Kanonična korelacija

Uporaba R pri seminarju iz Multivariatne analize

Podiplomski študij Statistika

A. Blejec

andrej.blejec@nib.si

11. junij 2009

Povzetek

⋈

Kazalo

1 Uvod

Dokumenti, ki opisujejo funkcije za generiranje in obvladovanje manjkajočih podatkov si oglejte dokumente iz prejšnjih let. Na razpolago so na spletnem naslovu http://ablejec.nib.si/R/MVA.

2 Korelacijska matrika

```
> R <- read.xls("../data/R.xls")
> R <- as.matrix(R)
> m <- dim(R)[1]
```

```
> m1 <- 3
> m2 <- m - m1
> R
    x1  x2  x3  y1  y2
x1 1.0 0.2 0.4 0.3 0.3
x2 0.2 1.0 0.3 0.4 0.4
x3 0.4 0.3 1.0 0.3 0.4
y1 0.3 0.4 0.3 1.0 0.3
y2 0.3 0.4 0.4 0.3 1.0
```

Vpis matrike

```
> R <- c(1, 0.2, 0.4, 0.3, 0.3, 0.2, 1, 0.3, 0.4, 0.4,
+     0.4, 0.3, 1, 0.3, 0.4, 0.3, 0.4, 0.3, 1, 0.3, 0.3,
+     0.4, 0.4, 0.3, 1)
> R <- matrix(R, 5, 5)
> varNames <- c(paste("x", 1:3, sep = ""), paste("y", 1:2,
+     sep = ""))
> varNames
[1] "x1" "x2" "x3" "y1" "y2"
> dimnames(R) <- list(varNames, varNames)
> R
    x1  x2  x3  y1  y2
x1 1.0 0.2 0.4 0.3 0.3
x2 0.2 1.0 0.3 0.4 0.4
x3 0.4 0.3 1.0 0.3 0.4
y1 0.3 0.4 0.3 1.0 0.3
y2 0.3 0.4 0.4 0.3 1.0
```

```
> det(R)
[1] 0.43118
```

Spremenimo eno od vrednosti

```
> R[4, 5] <- (-0.3)
```

2.1 Trikotniki

```
> R[upper.tri(R)]
 [1]  0.2  0.4  0.3  0.3  0.4  0.3  0.3  0.4  0.4 -0.3
> R[lower.tri(R)] <- NA
> R
   x1  x2  x3  y1   y2
x1  1 0.2 0.4 0.3  0.3
x2 NA 1.0 0.3 0.4  0.4
x3 NA  NA 1.0 0.3  0.4
y1 NA  NA  NA 1.0 -0.3
y2 NA  NA  NA  NA  1.0
```

Takole lahko preslikamo zgornji trikotnik v spodnjega

```
> R[lower.tri(R)] <- t(R)[lower.tri(R)]
> R
    x1  x2  x3   y1   y2
x1 1.0 0.2 0.4  0.3  0.3
x2 0.2 1.0 0.3  0.4  0.4
x3 0.4 0.3 1.0  0.3  0.4
y1 0.3 0.4 0.3  1.0 -0.3
y2 0.3 0.4 0.4 -0.3  1.0
```

