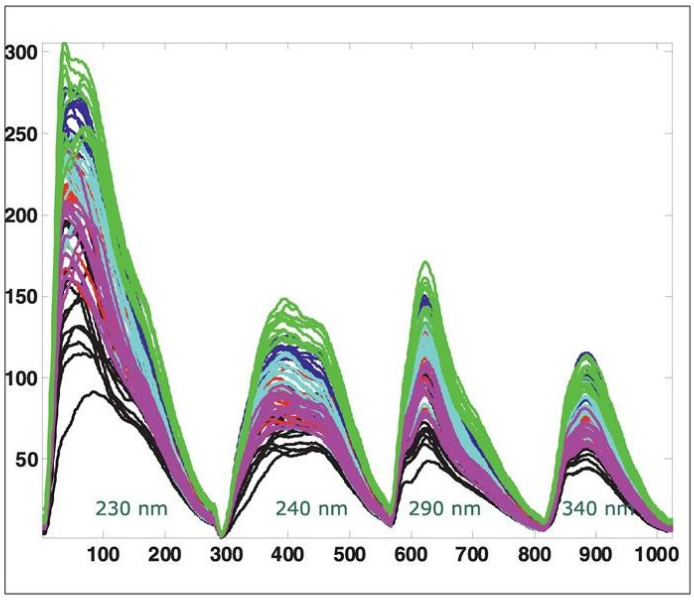
Find de gode retninger

I forrige klumme så vi, at ligningen

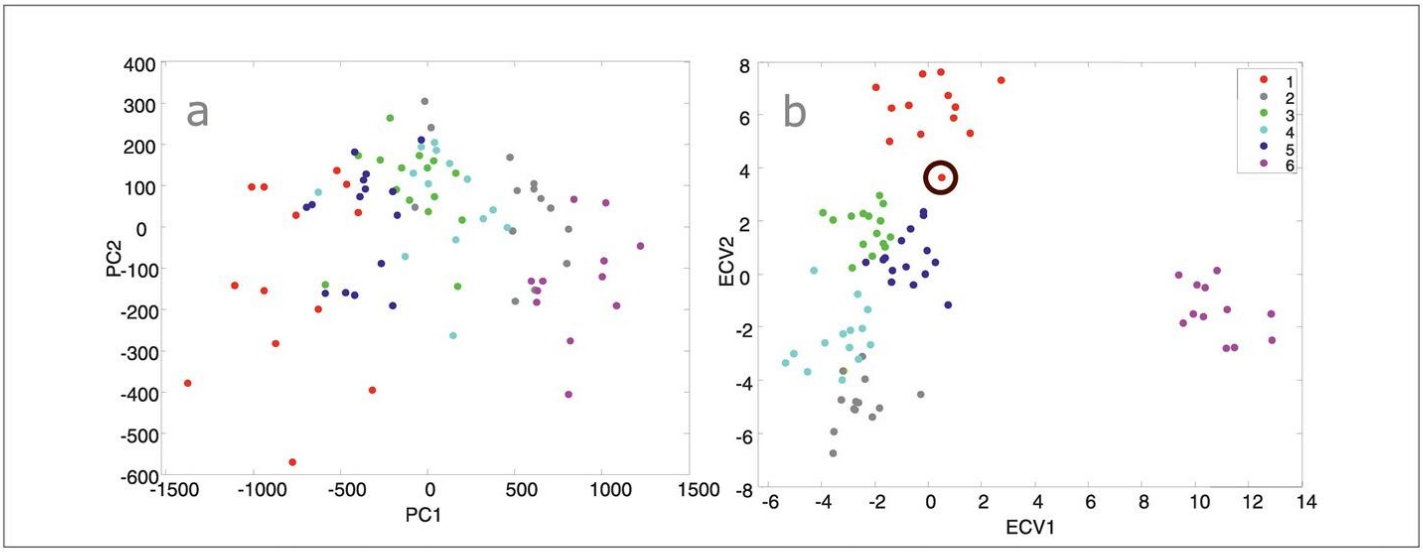
$$\mathbf{S}_{mellem\ grupper}\mathbf{W} = \lambda \mathbf{S}_{inden\ for\ grupper}\mathbf{W}$$

definerer en retning **w,** som er egnet til at bruge til klassifikation. En løsning af ligningen med hensyn til **w** vil give en retning, hvor prøver fra forskellige klasser er langt fra hinanden og samtidig veldefinerede.

