Selvstudie 1 - PCA/C4.5

Nicholas Fitzhugh

9/2/2017

Opgave 1 (PCA)

```
library(MASS)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.3.2
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
## Warning: package 'tidyr' was built under R version 3.3.2
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
## select(): dplyr, MASS
```

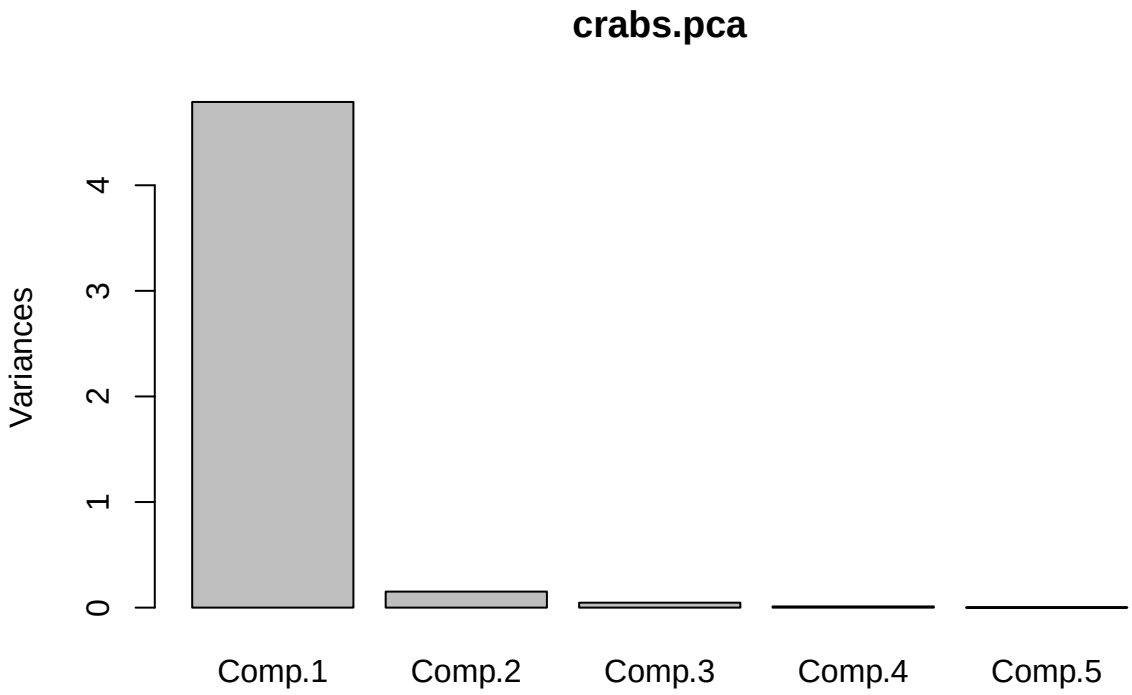
```
data(crabs); head(crabs)
```

```
##   sp sex index   FL  RW   CL   CW  BD
## 1  B   M     1  8.1 6.7 16.1 19.0 7.0
## 2  B   M     2  8.8 7.7 18.1 20.8 7.4
## 3  B   M     3  9.2 7.8 19.0 22.4 7.7
## 4  B   M     4  9.6 7.9 20.1 23.1 8.2
## 5  B   M     5  9.8 8.0 20.3 23.0 8.2
## 6  B   M     6 10.8 9.0 23.0 26.5 9.8
```