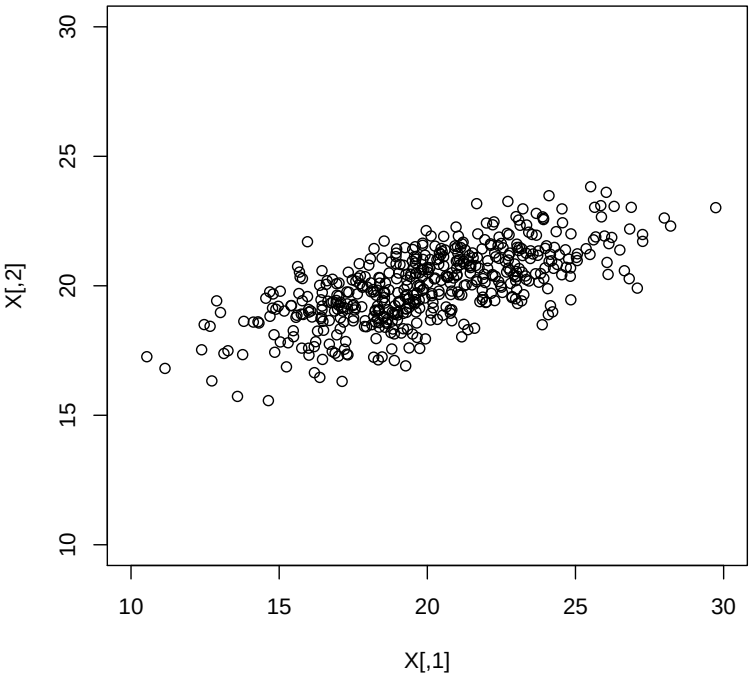
Funkcij za preverjanje simetričnosti

```
> is.symmetric <- function(x) {
+     identical(x, t(x))
+ }
> is.symmetric(R)
[1] TRUE
> U <- R
> U[1, 2] <- 100
> U
    x1    x2  x3   y1   y2
x1 1.0 100.0 0.4  0.3  0.3
x2 0.2   1.0 0.3  0.4  0.4
x3 0.4   0.3 1.0  0.3  0.4
y1 0.3   0.4 0.3  1.0 -0.3
y2 0.3   0.4 0.4 -0.3  1.0
> is.symmetric(U)
[1] FALSE
```

Narediti bi bilo dobro funkcijo (če je še ni??) za spremebo zgornje/spodnje trikotne matrike v simetrično.

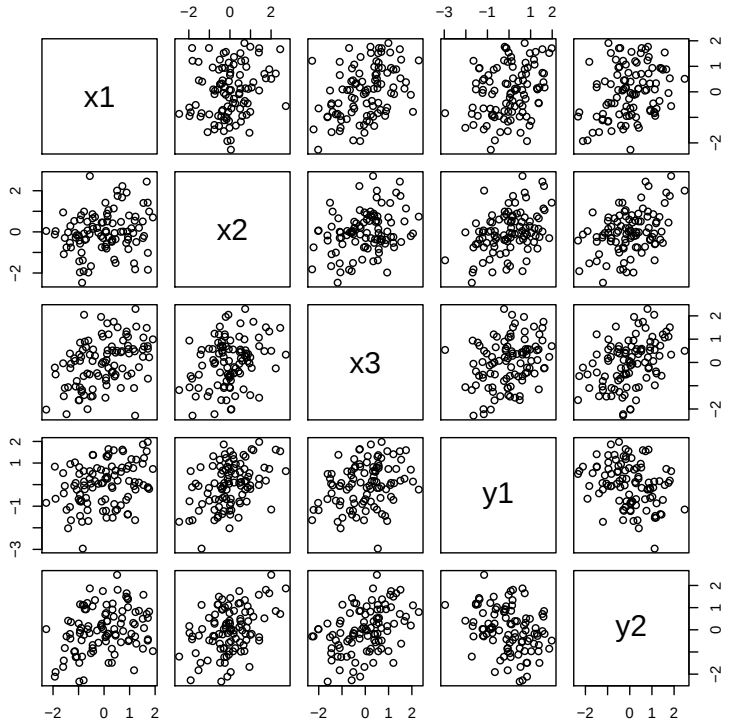
3 Generiranje podatkov

```
> library(MASS)
> (Sigma <- matrix(c(10, 3, 3, 2), 2, 2))
     [,1] [,2]
[1,]   10    3
[2,]    3    2
> X <- mvrnorm(50, c(20, 20), Sigma, emp = T)
> cov(X)
     [,1] [,2]
[1,]   10    3
[2,]    3    2
> X1 <- mvrnorm(500, c(20, 20), Sigma)
> cov(X1)
          [,1]     [,2]
[1,] 10.351849 3.216887
[2,]  3.216887 2.007822
> X <- mvrnorm(500, c(20, 20), Sigma, emp = T)
> plot(X, xlim = c(10, 30), ylim = c(10, 30))
```




```
> set.seed(1230)
> X <- mvrnorm(100, mu = rep(0, m), Sigma = R, emp = T)
```

```
> pairs(X)
```

