HDDA Tutorial: MDS : Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 5

For this tutorial we will use the `UScereal` dataset which is available if you install and load the package `MASS`. To load the dataset use the command `data(UScereal)`. Each observation is a brand of breakfast cereal and in total data are available on 11 different variables. Have a look at the help documentation for `UScereal` to familiarise yourself with the data.

1. Remove the non-metric variables `vitamins`, `shelf` and `mfr`.

```
#First load required packages
library(MASS) #MASS used for data
library(tidyverse)
cereal_metric<-select(UScereal,-vitamins,-shelf,-mfr) # Note use of minus
str(cereal_metric)#Confirm non-metric removed
```

```
## 'data.frame':    65 obs. of  8 variables:
##  $ calories : num  212 212 100 147 110 ...
##  $ protein  : num  12.12 12.12 8 2.67 2 ...
##  $ fat      : num  3.03 3.03 0 2.67 0 ...
##  $ sodium   : num  394 788 280 240 125 ...
##  $ fibre    : num  30.3 27.3 28 2 1 ...
##  $ carbo    : num  15.2 21.2 16 14 11 ...
##  $ sugars   : num  18.2 15.2 0 13.3 14 ...
##  $ potassium: num  848.5 969.7 660 93.3 30 ...
```

2. The intention is to use 8-dimensional Euclidean distance between the observations as an input to MDS. Should the data be scaled before computing the distance measure?

The data are measured in different units. Most variables are measured in grams but sodium is measured in micrograms, and, more importantly, calories is measured in calories. Without standardisation, if calories were converted to kilo Joules or the other variables converted to ounces then results would change.

3. Find the 2-dimensional classical MDS solution and plot it.

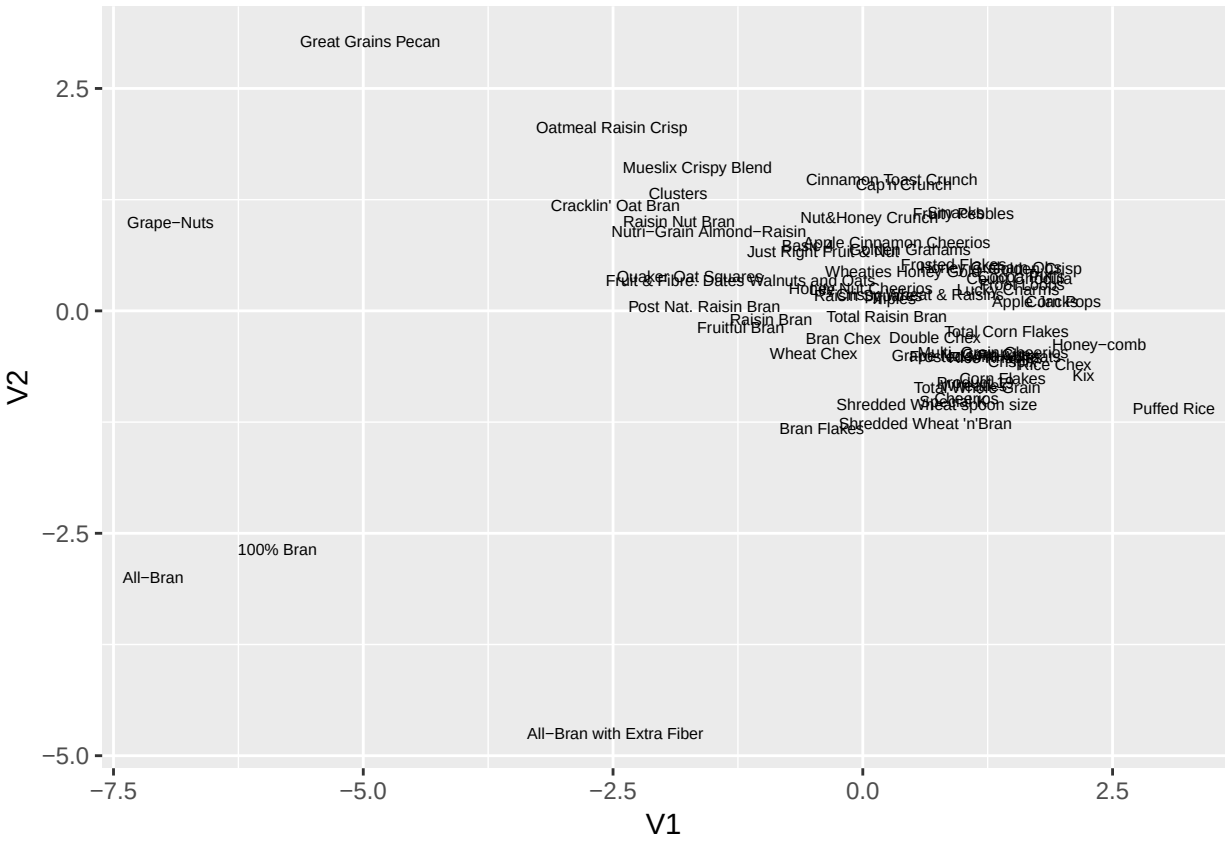
```
cereal_metric%>%
  scale%>% #standardise
  dist->dd #Compute distance

#Assign ceral names to dist object
rownames(UScereal)->attributes(dd)$Labels


#Compute classical MDS
cmds<-cmdscale(dd,eig = T) # Set eig=T for later questions

#Store representation in data frame
cmds$points%>%
  as.data.frame()%>%
  rownames_to_column(var = 'Cereal Name')->df

ggplot(df,aes(x=V1,y=V2,label=`Cereal Name`))+
  geom_text(size=2)
```



