

Principal Component Analysis – den kemometriske arbejdshest

Principal Component Analysis (PCA) er den helt centrale metode i kemometrien. Princippet anvendes inden for mange forskningsdiscipliner med forskellige formål, og har derfor mange forskellige navne. Her introducerer vi den kemometriske version af PCA og viser efterfølgende anvendelsen af PCA på næringsstofdata på udvalgte McDonald's produkter

Af Lars Norgaard, Soren Balling Engelsen og Rasmus Bro

Gennemsnit af variable er fornuftigt. Vi starter med et simpelt eksempel for at illustrere tankegangen i PCA: på et universitet får de studerende karakterer på en skala fra 1 til 10, hvor 10 er topkarakter. Karaktererne gives i følgende fem kurser: fysik, kemi, matematik, historie og musik, og vi har adgang til karaktererne for to studerende Albert og Amadeus:

Navn\Kursus	Fysik	Kemi	Matematik	Historie	Musik
Albert	8	10	8	6	5
Amadeus	5	5	5	9	10

En måde at rangordne studerende på i forbindelse med optag på en videregående uddannelse er som bekendt at beregne karaktergennemsnit. Gennemsnittet for Albert er:

$$\begin{aligned} \text{Gennemsnit-Albert:} \\ &= (8+10+8+6+5) / 5 \\ &= (8+10+8+6+5) \times 0,2 \\ &= 0,2 \times 8 + 0,2 \times 10 + 0,2 \times 8 + 0,2 \times 6 + 0,2 \times 5 \\ &= 7,40 \end{aligned}$$

Et gennemsnit er et eksempel på en vægtet sum; i PCA kaldes resultatet af en vægtet sum for en score.

Når man beregner et gennemsnit er alle vægte ens; eftersom der er fem karakterer, bliver alle vægtene 0,2. Hvis man ønsker en score, der afspejler de studerendes evner i primært de naturvidenskabelige fag, kan man f.eks. vælge følgende vægte:

$$\begin{aligned} \text{Natarsnit-Albert:} \\ &= 0,3 \times 8 + 0,3 \times 10 + 0,3 \times 8 + 0,05 \times 6 + 0,05 \times 5 \\ &= 8,35 \end{aligned}$$

I denne vægtede sum bliver de tre naturvidenskabelige kurser vægtet mere (0,3) end kurserne i historie og musik (0,05).

De to forskellige sæt af vægte kan samles i hver deres vektor. I PCA kaldes vægtene for loadings og vektorerne benævnes derfor loading-vektorer. Disse kan skrives som:

$$\begin{aligned} \mathbf{p}_1 &= (0,2 \ 0,2 \ 0,2 \ 0,2 \ 0,2) \\ \mathbf{p}_2 &= (0,3 \ 0,3 \ 0,3 \ 0,05 \ 0,05) \end{aligned}$$

Disse vektorer kan anvendes til at beregne scores for alle andre studerende, der har fulgt de samme kurser. For Amadeus bliver score-værdierne:

$$\begin{aligned} \text{Gennemsnit-Amadeus:} \\ &= 0,2 \times 5 + 0,2 \times 5 + 0,2 \times 5 + 0,2 \times 9 + 0,2 \times 10 \\ &= 6,80 \\ &= \text{score-værdi 1 for Amadeus} \end{aligned}$$

$$\begin{aligned} \text{Natarsnit-Amadeus:} \\ &= 0,3 \times 5 + 0,3 \times 5 + 0,3 \times 5 + 0,05 \times 9 + 0,05 \times 10 \\ &= 5,45 \\ &= \text{score-værdi 2 for Amadeus} \end{aligned}$$

Ud fra de beregnede score-værdier kan man f.eks. lave plot hvor Gennemsnit-scores plottes mod Natarsnit-scores for flere studerende, og på den måde kan man karakterisere de studerende.

Ovennævnte eksempel illustrerer princippet i PCA: nye variable dannes som vægtede summer af de oprindelige målte variable. Vægtene kaldes loadings (**P**) og de nye variable kaldes scores (**T**). Normalt kendes vægtene ikke på forhånd; de beregnes ved hjælp af en algoritme, der estimerer vægtene direkte fra data i stedet for en subjektiv vurdering som i eksemplet.

Principal Component Analysis

Den matematiske model, der opstilles i forbindelse med PCA, er i den simpleste form

$$\mathbf{X} = \mathbf{TP}$$

hvor **X** er de rå data, **T** er score-matricen indeholdende score-vektorerne og **P** er loading-matricen indeholdende loading-vektorerne. Man kan opfatte PCA modellen som en opsplitning af information: fra rådata (**X**), der kan illustreres som ét Excel ark, laver PCA to nye Excel-ark: ét der indeholder information om prøverne (scores, **T**), og ét der indeholder information om variablene (loadings, **P**).