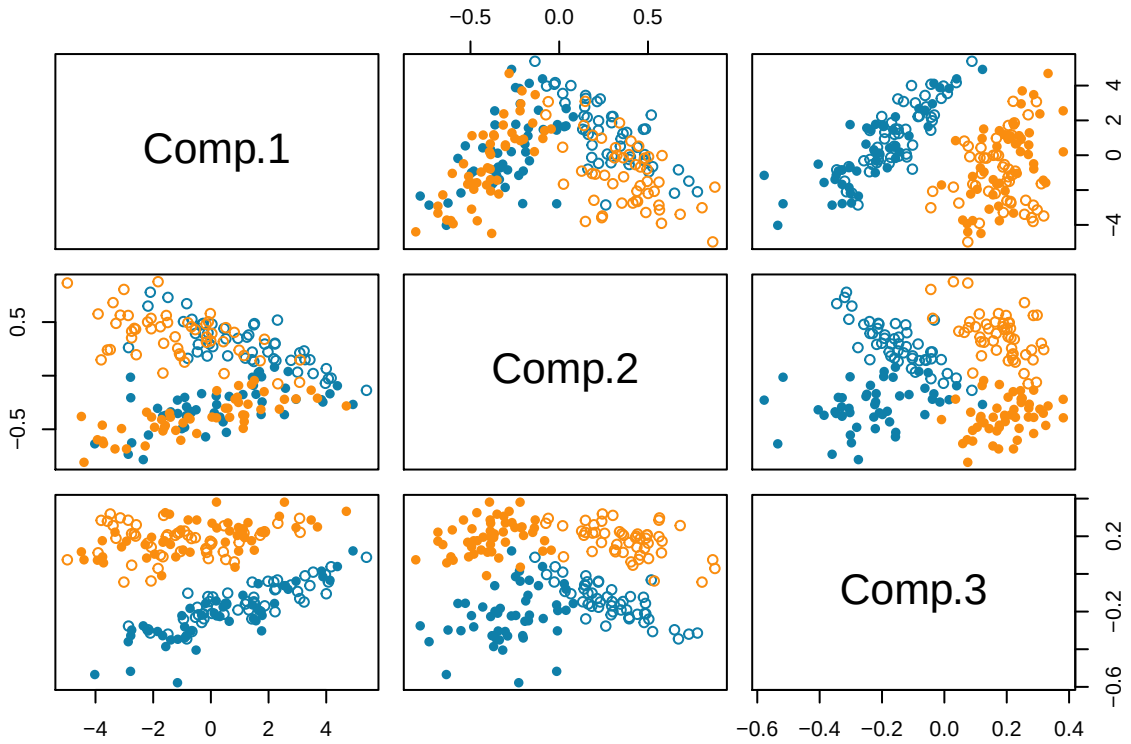
```
crabs.pca <- princomp(crabs[,4:8], cor = TRUE)
loadings(crabs.pca)
```

```
##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## FL -0.452 -0.138  0.531  0.697
## RW -0.428  0.898
## CL -0.453 -0.268 -0.310        -0.792
## CW -0.451 -0.181 -0.653         0.575
## BD -0.451 -0.264  0.443 -0.707  0.176
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