4. Does the plot indicate that one or more cereal brands could be outliers?

All Bran, All Bran with extra Fibre, 100% Bran, Grape Nuts and Great Grains Pecan are all potentially outliers. On closer inspection the first three of these are very high in fibre and protein. Also Grape Nuts and Grape Grain pecan are high in calories and carbohydrates.

5. What are the goodness of fit measures for this solution? Are they same or different?

```
cmds$GOF
```

```
## [1] 0.6982353 0.6982353
```

```
# The GOF measures are the same
```

6. Are there (non-negligible) negative eigenvalues? Why or why not?

```
#Check the minimum eigenvalue
min(cmds$eig)
```

```
## [1] -5.859887e-14
```

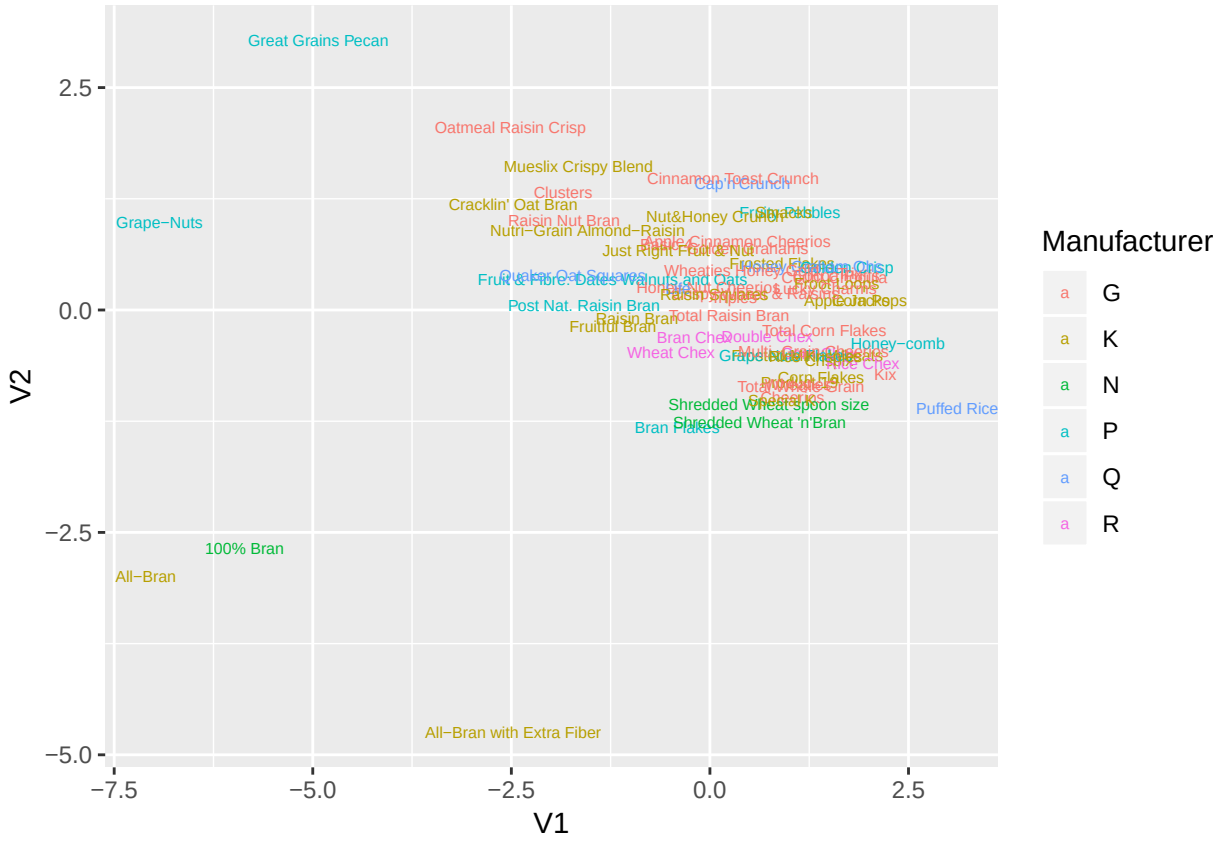
```
# This looks negative but the e-14 is the reciprocal of 1 with 14
#trailing zeros.  This number is indistinguishable from zero.  So
#all eigenvalues are non-negative. This is to be expected since
#input distance is Euclidean.
```

7. How would you expect your answer to questions 5 and 6 to change if Manhattan distances are used.

For Manhattan distances some eigenvalues can be negative and in this case the GoF measures may differ.

8. Re do the plot but with different coloured labels for each manufacturer. What conclusions do you draw from this analysis?

```
df<-add_column(df,Manufacturer=UScereal$mfr)
ggplot(df,aes(x=V1,y=V2,col=Manufacturer,label=`Cereal Name`))+
  geom_text(size=2)
```




```
#The two big manufacturers are General Mills and Kelloggs. Kelloggs
#brands are a bit more spread out (this is easier to see if we
#simply use points rather than the cereal names). General Mills may
#have too many similar brands competing with one another. General
#Mills may benefit from diversifying into a product similar to
#all-Bran or 100% bran. There are obvious limitations to this
#analysis for example the bran market may be too small to be
#profitable.
```