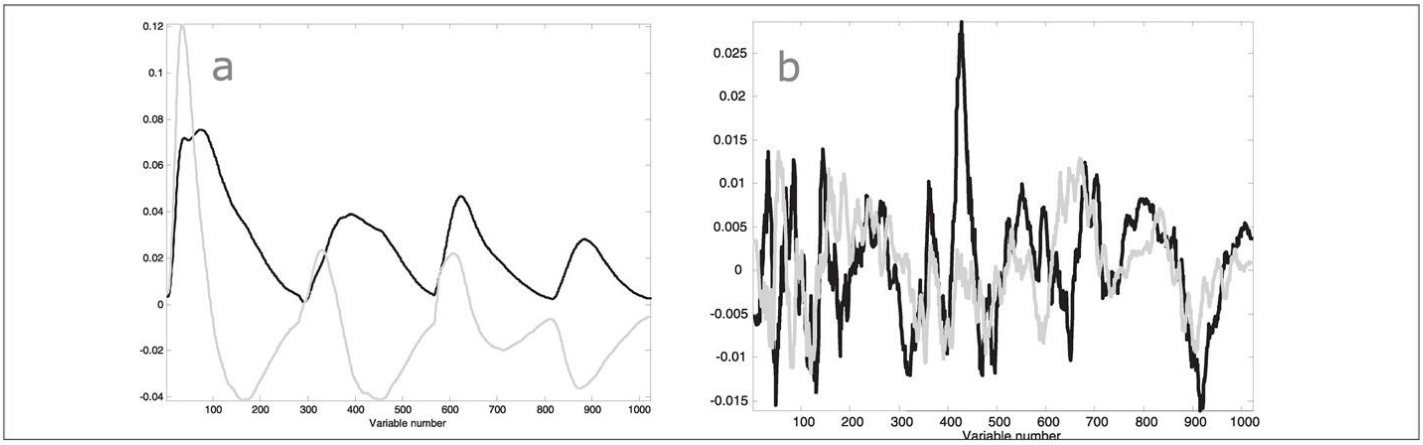
Desværre kan ovenstående ligning ikke umiddelbart løses for spektroskopiske data, da man skal beregne $\mathbf{S}^{-1}_{inden\ for\ grupper}$ for at løse ligningen. Denne kan ikke estimeres for kolineære data. Hvis der er analyseret f.eks. 83 prøver med fluorescensspektroskopi (1023 variable) vil dimensionen på $\mathbf{S}_{inden\ for\ grupper}$ være 1023 × 1023, og den inverse matrix kan ikke umiddelbart findes, da rangen maksimalt er 83.



Figur 1. Fluorescens-emissionsspektre af 83 sukkerprøver. Excitationsbølgelængderne er 230 nm, 240 nm, 290 nm og 340 nm. Farverne angiver tilhørsforhold til hver af de seks fabrikker, der har leveret prøver.



Figur 2. a) PCA-score 1 versus 2 fra en PCA på datamatricen, b) ECVA-score 1 versus 2 fra en ECVA på den samme datamatrice. Prøver markeret med en cirkel fejlklassificeres i ECVA-modellen.



Figur 3. a) PCA-loading 1 (sort) og 2 (grå) fra en PCA på datamatricen, b) ECVA-loading 1 (sort) og 2 (grå) fra en ECVA på datamatricen.

Hvis vi kun betragter to grupper, kan det matematisk vises [2], at ovenstående ligning kan skrives som

$$(\bar{x}_1 - \bar{x}_2)k = \lambda \mathbf{S}_{inden\ for\ grupper}\mathbf{W}$$

hvor $\bar{x}_1$ og $\bar{x}_2$ er hhv. gennemsnittet af alle spektre i klasse 1 og klasse 2.

