Når diskrimination er godt

Principal Component Analysis anvendes ofte som første skridt i den eksplorative multivariate analyse. Den tilsvarende statistiske metode til klassifikationsproblemer hedder Canonical Variates Analysis

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

ADVARSEL: Denne klumme er en smule tør; men fortvivl ikke: den efterfølgende klumme vil illustrere principperne på en relevant kemisk problemstilling.

Klassifikationsmetoderne SIMCA og PLS-DA (Dansk Kemi 8-10, 2009) er begge udviklet i en kemometrisk kontekst. I den klassiske statistik er der imidlertid også udviklet en anden og særdeles effektiv metode til diskrimination mellem klasser, Canonical Variates Analysis (CVA) [1,2]. Den vil vi se nærmere på her.

Hvad er CVA?

Som illustreret i figur 1 finder PCA i første komponent den retning - loadingvektor - **p**, der for det samlede datasæt maksimerer variansen; dette svarer til at den kvadrerede vinkelrette afstand for hvert enkelt punkt ind til linjen angivet ved **p**-retningen er mindst mulig set over alle prøver; altså at linjen ligger så tæt på alle punkter som muligt. Idéen bag CVA er derimod at finde den retning, **w**, for hvilket udtrykket

$$\frac{\text{"mellem gruppe spredning"}}{\text{"inden for gruppe spredning"}}$$

bliver maksimalt. Dvs., at CVA søger en retning hvor grupperne ligger langt fra hinanden svarende til høj "mellem gruppe spredning", og hvor grupperne samtidig er meget veldefinerede svarende til lav "inden for gruppe spredning". CVA har således direkte fokus på at skelne mellem kendte grupper.

Matematisk beskrivelse

Ovenstående kvalitative udtryk kan opskrives matematisk som at maksimere funktionen *J*, der er defineret ved

$$J(\mathbf{w}) = \frac{\mathbf{w}'\mathbf{S}_{mellemgrupper}\mathbf{w}}{\mathbf{w}'\mathbf{S}_{indenforgrupper}\mathbf{w}}$$

Antag at vi har en datamatrix **X** ($n \times v$) med målinger, hvor prøverne kommer fra g forskellige grupper med

n_i prøver i den *i*'te gruppe ($n = \sum_{i=1}^{g} n_i$) og *v* variable.

S-udtrykkene i ovenstående ligning kaldes scatter- eller spredningsmatricer og defineres som følger:

$$\mathbf{S}_{mellem\ grupper} = \frac{1}{(g-1)}\sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$$

$$\mathbf{S}_{inden\ for\ grupper} = \frac{1}{(n-g)}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$$

hvor *n* er total antal prøver, n_i antal prøver i den *i*'te gruppe og *g* er antal grupper. $\mathbf{x}_{ij}$ er f.eks. det *j*'te målte NIR-spektrum i gruppe *i*, $\bar{\mathbf{x}}_i$ er gennemsnitsspektret for den *i*'te gruppe, og $\bar{\mathbf{x}}$ er gennemsnitsspektret beregnet på alle prøver. Matrixdimensionerne for **S**-udtrykkene er antal variable × antal variable ($v \times v$).

Udtrykket $J(\mathbf{w})$ kan omskrives til et generaliseret egenværdiudtryk (ikke åbenlyst at det er således, men det kan bevises)