```
plot(crabs.pca)
```




```
num_cols <- c("FL","RW","CL","CW","BD")
ccol <- function(sp) ifelse(sp=="B","#0f7fa9","#fa8d0f")
cpch <- function(sx) 1+ 15*(crabs$sex=="M")
pairs(crabs.pca$scores[,1:3], col=ccol(crabs$sp), pch=cpch(crabs$sex))
```



##Op-
gave 2

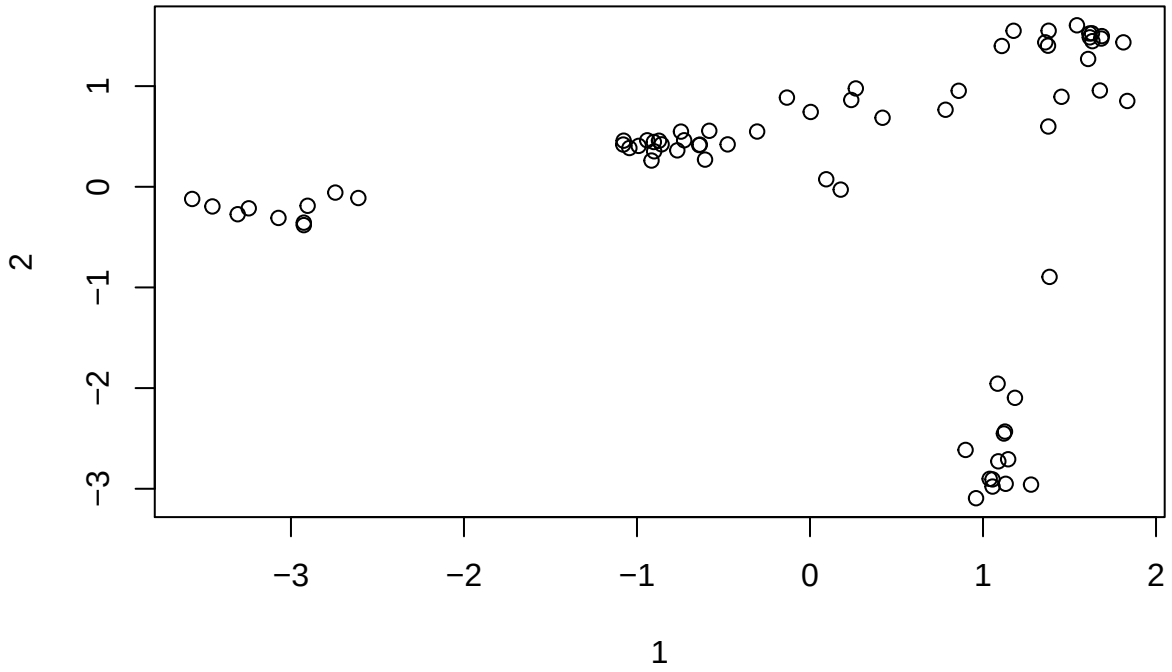
```
aims_freq <- read.csv("aims_freq.csv", header = TRUE)
d <- dist(aims_freq[,4:100])
d.mds <- cmdscale(d)
d.sammon <- sammon(d, y = d.mds)
```

```
## Initial stress        : 0.07568
## stress after  10 iters: 0.04823, magic = 0.092
## stress after  20 iters: 0.03480, magic = 0.213
## stress after  30 iters: 0.02931, magic = 0.150
## stress after  40 iters: 0.02914, magic = 0.500
## stress after  50 iters: 0.02909, magic = 0.500
```

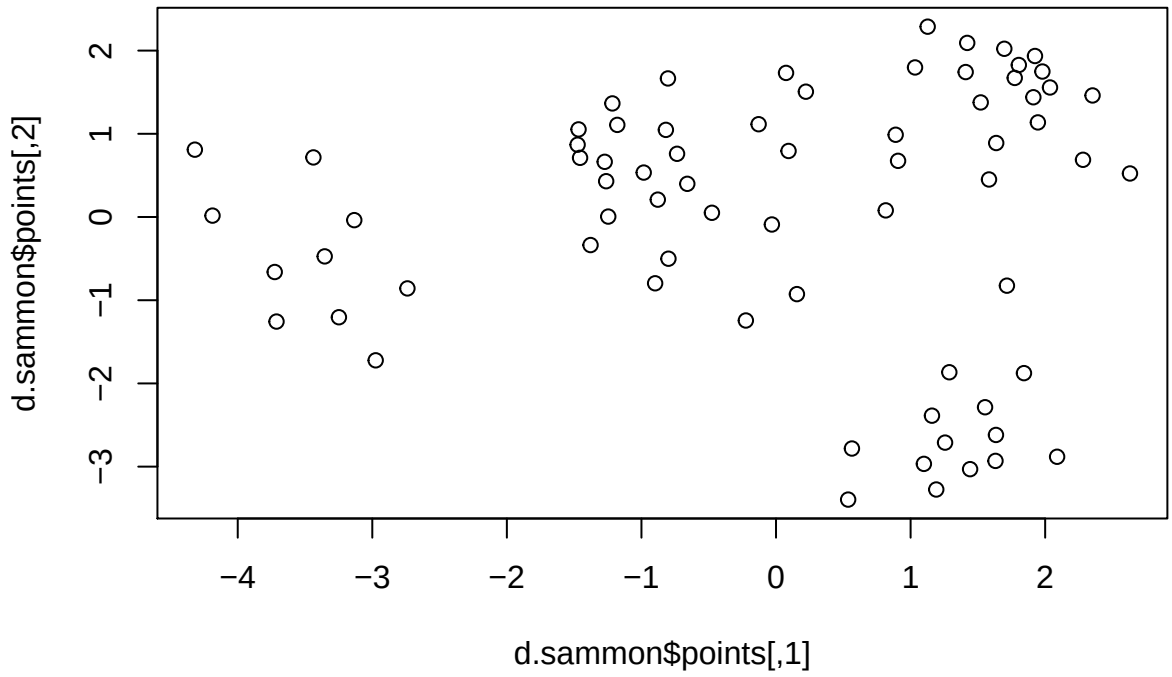
```
d.isomds <- isoMDS(d)
```

