Klassifikation med Principal Component Analysis

Principal Component Analysis er ofte anvendt som første skridt i den eksplorative multivariate analyse, men den finder også anvendelse i klassifikationsproblemer

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

Stammer den analyserede blodprøve fra en rask eller syg person? Stammer oliespildet fra skib A, B eller C? Den slags udfordringer hører under temaet klassifikation, hvor det undersøges om prøver med ukendt tilhørsforhold kan tilegnes kendte grupper af prøver; f.eks. cancer eller kontrol.

PCA er et effektivt værktøj i den eksplorative dataanalyse, hvor man ønsker et overblik over sammenhæng mellem prøver og variable, og når det gælder klassifikation, kan principperne i PCA også anvendes. Klassifikationsmetoden som baserer sig på PCA hedder *Soft Independent Modeling of Class Analogy*, som forkortes SIMCA [1].

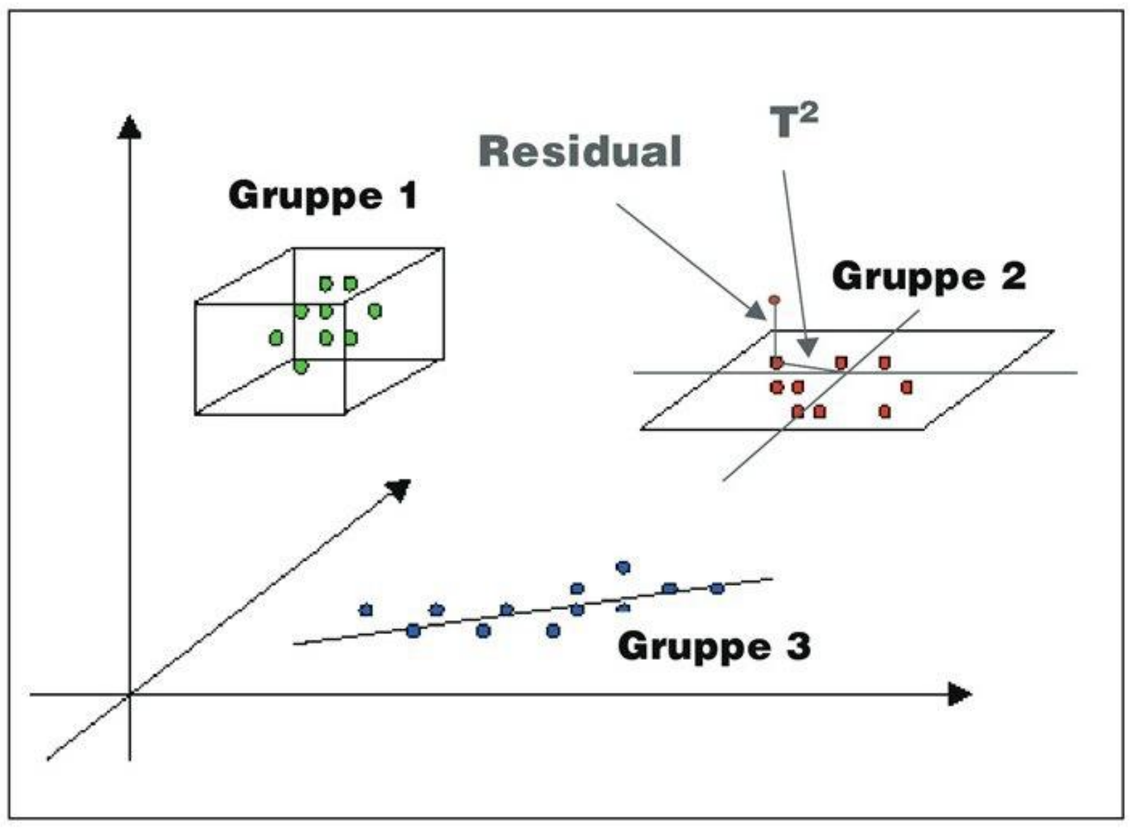
Princip i SIMCA

SIMCA-metoden består af to trin

1. Beregn en PCA model af data fra hver kendt klasse separat; disse data kaldes tilsammen for træningssættet.

2. Ukendte prøver, der skal klassificeres, sammenlignes med hver enkelt PCA model og tilegnes til klassen, hvis de passer ind i denne baseret på to afstandsmål: Residualvariation og Hotelling's T^2 (se klumme i Dansk Kemi nr. 3, 2008, for en nærmere beskrivelse af disse). Ordet *analogy* i den oprindelige titel refererer altså til analogi med træningsprøverne og PCA modellen af disse.

I figur 1 er princippet illustreret for én ukendt prøve.



Figur 1. Illustration af princippet i SIMCA. En ukendt prøves analogi til en PCA-model for hver klasse måles vha. residualvariation og Hotelling's T^2.