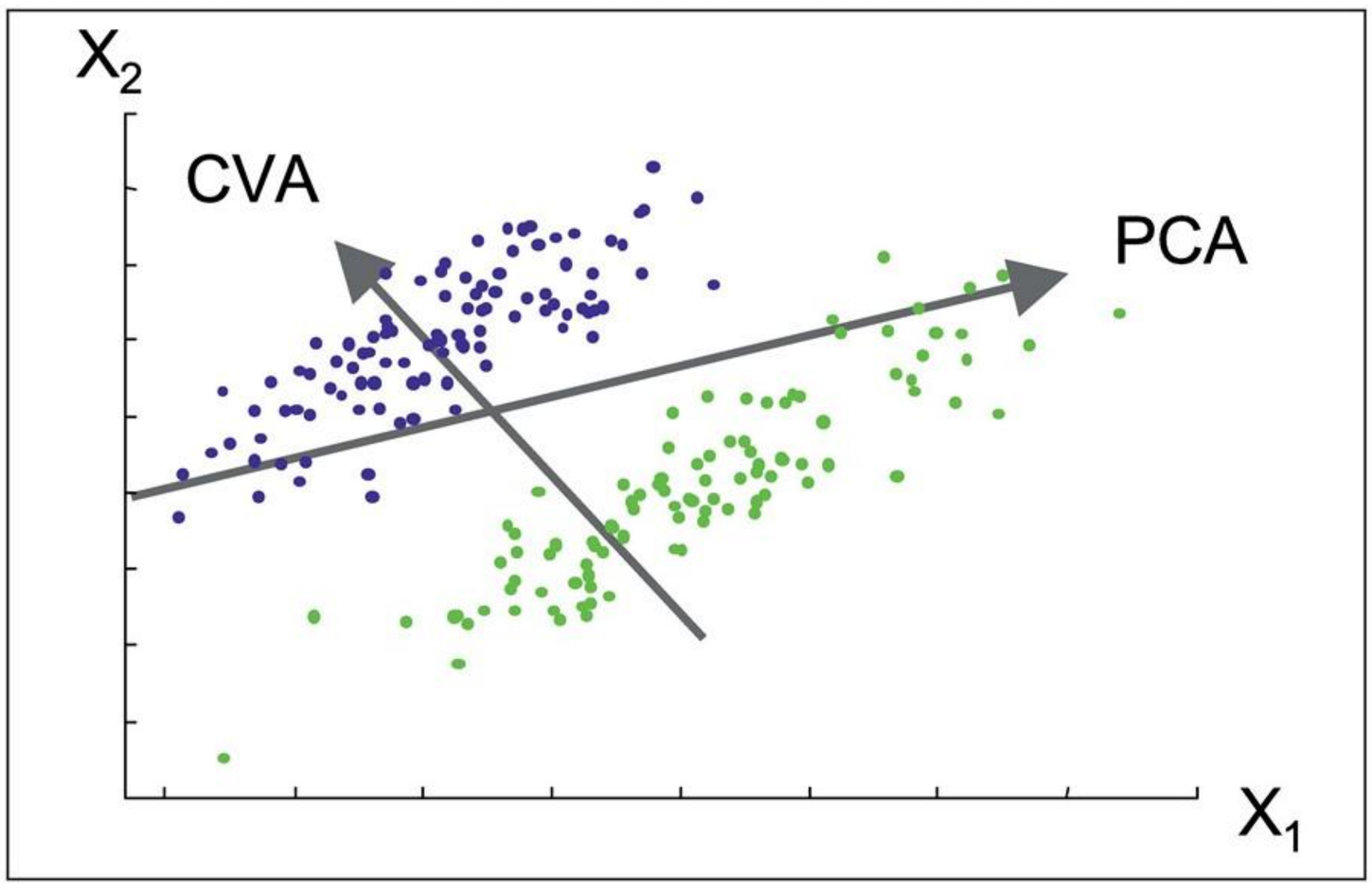
$$\mathbf{S}_{mellem\ grupper}\,\mathbf{w} = \lambda\mathbf{S}_{inden\ for\ grupper}\,\mathbf{w}$$

Dette kan omskrives til et standard egenværdiproblem ved at gange med den inverse for $\mathbf{S}_{\text{inden for grupper}}$

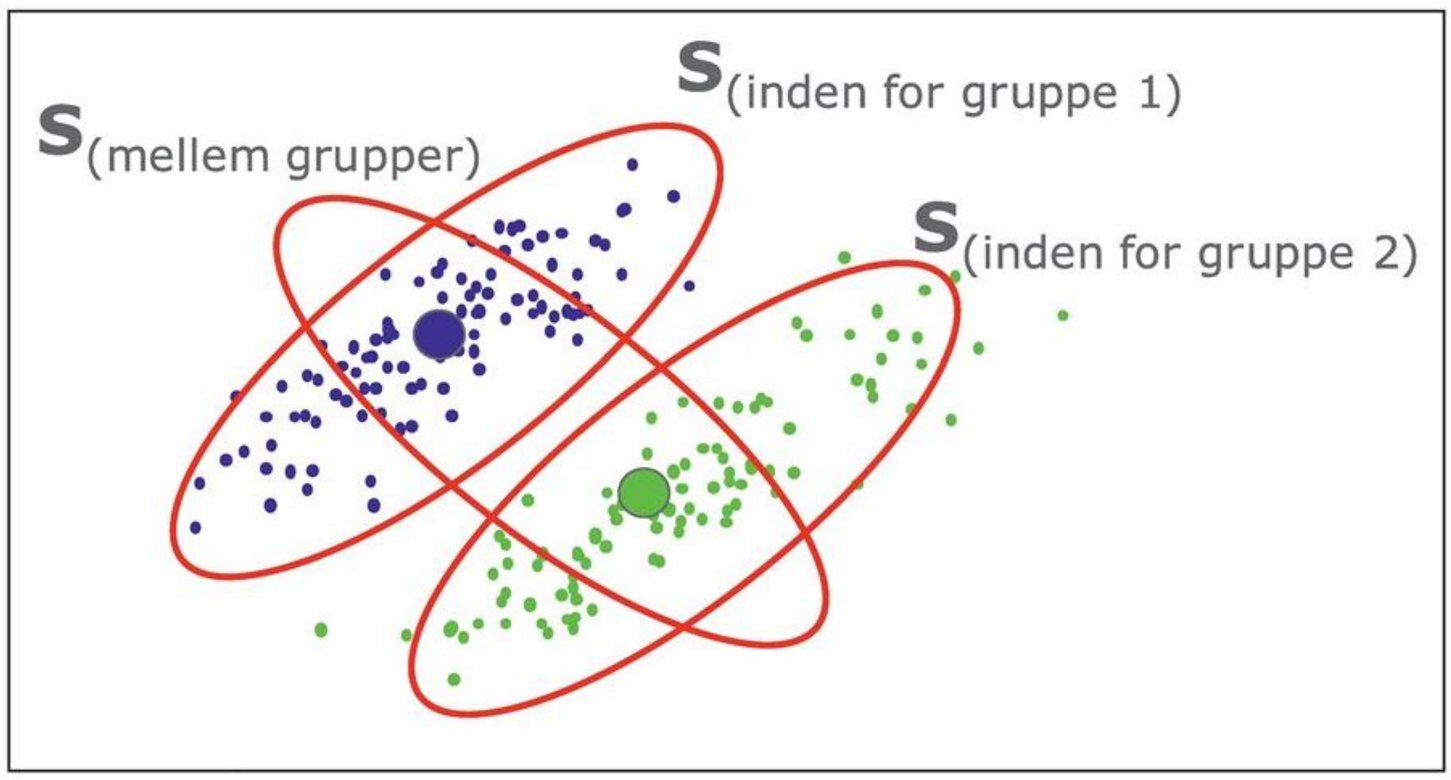
$$\mathbf{S}^{-1}_{indenforgrupper}\mathbf{S}_{mellemgrupper}\mathbf{w} = \lambda\mathbf{w}$$