```
## initial  value 9.861148
## iter   5 value 7.054416
## final  value 6.902732
## converged
```

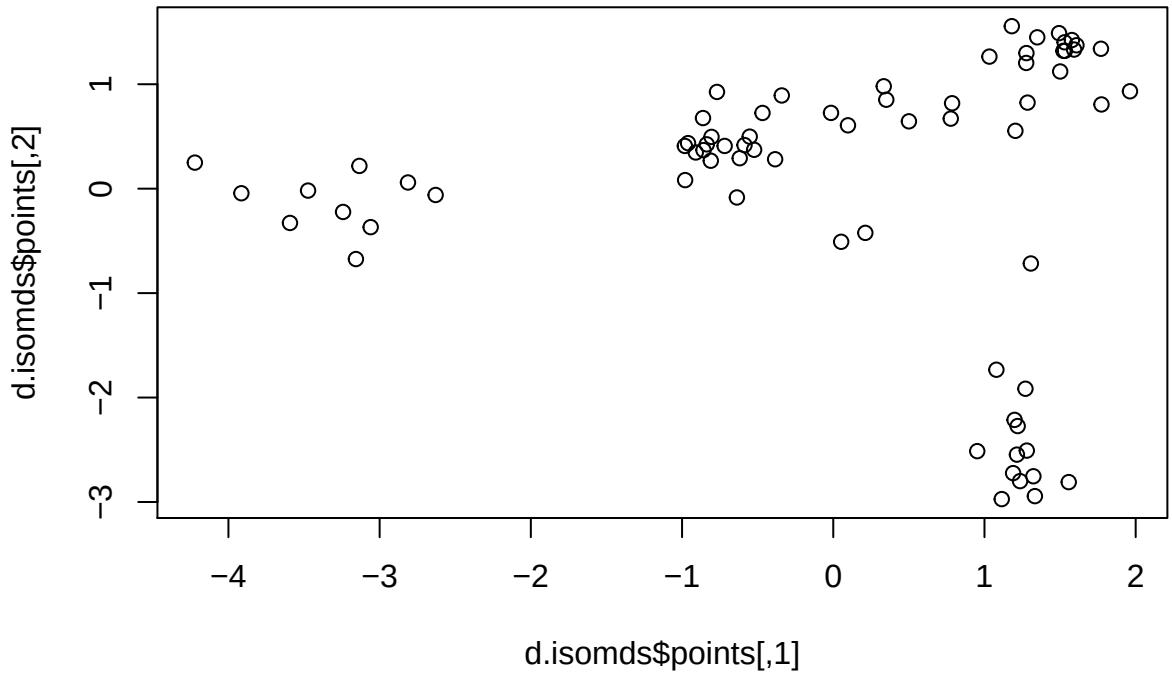
```
plot(d.mds, xlab = "1", ylab = "2")
```