Tænkt eksempel - oliespild

Antag at vi har to kendte klasser: 20 prøver af olie fra skib A og 25 prøver af olie fra skib B. Prøverne er analyseret med en multivariat instrumentel metode med 50 variable (f.eks. gaskromatografi med 50 toppe). På begge sæt prøver er alle de samme variable målt. Der er spildt olie i Østersøen, og vi laver derfor gaskromatografisk analyse af en ny prøve fra dette oliespild. Formålet er at finde ud af, om synderen er skib A eller skib B eller ingen af disse?

Vi beregner en PCA-model for hver af de to kendte klasser:

$$\textbf{Skib A: } \mathbf{X}_A = \mathbf{T}_A\mathbf{P'}_A + \mathbf{E}_A \qquad \text{Ligning 1}$$

$$\textbf{Skib B: } \mathbf{X}_B = \mathbf{T}_B\mathbf{P'}_B + \mathbf{E}_B \qquad \text{Ligning 2}$$

Indeks angiver skibet, og i dette tilfælde ikke antal komponenter beregnet i modellen. Sidstnævnte er i øvrigt vigtigt at kunne estimere, og dette kan være forskelligt for de to modeller. Man kan altså f.eks. have en 4-komponent PCA-model for skib A og en 2-komponent PCA-model for skib B. Denne fleksibilitet er en fordel i SIMCA-metoden ift. andre metoder, som er baseret på beregninger på alle prøver (f.eks. PLS diskriminant-analyse, som vi vil beskrive senere). I ligning 1 og 2 vil også indgå forbehandling (f.eks. centrering og skalering), men dette er udeladt for at simplificere ligningerne.

Antallet af komponenter i modellerne estimeres ud fra f.eks. krydsvalidering eller testsæt-validering.

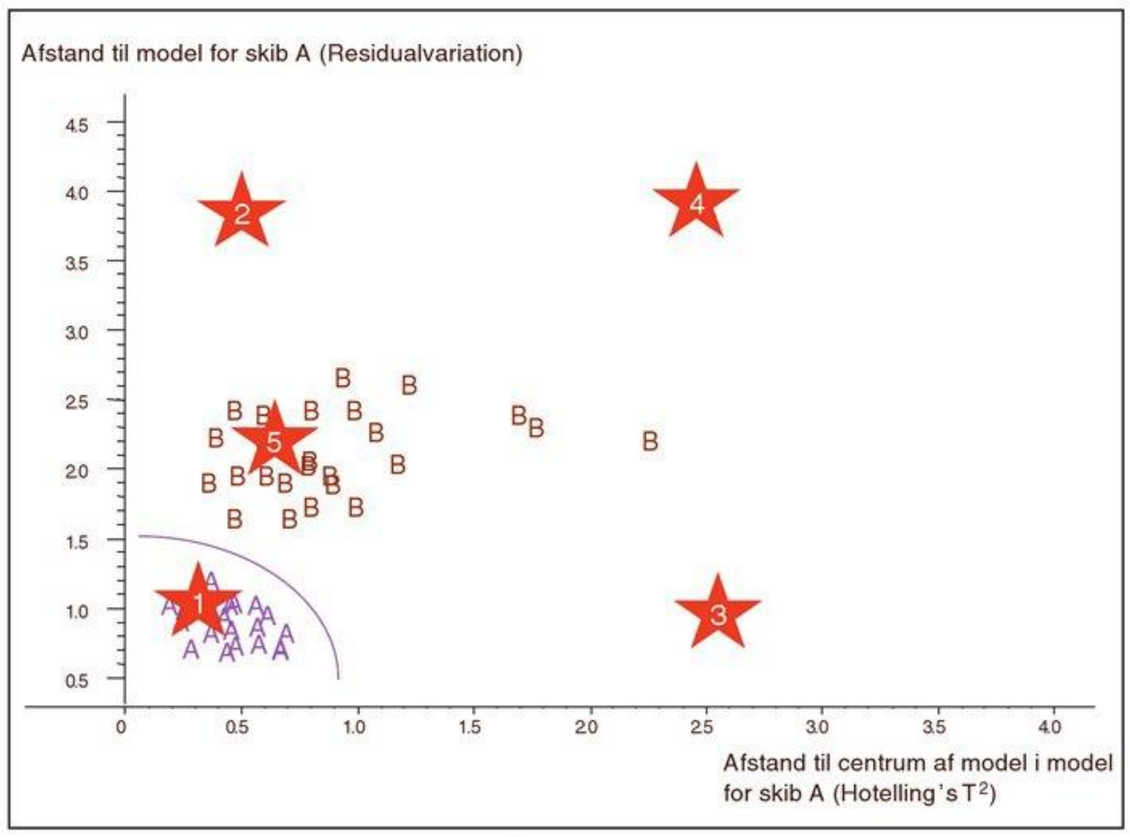
Hypotese

Vi opstiller nu den hypotese, at den ukendte prøve kommer fra skib A, og derfor må det antages, at prøven indeholder de samme kemiske variationer som prøverne fra skib A. En måde at formulere dette på er, at den ukendte prøve indeholder de samme loadings, som prøverne fra skib A.