Denne ligning kan umiddelbart løses med kendte matematiske metoder, og retningen **w** kan estimeres. Det er indbygget i

Det Kemometriske Rum



Figur 1. Illustration af princippet i PCA og CVA i et eksempel med to grupper i data (blå og grøn). I PCA estimeres den retning, som beskriver mest mulig variation i hele datasættet uden hensyn til den kendte gruppering. I CVA estimeres den retning, der ved projektion ind på retningen optimerer adskillelse af de kendte grupper (efter et givet matematisk udtryk).



Figur 2. Illustration af $S_{\text{mellem grupper}}$ og $S_{\text{inden for grupper}}$ som anvendes i CVA. $S_{\text{inden for grupper}}$ er sammensat af bidrag fra $S_{\text{inden for gruppe 1}}$ og $S_{\text{inden for gruppe 2}}$.

CVA-metoden, at der estimeres netop én retning (**w**), når klassifikationsproblemet omhandler to grupper; ved tre grupper findes to **w**-retninger og generelt findes der g-1 **w**-retninger for g grupper.

Selve klassifikationen

CVA finder den retning der bedst diskriminerer mellem f.eks. to grupper. For at kunne bruge metoden aktivt til klassifikation af nye, ukendte prøver er proceduren:

1) Projicér spektret af den nye prøve $\boldsymbol{x}_{ny}$ ind på den fundne retning **w**. På den måde estimeres en CVA-scoreværdi $\boldsymbol{t}_{ny} = \boldsymbol{x}'_{ny}\boldsymbol{w}$.

2) CVA-scoreværdien kan bruges til at klassificere den nye prøve ved at sammenligning med CVA-scoreværdierne for de kendte klasser. I praksis kan man f.eks. udnytte en standard Lineær Diskriminant Analyse (LDA) til dette.

Problem med spektrale data

CVA fungerer fint på såkaldte fuld-rang datamatricer; i kemiens verden vil det f.eks. være få uafhængige målinger på mange prøver. I et sådant tilfælde kan den inverse til $\mathbf{S}_{\text{inden for grupper}}$ let findes. Men hvis man har med NIR-spektre at gøre, vil dimensionen af $\mathbf{S}_{\text{inden for grupper}}$ blive f.eks. 700 × 700, og den inverse til en sådan stærkt kolineær matrix, kan ikke umiddelbart estimeres. Det svarer populært sagt til at dividere med nul. Tilsvarende gælder hvis man i datasættet har flere variable end prøver.

Alternativt kan man beregne en PCA på de rå NIR-data (**X** = **TP'** + **E**) og efterfølgende anvende CVA-metoden på scores (T). Beregnes ti scorevektorer i en PCA, vil dimensionen på **S**-udtrykkene blive 10 × 10 og med fuld rang, således at egenværdiligningen kan løses uden problemer. Ulempen er at fortolkningen bliver vanskeligere, da den sker på ti CVA-weights (**w**), som skal forbindes til de rå NIR-data gennem ti PCA loading-vektorer. Endvidere kan man miste relevant information, som ikke bliver inkluderet i de udvalgte score-vektorer.

Outro

CVA er en glimrende og effektiv metode til at finde diskriminative retninger i et multivariat datasæt, når grupperne på forhånd er kendte. Ulempen ved CVA er, at den ikke kan håndtere stærkt korrelerede variable og datasæt med flere variable end prøver; hvilket gør at den ikke direkte kan anvendes på f.eks. spektroskopiske datasæt. I næste klumme vil se på en lille ændring i løsningen af egenværdiproblemet, som betyder, at CVA kan anvendes direkte på spektroskopiske data.

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

Referencer
1. Krzanowski WJ. Principles of Multivariate Analysis (Revised edn). Oxford University Press: New York, 2000.
2. Rao CR. Advanced Statistical Methods in Biometric Research. Wiley: New York, 1952.