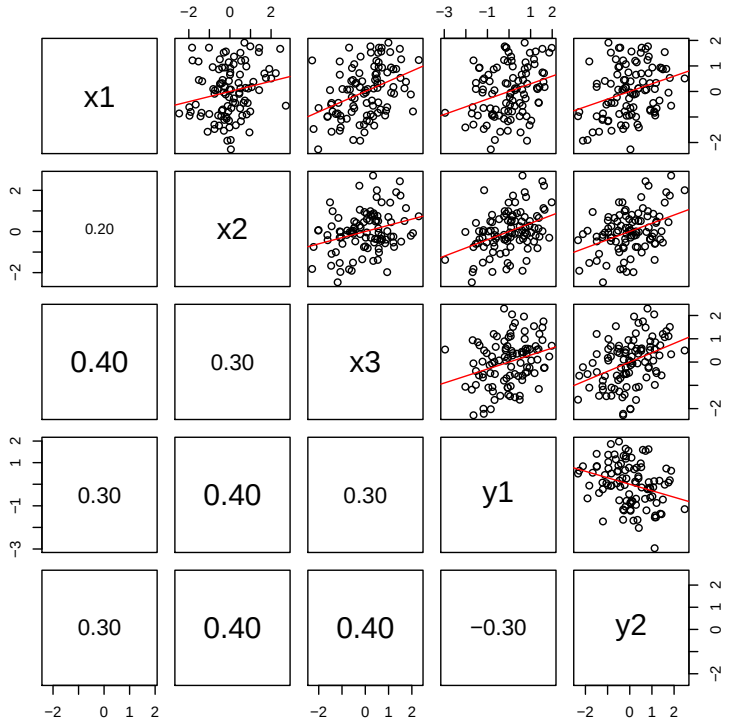


3.1 Dopolnitev prikaza `pairs`

```
> panel.cor <- function(x, y, digits = 2, prefix = "",
+     cex.cor) {
+     usr <- par("usr")
+     on.exit(par(usr))
+     par(usr = c(0, 1, 0, 1))
+     r <- (cor(x, y, use = "pairwise.complete.obs"))
+     txt <- format(c(r, 0.123456789), digits = digits)[1]
+     txt <- paste(prefix, txt, sep = "")
+     if (missing(cex.cor))
+         cex.cor <- 0.8/strwidth(txt)
+     text(0.5, 0.5, txt, cex = cex.cor * abs(r))
+ }


> panel.abline <- function(x, y, col.ab = "red", lwd.ab,
+     ...) {
+     points(x, y)
+     if (missing(lwd.ab))
+         lwd.ab <- 1
+     abline(lm(y ~ x), col = col.ab, lwd = lwd.ab, ...)
+ }


> pairs(X, lower.panel = panel.cor, cex.cor = 5, upper.panel = panel.abline)
```



Shranimo generirane podatke za kasnejšo rabo

```
> data <- X
```

4 Kategorizacija podatkov

Za kategorizacijo uporabimo funkcijo `cut`. Kot argument lahko vpišemo število enako širokih razredov ali pa meje (uporabno za neenako široke razrede).

```
> x <- X[, 1]
> range(x)
[1] -2.267590  1.911885
> table(cut(x, 5))
 (-2.27,-1.43] (-1.43,-0.597] (-0.597,0.241]   (0.241,1.08]
             7             26             26             24
  (1.08,1.92]
            17

> table(cut(x, breaks = c(-3, -2, -1, 1, 2, 3)))
(-3,-2] (-2,-1]  (-1,1]   (1,2]   (2,3]
      1      14      67      18       0
```

Razrez na intervale z enako frekvenco.

```
> quantile(x)
         0%         25%         50%         75%        100%
-2.26759014 -0.81037719  0.06219079  0.77638817  1.91188462
> table(cut(x, quantile(x)))
 (-2.27,-0.81] (-0.81,0.0622] (0.0622,0.776]   (0.776,1.91]
            24             25             25             25

> quantile(x)
         0%         25%         50%         75%        100%
-2.26759014 -0.81037719  0.06219079  0.77638817  1.91188462
> table(cut(x, quantile(x, c(0.01, 0.05, 0.25, 0.75, 0.95,
+     0.99))))
(-1.93,-1.54] (-1.54,-0.81] (-0.81,0.776]  (0.776,1.67]   (1.67,1.77]
            4            20            50            20             4
```

