1. Tredje obligatoriske læringsaktivitet i MAT102 - innlevering av kode om PCA - frist 30/10

Innlevering 3 har oppgaver om forholdsvis store utregninger som skal utføres på datamaskin, og så tolkes. Dere har lov til å kalle funksjoner som er lagt ut på canvas om PCA. Koden skal leveres. Der det er spørsmål etter tallsvar fra utregninger kan dere legge svaret som kommentar i koden. På oppgave (1), for eksempel, betyr det at hele den preprosesserte matrisen skal limes inn som en kommentar etter koden som genererer den. På oppgaver som (2) skal hele svaret stå som en kommentar i besvarelsen (NB: Prøv å være kort og presis).
Inndata ligger i filen Arbeidskrav3.mat. I Matlab importeres denne filen med

```
load Arbeidskrav3.mat
```

I Python må også et bibliotek på plass, og litt ekstra formatering for å få rette datatyper:

```
import scipy.io
data = scipy.io.loadmat("Arbeidskrav3.mat")
X1 = np.array(data['X1'])
X2 = np.array(data['X2'])
```

Dere trenger matrisen X1 og navnelistene objNames1 og varNames1. (7) og (8) er frivillige (men ikke så mye ekstraarbeid), der trenger dere også matrisen X2 og navnelistene objNames2 og varNames2. NB: I Python vil `data` ha type `dict`. Det innebærer at f.eks. X1 kan plukkes frem ved å skrive `data[‘X1’]`, se eksempellinjene over. objNames og varNames kan brukes direkte på denne måten, og vil virke sammen med annotate i plotting i Python.

I artikkelen Folkenberg et al: What is mouthfeel?... som ligger på canvas analyseres ulike målinger fra automatkakao. I denne oppgaven skal vi kontrollregne på deres data. Vi starter med et utdrag av dataene fra artikkelen. Disse dataene er samlet i filen Arbeidskrav3.mat; det er alle delene som slutter på 1. Variabelnavnene som svarer til søylene/kolonnene i matrisen er i rekkefølge

```
'%COCOA';'%SUGAR';'%MILK';'COLOUR(L)';'VISCOSITY/10';
'colour';'cocoa-odour';'smooth-txtr';'milk-taste';'sweet';
```

De første fem er fysiske målinger, de siste fem er input fra et smakspanel. Prøvene er tatt for sju ulike innstillinger på kakaoautomaten. Innstilling 5 er foretatt to ganger, som en test på om smakspanelet er konsistent fra forsøk til forsøk. Dermed er det totalt åtte rekker i matrisen, med objektnavn

```
'1:Milk+';'2';'3:Sugar+';'4';'5a';'5b';'6';'7:Cocoa+';
```

(1) Preprosesser matrisen. Altså gjør den til en matrise med null i gjennomsnitt og 1 i standardavvik for hver søyle/kolonne.
(2) Forklar med egne ord hvorfor standardiseringen er viktig i denne situasjonen.
(3) Regn ut de to første prinsipalkomponentene, og plott de åtte objektene mot prinsipalkomponentene (score plot).
(4) Som en test, er de to forsøkene med innstilling 5 nær hverandre?

(5) Hvor stor del av variasjonen er forklart?
(6) Plot så de ti variablene (loadings plot). Kom med minst en kommentar til dataene basert på analysen og plottene. NB: Når score plot og loadings plot skal vurderes sammen, tegn begge basert på samme pca-utregning (pga. random startvektor).
(7) (**(7) og (8) er frivillige ekstraoppgaver**) Nå ser vi på hele datamengden fra artikkelen, samlet i Arbeidskrav3.mat; det er alle delene som slutter på 2. Forsøksrekken er foretatt en gang til, men denne gangen er det tilsatt et ekstra stoff (for alle innstillingene). Innstilling fem ble ikke dublisert i den nye forsøksrekken. Det er også foretatt noen flere målinger, så datamatrisen er litt større. Foreta PCA igjen, også nå med to prinsipalkomponenter.
(8) Kan du se fra dataene en måte å skille de gamle (uten tilsetting) fra de nye (med tilsetting)? Sammenlign med figur 2B på side 188 i artikkelen (som vårt plott bør ligne på, men ikke være helt identisk med).

Preben og Jon Eivind