Dette kan testes ved at *projicere* den ukendte prøve ned på modellen for skib A:

$$\textbf{Skib A: } \mathbf{x}_u' = \mathbf{t}_{uA}'\mathbf{P}_A' + \mathbf{e}_{uA}' \qquad \text{Ligning 3}$$

Her indeholder $\mathbf{x}_u$ den ukendte prøves gaskromatogram (en rækkevektor med 50 elementer) og $\mathbf{P}_A$ er loadings fundet i ligning 1 baseret på prøverne fra skib A. Ved at dividere $\mathbf{x}_u$ med $\mathbf{P}_A$ fås et estimat af prøvens scoreværdier for det antal komponenter PCA-modellen for skib A er beregnet med. Er den baseret på fire komponenter vil $\mathbf{t}_{uA}$ indeholde fire tal, som er scoreværdierne for den ukendte prøve i skib A PCA-modellen. Tilsvarende vil residualet $\mathbf{e}_{uA}$ indeholde et



Figur 2. Residualvariation versus T^2 med eksempler på placering af ukendte prøver. Se tekst for forklaring.

tal for hver variabel i gaskromatogrammet, som består af 50 variable.

Analogi

For at kunne sammenligne den ukendte prøve med de kendte prøver fra skib A beregnes Hotelling's T^2 og kvadratsummen af residualet for de kendte prøver og for den ukendte prøve. For den ukendte prøve beregner vi altså (se figur 1):

Residualvariation: Den ukendte prøves afstand til modellen (kvadratsum af residualvektoren)

T^2: Den ukendte prøves afstand til centrum af modellen i modellen (Hotelling's T^2 baseret på scorevektoren)

Tilsvarende afstandsmål beregnes for hver enkelt af de kendte prøver, og størrelsen af den ukendtes prøves residualvariation og T^2 sammenlignes med de kendte prøvers residualer og T^2-værdier. Dette foregår i et Residualvariation versus T^2 plot som vist i figur 2.

I figuren ses, at A-prøverne har de mindste residualer og T^2 afstande; det giver god mening, da modellen er udviklet til at beskrive netop disse prøver. Vi har i figuren også vist, hvorledes B-prøverne opfører sig, når de projiceres ind i PCA-modellen for skib A. I dette tilfælde er der fin adskillelse mellem A- og B-prøver.

Klassifikation

I figur 2 ses (med stjerner) fem ukendte prøvers placering i plottet:

- Ukendt 1 viser sig tydeligt at være en A-prøve, da den tydeligt er placeret blandt A-prøverne.

- Ukendt 2 har et højt residual ift. A-prøverne, og det tyder på et andet gaskromatografisk mønster, og dermed at prøven ikke stammer fra skib A.

- Ukendt 3 har en høj T^2-værdi, mens residualet svarer til residualet fra en skib A-prøve. Det betyder, at det gaskromatografiske mønster ligner skib A-prøverne, men intensiteten/koncentrationen for den ukendte prøve er enten meget høj eller meget lav ift. A-prøverne (dette bør selvfølgelig inspiceres i rådata). Det kunne give anledning til at undersøge, om der er sket en fejl i analyseprocessen.

- Ukendt 4 har både høj residualvariation og T^2, og det er usandsynligt, at prøven er fra skib A.

- Ukendt 5 er placeret blandt skib B-prøverne, og det er et hint om, at denne prøve kan være en skib B-prøve; dette bør dog altid testes i en tilsvarende PCA-model for skib B.

Den indlagte grænse for A-prøverne er fastsat af forfatterne, men det er muligt at lave automatiske SIMCA-klassifikationsalgoritmer med prædefinerede signifikansgrænser.

De ukendte prøver er nu testet ift. skib A, og den samme proces skal nu laves for skib B. Det sparer vi læserne for i dette tænkte eksempel og vil i stedet i næste klumme vise, hvordan SIMCA kan anvendes på et virkeligt datasæt.

Outro

SIMCA er en effektiv metode til multivariate klassifikationsproblemer. Det svære i SIMCA-metoden er at bestemme det korrekte antal komponenter i de enkelte PCA-modeller. Det er ofte er en større udfordring end f.eks. i PLS-modellering.

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

Referencer
1. Wold S. *Pattern-Recognition by Means of Disjoint Principal Components Models*. Pattern Recognition, 8(3): 127-139, 1976.