Razrez vseh spremenljivk v matriki.

```
> Y <- cut(X, 5)
> str(Y)
 Factor w/ 5 levels "(-2.97,-1.83]",..: 3 3 3 2 3 2 3 2 2 3 ...
> Y <- as.numeric(Y)
> dim(Y) <- dim(X)
> head(Y)
     [,1] [,2] [,3] [,4] [,5]
[1,]    3    4    2    3    3
[2,]    3    1    4    2    3
[3,]    3    3    5    4    3
[4,]    2    3    2    3    2
[5,]    3    3    2    4    2
[6,]    2    2    3    2    3

> apply(X, 2, min)
       x1        x2        x3        y1        y2
-2.267590 -2.469638 -2.284331 -2.963278 -2.334583
```

5 Kanonična korelacija v R

V osnovnem paketu **stats** je za kanonično korelacijo predvidena funkcija `cancor`.
Vzemimo primer iz pomoči za `cancor`

```
> pop <- LifeCycleSavings[, 2:3]
> oec <- LifeCycleSavings[, -(2:3)]
> cancor(pop, oec)
$cor
[1] 0.8247966 0.3652762

$xcoef
              [,1]        [,2]
pop15 -0.009110856 -0.03622206
pop75  0.048647514 -0.26031158

$ycoef
             [,1]           [,2]          [,3]
sr   0.0084710221  3.337936e-02 -5.157130e-03
dpi  0.0001307398 -7.588232e-05  4.543705e-06
ddpi 0.0041706000 -1.226790e-02  5.188324e-02

$xcenter
  pop15   pop75
35.0896  2.2930

$ycenter
       sr       dpi      ddpi
   9.6710 1106.7584    3.7576
```

6 Paket CCA

Za razširitev funkcionalnosti lahko uporabimo funkcije iz paketa **CCA** (**?**).