Opsplitningen af informationen laves på en sådan måde, at de to dele, **T** og **P**, forklarer mest mulig variation i rådata (**X**); PCA algoritmen finder vægtene (loadings), således at dette sker. Ingen andre vægte vil kunne beskrive mere af den systematiske variation i det givne datasæt. I karakter eksemplet ovenfor kan ▶

	(Energi (kJ/g) - 9,6) / 1,7	(Protein (%) - 9,3) / 4,6	(Kulhydrat (%) -25,0) / 7,1
Apple Pie	1.1	-1.4	1.0
Big Mac	0.0	0.7	-0.8
Cheeseburger	0.6	1.0	0.1
Filet-O-Fish	1.3	0.3	0.2
Grilled Chicken	-0.9	0.7	-1.4
Hamburger	0.3	0.7	0.5
McChicken	-0.2	0.5	-0.8
McFeast	-0.6	0.5	-1.5
Pommes Frites	1.6	-0.9	1.7
Quarter Pounder m/ost	0.6	1.4	-1.0
Sundae Chokolade	-1.0	-1.1	0.5
Sundae Jordbær	-1.6	-1.2	0.5
Sundae Karamel	-1.0	-1.2	0.9

dette tolkes således, at den første loading er de vægte, der bedst ligner flest mulig karakterprofil – altså en slags gennemsnitsprofil. Og score for hver person bliver så et tal, der angiver, hvor meget personen har af denne gennemsnitsprofil. Den næste loading kan derefter findes ved en opvægtning af elevernes naturvidenskabelige evner, og score 2 angiver derfor hvor meget personen har af disse færdigheder.

McDonald's data

I første klumme (Dansk Kemi, december 2007) gav vi læserne den udfordring, at de skulle karakterisere/gruppere udvalgte produkter fra McDonald's i forhold til indholdet af givne næringsstoffer og energi. I tabel 1 ses de niveau- og varians-korrigerede data. Denne korrektion kaldes autoskalering, og den sikrer, at alle målinger kan bidrage på lige fod i dataanalysen. For denne demonstration er det blot vigtigt at vide, at f.eks. protein-procenten er angivet som $(\text{Protein } \% - 9,6) / 1,7$.

Vi anvender nu et program* til at beregne en PCA model af data angivet i tabel 1. På en moderne PC tager dette et split-sekund. Resultatet er en loading-matrix, der indeholder vægtene og en tilhørende scores matrix med gennemsnittene baseret på vægtene.

De to første loading-vektorer ser ud som følger:

$$\mathbf{p}_1 = (-0,56 \quad -0,42 \quad 0,19 \quad -0,63 \quad -0,27)$$

$$\mathbf{p}_2 = (-0,42 \quad 0,54 \quad -0,71 \quad -0,19 \quad -0,02)$$

Det betyder, at den første score-værdi (vægtet sum) for et produkt findes som

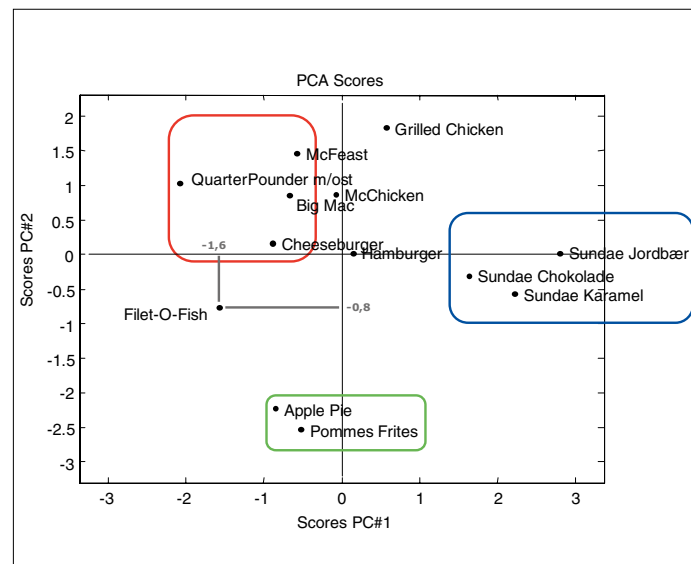
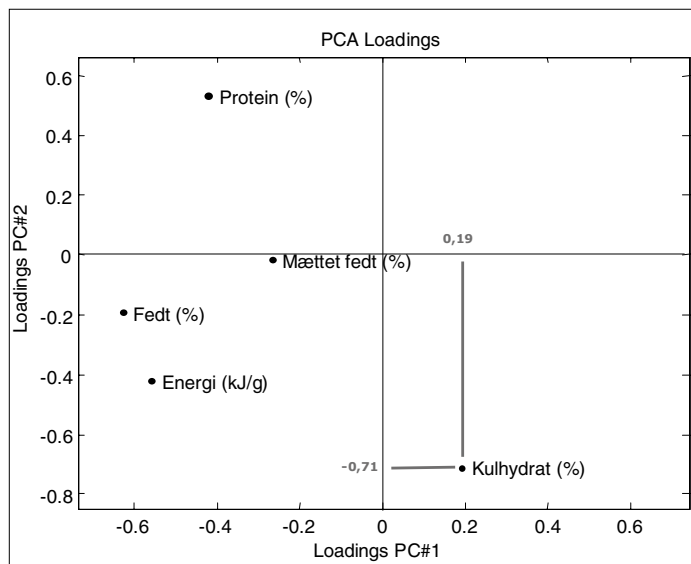
$$\text{Score 1} = -0,56 \times \text{Energi} - 0,42 \times \text{Protein} + 0,19 \times \text{Kulhydrat} - 0,63 \times \text{Fedt} - 0,27 \times \text{Mættet fedt}$$