9. Re-do the analysis using the Sammon mapping. Do your conclusions change?

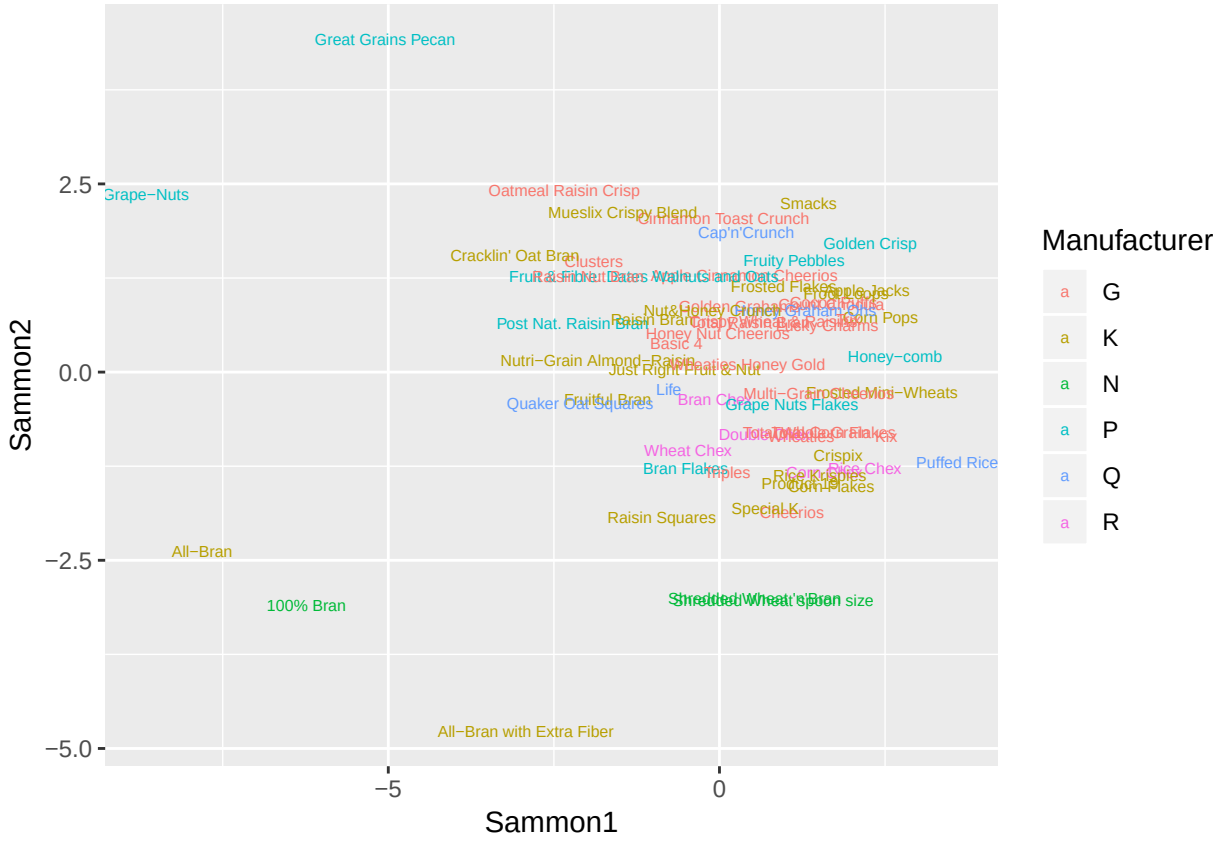
```
smds<-sammon(dd)
```

```
## Initial stress        : 0.09834
## stress after  10 iters: 0.06473, magic = 0.018
## stress after  20 iters: 0.03582, magic = 0.213
## stress after  30 iters: 0.02990, magic = 0.500
## stress after  40 iters: 0.02928, magic = 0.500
## stress after  50 iters: 0.02900, magic = 0.500
## stress after  60 iters: 0.02897, magic = 0.500
```

```
#For Sammon the coordinates are in points and are in matrix form
#Using [,1] and [,2] allows us to use the first and second columns
#respectively.
df<-add_column(df,Sammon1=smds$points[,1],
```

```
              Sammon2=smds$points[,2])

ggplot(df,aes(x=Sammon1,y=Sammon2,col=Manufacturer,label=`Cereal Name`))+
  geom_text(size=2)
```




```
# The results are fairly similar. The conclusion that Kelloggs
#brands are more diverse is perhaps a bit clearer when the Sammon
#mapping is used.  In a report it would be worth only presenting
#one solution while stating that the other solution gave mostly
#similar results.
```