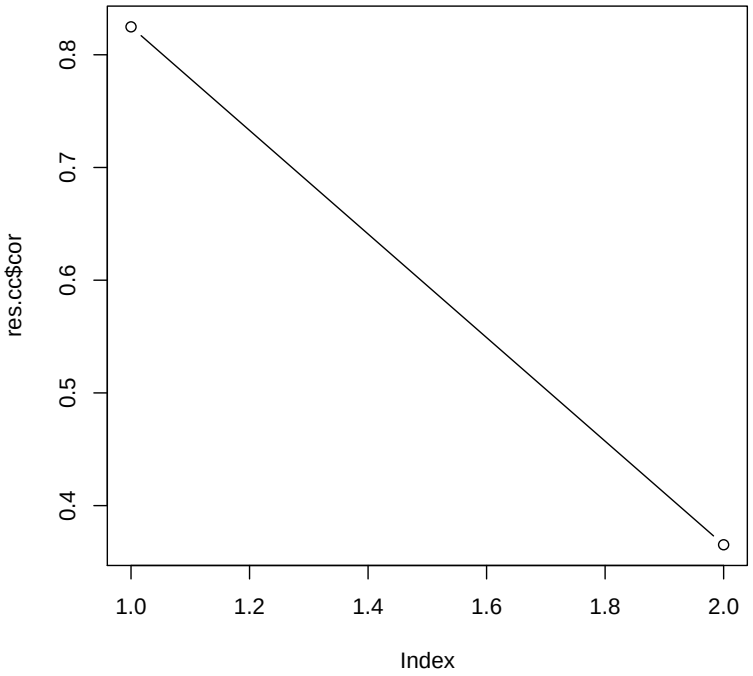
```
> library(CCA)
```

7 Primer iz pomoči za CCA

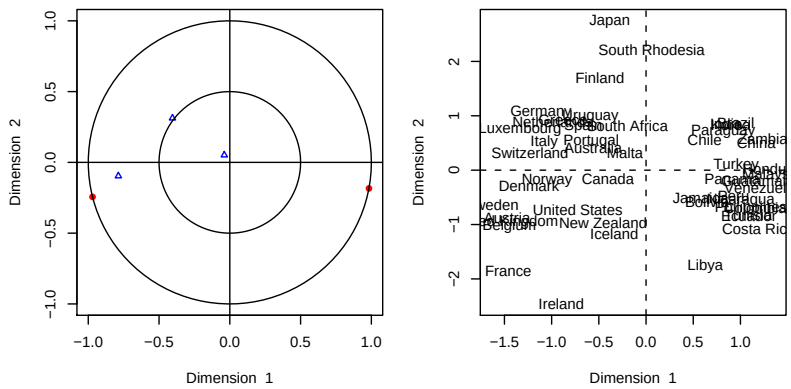
```
> data(nutrimouse)
> X = as.matrix(nutrimouse$gene[, 1:10])
> Y = as.matrix(nutrimouse$lipid)
> X <- pop
> Y <- oec
> res.cc <- cc(X, Y)

> names(res.cc)
[1] "cor"    "names"  "xcoef"  "ycoef"  "scores"

> plot(res.cc$cor, type = "b")
```

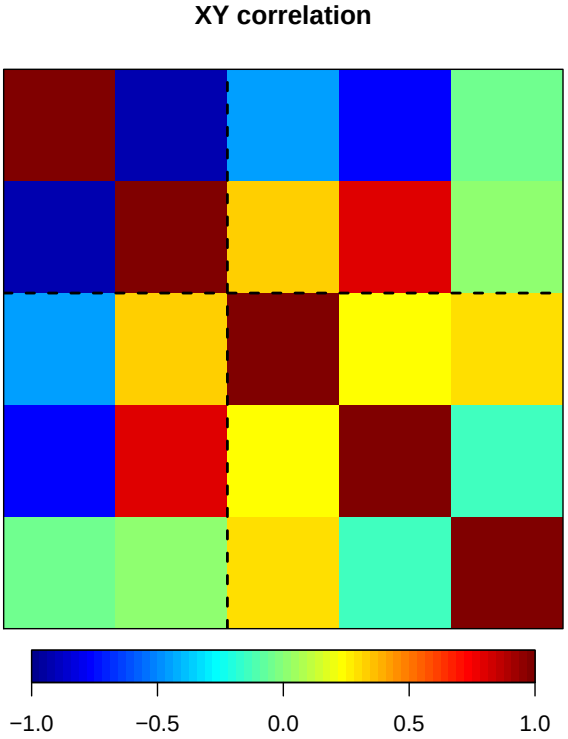

res.cc$cor
0.4
0.5
0.6
0.7
0.8
1.0
1.2
1.4
1.6
1.8
2.0
Index

```
> plt.cc(res.cc)
```

