

Når PARAFAC ikke virker

PARAFAC-modellen kan give unik indsigt i data, når den virker. Desværre har PARAFAC også nogle helt specielle matematiske egenskaber, som gør, at den ikke altid giver brugbare resultater.

Af Rasmus Bro, Søren Balling Engelsen,
Institut for Fødevarevidenskab,
Københavns Universitet og Lars Nørgaard, FOSS

Når PARAFAC virker, så kan den give næsten magiske resultater. PARAFAC kan "afkode" komplekse data og finde de rene underliggende komponenter, hvilket giver utallige nye muligheder. PARAFAC-modellen virker aldeles glimrende på data, som følger modellen. Det har vi bl.a. set med fluorescens- og NMR-data i tidligere klummer.

Det er desværre sådan, at PARAFAC ikke altid virker. Faktisk er matematikere en smule "skræmt" over PARAFAC-modellen, da det viser sig, at generelle løsninger på PARAFAC-problemer er meget vanskelige at garantere. Man siger, at PARAFAC-problemet er NP-hårdt [1], og det har praktisk betydning. Der er situationer, hvor ingen PARAFAC-algoritme kan beregne en model, der giver den bedste løsning. Og den "ikke-løsning", man i stedet får, vil i praksis være matematisk og kemisk nonsens. Det er således vigtigt, at kende til de praktiske implikationer af dette problem, når man vil anvende PARAFAC. Prisen for de fantastiske muligheder med PARAFAC er, at den er numerisk spidsfindig.

Hvordan ser en ikke-løsning ud?

Når PARAFAC-problemet ikke lader sig løse opstår der, hvad der kaldes en tofaktor-degenerering [2]. Det er en situation, hvor to PARAFAC-komponenter bliver næsten helt identiske, men med omvendt fortegn. Lad os sige, at vores data ligger i en trevejs kasse $\underline{X}$, og at vi har beregnet en trekomponent PARAFAC-model. Hver komponent består af en scorevektor for

prøveretningen og en loading-vektor for hver variabelretning (figur 1).

Når man har tofaktor-degenerering, så vil to af komponenterne; eksempelvis $\underline{Z}_1$ og $\underline{Z}_2$ være næsten identiske, men med omvendt fortegn. Der gælder således, at

$$\underline{Z}_1 = -\underline{Z}_2 + \underline{\varepsilon},$$

hvor $\underline{\varepsilon}$ er en trevejskasse med små tal. Derfor vil summen

$$\underline{Z}_1 + \underline{Z}_2,$$

blive cirka nul og trekomponent-modellen bliver i praksis til en énkomponent-model, fordi

$$\underline{Z}_1 + \underline{Z}_2 + \underline{Z}_3 \sim \underline{Z}_3.$$

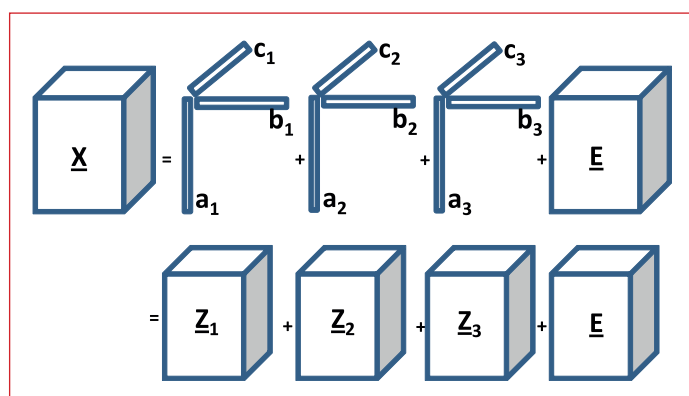
Det kan vises, at jo længere man beregner (itererer), jo mere lig bliver de to faktorer, og efter uendelig mange iterationer vil deres samlede bidrag blive eksakt nul. Samtidig vokser variansen i de to faktorer mod uendelig. Faktorerne vokser altså mod uendelig store tal, imens deres sum går mod nul.

Det skal understreges, at denne opførsel er et matematisk artefakt, som ikke afspejler nogen reel kemisk information i data. Desværre gør sådan et artefakt det umuligt at bruge den pågældende model til noget fornuftigt. Et eksempel er vist i figur 2, hvor en trekomponent-model er beregnet på sensoriske data. I dette tilfælde er 30 forskellige oste blevet bedømt af otte dommere mht. 23 forskellige deskriptorer, som afspejler smag, aroma, mundfylde, udseende m.m. [3]. Sensorisk analyse er en kvantitativ og objektiv målemetode, som anvendes i stor stil i bl.a. fødevareindustrien [4]. I dette tilfælde har vi en $30 \times 8 \times 23$ trevejs kasse. PARAFAC vil under normale omstændigheder være en ideel model for sådan et datasæt, for det giver mening at antage, at sensorikken afspejler latente, underliggende variable af mere fundamental karakter.