```
plot(d.sammon$points)
```



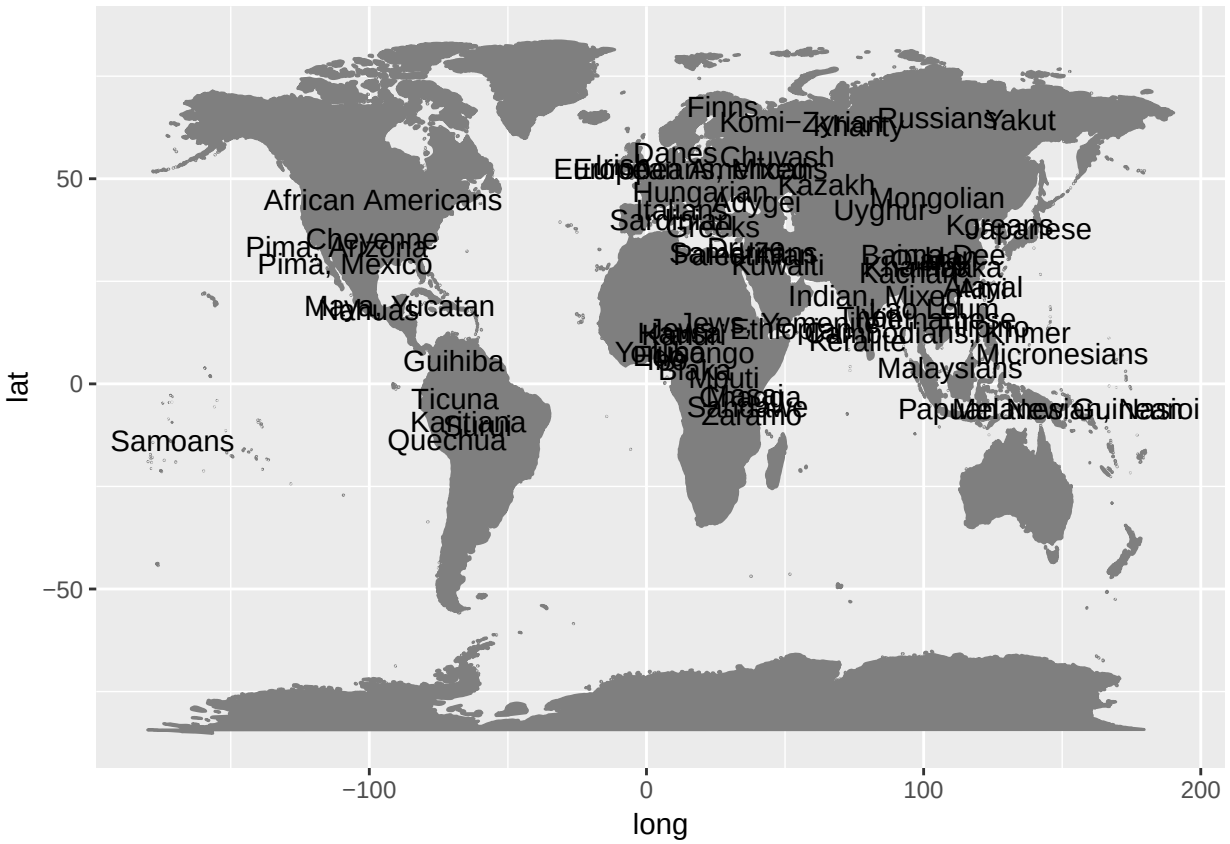

```
plot(d.isomds$points)
```




```
ggplot() + borders("world", colour="gray50", fill="gray50") +
geom_text(data = aims_freq, aes(x = long, y = lat, label = pop))
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```