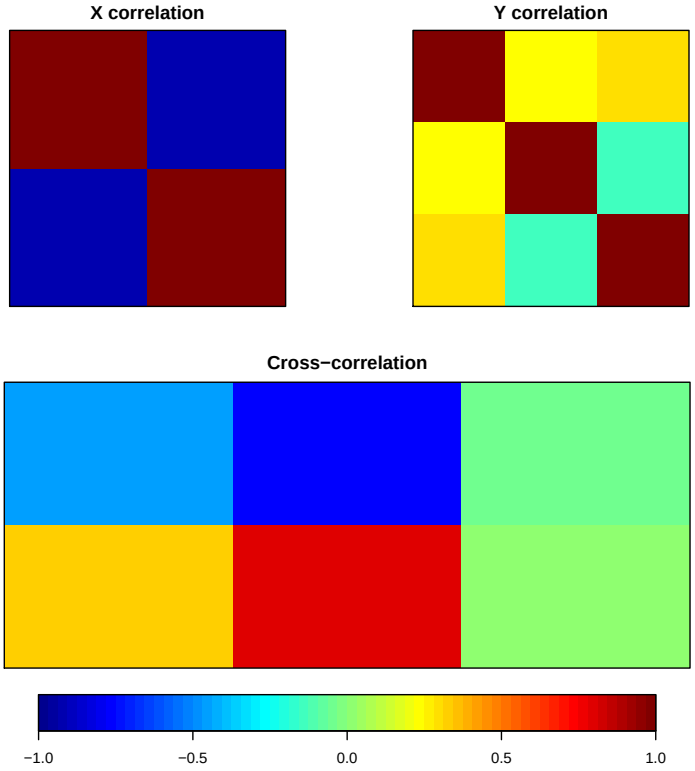

Dimension 2
−1.0
−0.5
0.0
0.5
1.0
Dimension 1
Dimension 2
−2
−1
0
1
2
Japan
South Rhodesia
Finland
Switzerland
Italy
Portugal
Australia
Malta
Chile
China
Turkey
Norway
Canada
Denmark
United States
New Zealand
Iceland
Libya
France
Ireland
−1.5
−1.0
−0.5
0.0
0.5
1.0
Dimension 1

```
> names(matcor(X, Y))
[1] "Xcor"  "Ycor"  "XYcor"

> img.matcor(matcor(X, Y))
```



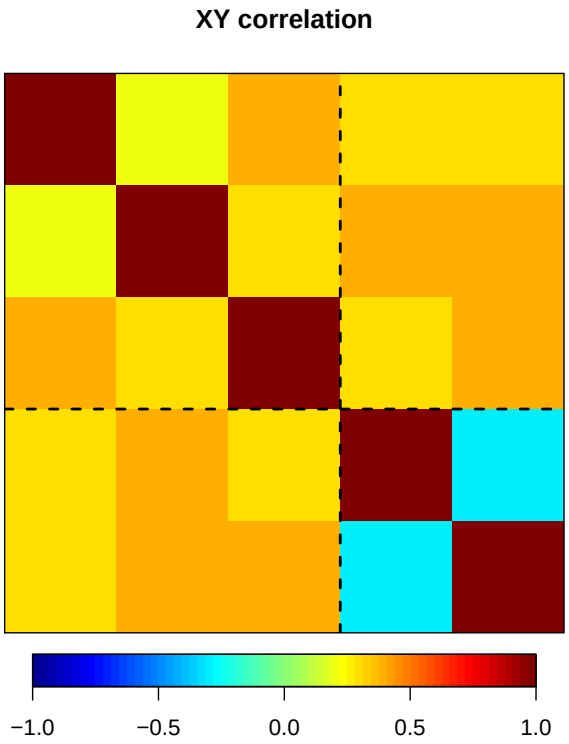

```
> img.matcor(matcor(X, Y), type = 2)
```