Dette udtryk kan omskrives til en multivariat regressionsmodel ved sætte $\mathbf{y} = (\bar{x}_1 - \bar{x}_2)k/\lambda$ og $\mathbf{X} = \mathbf{S}_{inden\ for\ grupper}$ og problemet, der skal løses, kan herefter præsenteres som at finde en regressionsvektor **w**, der giver den lavest mulige fejl **f** i udtrykket

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{f}$$

Analogien til PLS (Dansk Kemi 11, 2008) er slående, og vi kan benytte PLS til at bestemme regressionsvektoren **b** = **w**, som er den retning, der bedst adskiller de to klasser i det oprindelige multivariate rum. PLS optimerer prædiktionen af **y**, og kernefunktionen i ECVA er dermed ændret en smule ift. udgangspunktet CVA. Problemstillingen kan generaliseres til flere grupper som beskrevet i [1].

Eksempel – klassifikation af sukkerprøver

Relevansen af og baggrunden for at analysere sukkerprøver opløst i vand er tidligere beskrevet (Dansk Kemi 9, 2009). Vi vil analysere et datasæt bestående af 83 prøver fra seks forskellige sukkerfabrikker (1-6). Prøveforberedelsen består i at opløse 2,25 gram sukker i 15,0 mL ionbyttet vand. Fluorescens-emissionsspektre med i alt 1023 spektrale variable optages på opløsningen ved fire forskellige excitationsbølgelængder (230, 240, 290, 340 nm) med et LS50B instrument fra Perkin-Elmer (figur 1). Det betyder, at datamatricen **X**, der skal analyseres med ECVA, har dimensionen 83 × 1023.

Der beregnes nu en PCA og en ECVA på datasættet. Det fremgår tydeligt ved sammenligning af figur 2a og b, at ECVA udfører det job, som den er udviklet til at lave; nemlig at diskriminere mest muligt mellem spektrene fra de seks fabrikker. PCA udnytter ikke information om fabrikkerne, men finder de retninger, der beskriver mest muligt af totalvariationen i datasættet. Som det fremgår, er PCA-retningerne ikke specielt relevante ift. diskrimination af fabrikkerne.

Hvis man er interesseret i at tolke de spektrale data og få information om hvilke områder i spektrene, der bidrager til diskriminationen mellem fabrikkerne, kan man undersøge loading-vektorerne (figur 3a og b). Det ses, at der er store forskelle mellem PCA-loadings og ECVA-loadings; sidstnævnte er mere støjfyldte, da de retninger, der adskiller fabrikkerne, ikke beskriver den største variation i datasættet. Til gengæld er det muligt at finde de områder i spektrene, hvor den diskriminative information findes. I det givne eksempel er det f.eks. primært variablene omkring variabel 425 (emissioner for ex. 240 nm), der bidrager til adskillelsen af fabrik 6 fra de øvrige fabrikker (ECVA loading 1).

Selve klassifikationen

Én ting er tolkningen af data én anden er selve klassifikationsfejlen. Hovedformålet med at udvikle modellen er at være i stand til at klassificere spektre af ukendte prøver ift. de seks fabrikker. Hvis man udfører en segmenteret krydsvalidering (fem segmenter), begås der kun én fejlklassifikation; denne er markeret i ECVA-plottet i figur 2 med en cirkel.

Outro

Med en lille modifikation har vi udviklet CVA til at kunne håndtere multivariate kolineære data, uden at der skal udføres en pre-kompression af data med f.eks. PCA med fare for at miste relevant information i selve klassifikationen. ECVA har derfor et stort potentiale inden for spektroskopisk klassifikationsanalyse som f.eks. autenticitet og metabolomics. Man skal imidlertid være opmærksom på, at der ikke nødvendigvis er stor forskel på selve klassifikationsfejlen ved anvendelse af forskellige metoder som SIMCA, PLS-DA, PCA-CVA og ECVA, der hver har deres fordele og ulemper.

Referencer

1. Nørgaard L, Bro R, Westad F, Engelsen SB, A modification of canonical variates analysis to handle highly collinear multivariate data, Journal of Chemometrics 2006; 20: 425–435.
2. Duda RO, Hart PE, Stork DG. Pattern Classification (2nd edn). John Wiley & Sons: New York, 2001.

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