Desværre ser komponent et og to fra en trekomponent-model ud som vist i figur 2. Komponent et er korreleret med komponent to i alle tre retninger. For scores er korrelationen negativ, mens den er positiv for de to andre. Dette betyder eksplicit, at de to komponenter cirka udslukker hinanden.

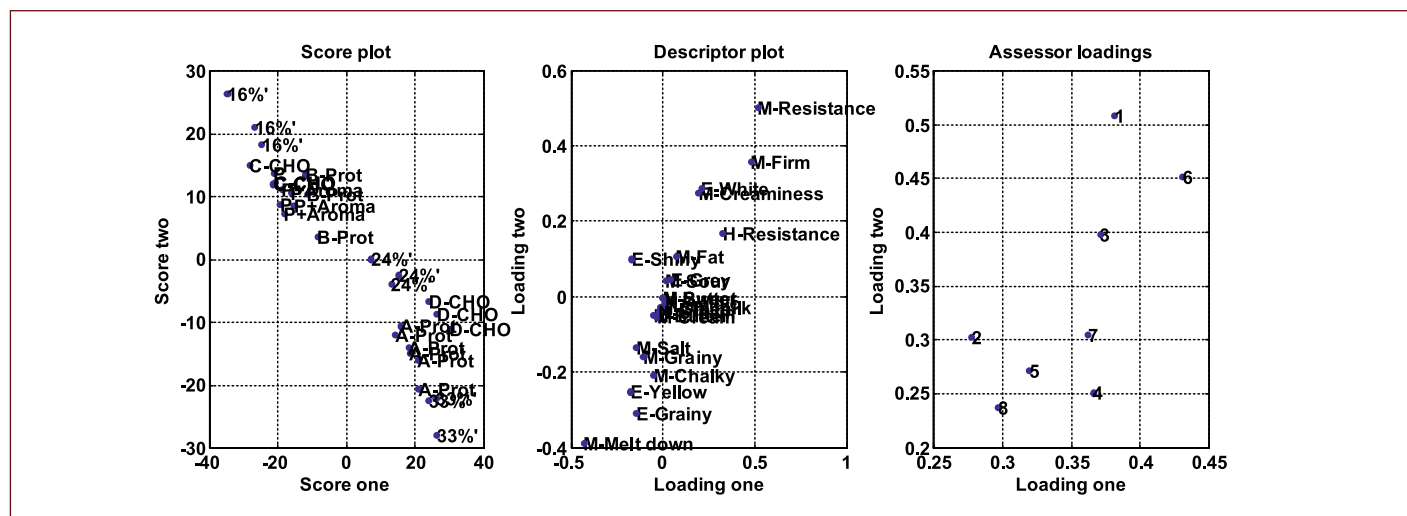
Kongruens kan måle det

For at kunne måle i hvor høj grad man har problemer med degenerering, kan man måle, hvor ens de forskellige komponenter er. Dette kan gøres ved at beregne korrelationen mellem komponenter for de forskellige $\underline{Z}$ trevejsstrukturer, efter man har foldet dem ud til en lang vektor. Korrelationer kan dog snyde lidt, for de kan blive perfekt én, selv når vektorerne er forskudt. Hvis eksempelvis vektorerne $\underline{a}$ og $\underline{b}$ har en korrelation på én, så vil $\underline{a}$ og $(\underline{b} + 10)$ også have en korrelation på én. I stedet for ►



Figur 1. Illustration af trekomponent PARAFAC-model. Hver komponent består af en score og to loading-vektorer. Ganger man disse sammen, får man den aktuelle énkomponent model/approximation, som er vist nederst i figuren for hver komponent. De to repræsentationer er således identiske.

■ DET KEMOMETRISKE RUM



Figur 2. Score og loading plots af trekomponent PARAFAC-model for sensoriske data. I dette tilfælde er 30 forskellige oste blevet bedømt af otte dommere mht. til 23 forskellige deskriptorer. Komponent et er plottet mod komponent to.

korrelation anvendes derfor en såkaldt Tucker Kongruens [5], som er identisk med korrelation, dog uden at korrigere for en konstant.