8 Simulirani podatki

```
> X <- data[, 1:3]
> Y <- data[, 4:5]
```

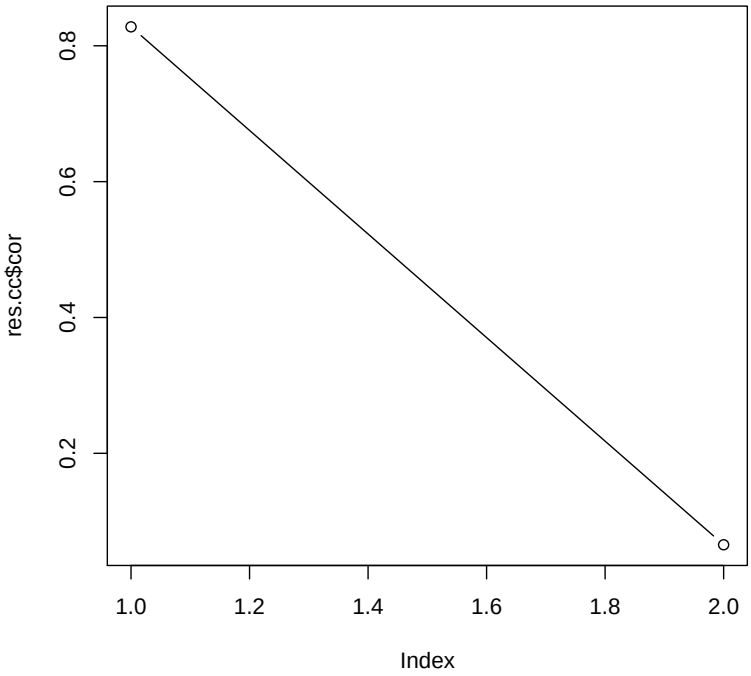
```
> img.matcor(matcor(X, Y))
```



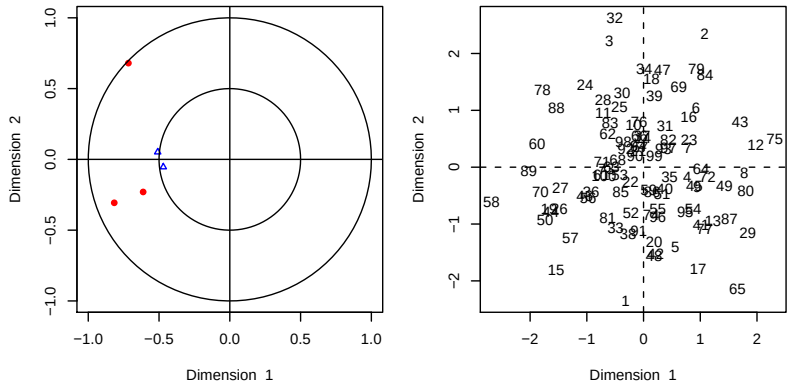

```
> res.cc <- cc(X, Y)
```

```
> names(res.cc)
[1] "cor"    "names"  "xcoef"  "ycoef"  "scores"
```

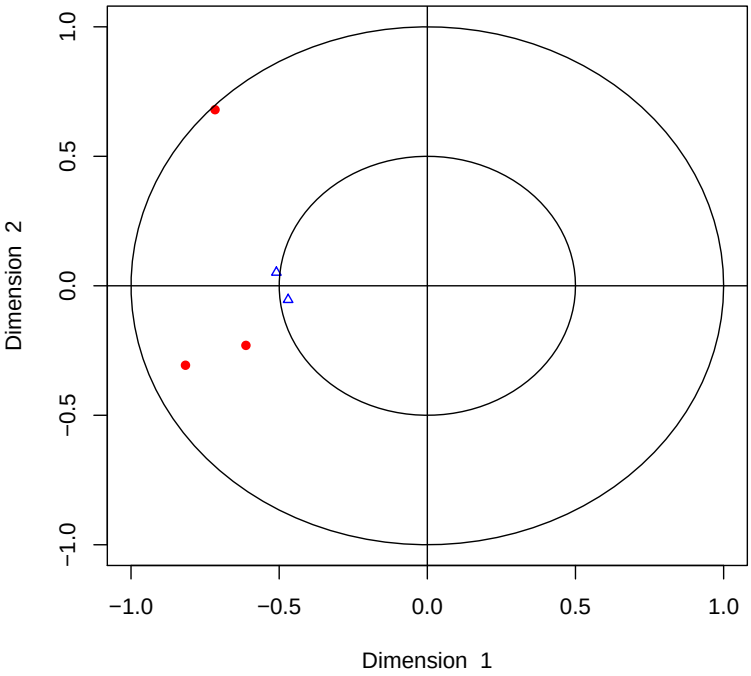
```
> plot(res.cc$cor, type = "b")
```




```
> plt.cc(res.cc)
```




```
> plt.var(res.cc, 1, 2)
```


Dimension 2
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Dimension 1

SessionInfo

Windows XP (build 2600) Service Pack 3

- R version 2.8.0 (2008-10-20), `i386-pc-mingw32`
- Locale: `LC_COLLATE=Slovenian_Slovenia.1250;LC_CTYPE=Slovenian_Slovenia.1250;LC_MON`
- Base packages: base, datasets, graphics, grDevices, methods, splines, stats, utils
- Other packages: CCA 1.2, fda 2.1.2, fields 5.02, Hmisc 3.5-2, MASS 7.2-46, spam 0.15-4, xlsReadWrite 1.3.2, zoo 1.5-5
- Loaded via a namespace (and not attached): cluster 1.11.13, grid 2.8.0, lattice 0.17-22, tools 2.8.0