For Filet-O-Fish bliver score-værdien således:

$$\text{Score 1 Filet-O-Fish} = -0,56 \times 1,3 - 0,42 \times 0,3 + 0,19 \times 0,2 - 0,63 \times 1,3 - 0,27 \times (-0,3) = -1,6$$

Tilsvarende bliver anden score-værdi $-0,8$. Filet-O-Fish kan nu placeres i et score-plot af score 1 mod score 2. Helt analogt kan alle andre produkter placeres i score-plottet ud fra deres score-værdier og resultatet bliver som vist i figur 1.

Bemærk at det er de autoskalerede værdier, der indgår i den viste beregning.



Autoskalerede McDonald's data.

(Fedt (%) - 10,1) / 3,7 (Mættet fedt (%) - 3,3) / 1

1.3	0.6
0.2	0.4
0.0	0.7
1.3	-0.3
-0.2	-2.0
-0.6	-0.5
0.2	-1.0
0.3	0.8
0.9	-0.8
0.8	1.9
-1.1	0.7
-1.8	-0.5
-1.5	-0.1

På tilsvarende vis kan man afbilde loadings for at se hvilke variable, der har givet anledning til score-plottet. Loading-plottet, figur 2, laves ved at plote vægtene mod hinanden: kulhydrat placeres ved koordinaterne (0,19, -0,71) osv.

Kendetegnende for kemometrien er at data og modelresultater præsenteres grafisk, og de grundlæggende plots er netop et score-plot og et loading-plot.

For scoreplottet gælder at produkter, der placerer sig tæt ved hinanden, har en energi- og næringsstofsammensætning, der minder om hinanden set over alle målinger. Dette følger direkte fordi score-værdien angiver hvor 'meget' produktet indeholder af den vægtede sum eller profil, som er mest kendetegnende for produkterne som sådan. Produkter, der ligger langt fra hinanden set i forhold til (0,0), er forskellige fra hinanden.

Øverst til venstre ses en samling af burger-produkter, til højre Sundae desserter og nederst Apple Pie og Pommes Frites. Det fremgår desuden, at McDonald's oprindelige kerneprodukt, Hamburgeren, er placeret tæt på (0,0), hvilket svarer til det gennemsnitlige indhold af næringsstoffer for de undersøgte produkter.

PCA giver overblik

Loading-plottet forklarer *hvorfor* prøverne placerer sig som de gør: øverst til venstre ligger proteinindhold, og det betyder, at prøver, som har en høj scoreværdi i denne retning, har højt proteinindhold. Det giver intuitiv mening i forhold til de kødholdige burger-produkter.

Der er til gengæld forholdsmeæssigt mindre kulhydrat i burgerne; her er det desserterne og Pommes Frites, der har stort indhold (sukker henholdsvis stivelse). Det er vigtigt at forstå, at PCA er en relativ dataanalyse: har man i sin analyse kun medtaget produkter med et kulhydrat indhold, der varierer mellem f.eks. 10 og 11%, vil PCA modellen udspænde netop denne variation.

Det, som er fantastisk ved PCA, er at man får en intuitiv grafisk afbildning som meget præcist fortæller om sammenhængen mellem mange prøver og mange variable: Disse prøver ligner hinanden; disse er meget forskellige; disse prøver er ens, fordi de indeholder meget af disse variable osv. Et sådant overblik er ofte svært at få, men ved hjælp af PCA kan man nemt få overblik over selv meget komplicerede datasæt.

I dette eksempel med McDonald's data havde vi kun få prø-

ver og få variable, og selv her, er det tydeligt at PCA øjeblikkeligt giver en overskuelig forklaring af hvorledes målingerne relaterer sig til hinanden. I de fleste reelle situationer indenfor videnskab og teknologi, er det normalt at have tusindvis af målinger og så er det endnu mere essentielt, at man ikke overser vigtige sammenhænge, fordi man kun kigger på én variabel ad gangen.

I den næste klumme vil vi gå lidt mere i dybden med PCA, og demonstrere forskellige aspekter af PCA på multivariable data fra nær-infrarød (NIR) spektroskopi.

Fodnote

* I denne klumme er anvendt programmet LatentiX (www.latentix.com) (freeware til PCA).

McDonald's data kan downloades
fra www.model.life.ku.dk