I praksis følger man kongruensen mellem alle komponenter og hvis en af disse bevæger sig mod -1, så skal man tage sine forholdsregler. Ofte vil man ikke reagere, før kongruensen er omkring -0.85 og kun hvis den vedvarende falder undervejs i iterationerne.

Hvad kan man gøre?

Som nævnt, så er en model med tofaktor-degenerering ikke valid. Og det er af matematiske årsager, så det har ingen konsekvenser for kvaliteten af ens data. Det er ikke data, der skal "bebrejdes". I praksis oplever man aldrig problemet for data, der følger PARAFAC-modellen fint, mens det eksempelvis i sensorik forekommer oftere.

Der er adskillige måder man kan prøve at håndtere problemet på. Man kan ændre antallet af komponenter op eller ned, og se om det fjerner problemet (og stadig giver en brugbar model). Man kan også ændre lidt på data. F.eks. kan man fjerne variable, der alligevel antages at være mindre væsentlige, eller man kan ændre på forbehandlingen af data ved f.eks. at skalere på anden vis. Endelig kan man også overveje at anvende andre modeller, som ikke har det samme problem.

Outro

I næste klumme vil vi se på en anden multivejs-model, kaldet Tucker, som ikke har samme problem med tofaktor-degenerering. Til gengæld er disse modeller ikke helt så lette at anvende og fortolke som PARAFAC-modeller.

E-mail-adresser

Rasmus Bro: rb@life.ku.dk

Søren Balling Engelsen: se@life.ku.dk

Lars Nørgaard: LNO@foss.dk

Referencer

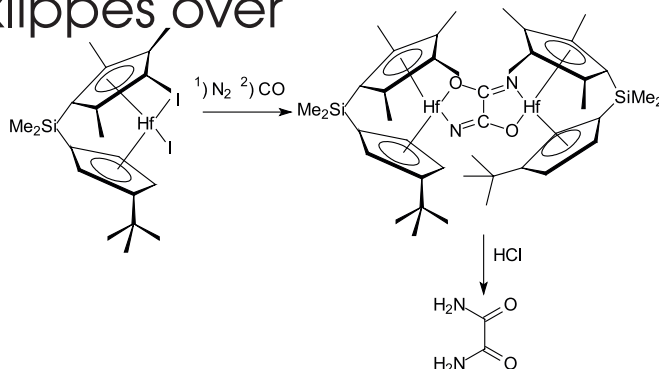
1. V. de Silva, L. H. Lim. Tensor Rank and the Ill-Posedness of the Best Low-Rank Approximation Problem. *Siam Journal on Matrix Analysis and Applications* 30 (3):1084-1127, 2008.
2. J. B. Kruskal, R. A. Harshman, M. E. Lundy. How 3-MFA data can cause degenerate PARAFAC solutions, among other relationships. In: *Multiway Data Analysis*, edited by R. Coppi and S. Bolasco, Amsterdam:Elsevier, 1989, p. 115-122.
3. R. Bro, E. M. Qannari, H. A. L. Kiers, Tormod Næs, M. B. Frøst. Multiway models for sensory profiling data. *J.Chemom.* 22:36-45, 2008.
4. H. Martens, M. Martens. *Multivariate Analysis of Quality: An Introduction*, Chichester:Wiley & Sons, 2001.
5. G. Lorho, F. Westad, R. Bro. Generalized correlation loadings - Extending correlation loadings to congruence and to multi-way models. *Chemom.Intell.Lab.Syst.* 84 (1-2):119-125, 2006.

■ Nyt om...

Bindingen i elementært N_2 klippes over

Luftens nitrogen udnyttes ikke i ret mange industrielle processer pga. vanskeligheden med at bryde den stærke nitrogen-nitrogen-binding. En undtagelse er Haber-Bosch-processen til fremstilling af NH_3 ud fra N_2 ; men den kræver også højt tryk og temperatur. Forskere ved Cornell University i USA har nu brudt N-N-bindingen, samtidig med brydningen af bindingen i et andet meget stabilt molekyle CO under dannelse af oxamid, som vist til højre. Hafniumsandwichforbindelsen til venstre reageres successivt med N_2 og CO ved stuetemperatur og et tryk på 1 atm. Den nye sandwichforbindelse spaltes af HCl til bl.a. oxamid.

Carl Th.



Dinitrogen cleavage and functionalization by carbon monoxide promoted by a hafnium complex, *Nature Chemistry* 2, January 2010, side 30.