C4.5

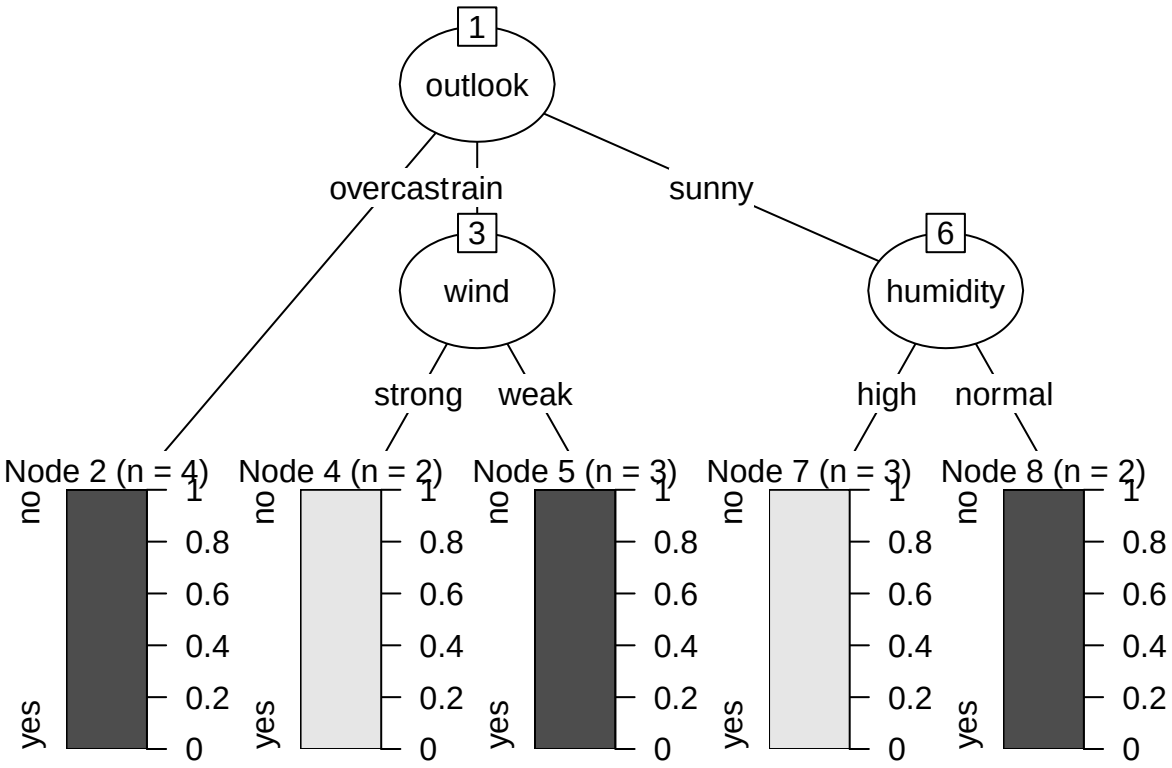
Opgave 3

```
golf <- read.csv("golf.csv", header = TRUE)
```

a)

```
library(C50)
cgolf <- golf[,c("outlook","temperature","humidity","wind","play")]

a <- C5.0(play~.,data=cgolf)
plot(a)
```



Samme med numeriske variable:

```
bgolf <- golf[,c(1,2,4,6,7,8)]
b <- C5.0(play~., data = bgolf, control = C5.0Control(sample = 0.50))
summary(b)
```

```
##
## Call:
## C5.0.formula(formula = play ~ ., data = bgolf, control
##  = C5.0Control(sample = 0.5))
##
##
## C5.0 [Release 2.07 GPL Edition]      Fri Feb 10 13:45:08 2017
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 7 cases (6 attributes) from undefined.data
##
## Decision tree:
##
## outlook = overcast: yes (3)
## outlook in {rain,sunny}: no (4/1)
##
##
## Evaluation on training data (7 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
```

```
##       2    1(14.3%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##       3          (a): class no
##       1     3    (b): class yes
##
##
##  Attribute usage:
##
##  100.00% outlook
##
##
## Evaluation on test data (7 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       2    4(57.1%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##       2          (a): class no
##       4     1    (b): class yes
##
##
## Time: 0.0 secs
```

Opgave 3.B

Se nederst 3.A

Opgave 4

```
sp_sex <- c()
for(i in 1:200){
  if(crabs[i,2]=="M"){
    if(crabs[i,1]=="B"){
      sp_sex <- c(sp_sex,"MB")
    }
    else{
      sp_sex <- c(sp_sex, "MO")
    }
  }
   if(crabs[i,2]=="F"){
    if(crabs[i,1]=="B"){
      sp_sex <- c(sp_sex,"FB")
    }
    else{
```

```
      sp_sex <- c(sp_sex, "FO")
    }
  }
}

crabs2 <- data.frame(crabs, sp_sex)
```

De numeriske variable er:

```
crabs3 <- crabs2[4:9]
e <- C5.0(as.factor(sp_sex)~., data = crabs3)
plot(e)
```