CCA exmaples

STAT 32950-24620

Spring 2023 (4/6)

Canonical Correlation Analysis

Example: Olympic records of 33 decathletes

```
Events:
100 meters (100), long jump (long), shotput (poid),
high jump (haut), 400 meters (400),
110-meter hurdles (110), discus throw (disq),
pole vault (perc), javelin (jave),
1500 meters (1500).
```


Check data format

```
library(ade4); data(olympic); str(olympic)
```

```
## List of 2
##  $ tab  :'data.frame':   33 obs. of  10 variables:
##   ..$ 100 : num [1:33] 11.2 10.9 11.2 10.6 11 ...
##   ..$ long: num [1:33] 7.43 7.45 7.44 7.38 7.43 7.72 7.(
##   ..$ poid: num [1:33] 15.5 15 14.2 15 12.9 ...
##   ..$ haut: num [1:33] 2.27 1.97 1.97 2.03 1.97 2.12 2.(
##   ..$ 400 : num [1:33] 48.9 47.7 48.3 49.1 47.4 ...
##   ..$ 110 : num [1:33] 15.1 14.5 14.8 14.7 14.4 ...
##   ..$ disq: num [1:33] 49.3 44.4 43.7 44.8 41.2 ...
##   ..$ perc: num [1:33] 4.7 5.1 5.2 4.9 5.2 4.9 5.7 4.8 4
##   ..$ jave: num [1:33] 61.3 61.8 64.2 64 57.5 ...
##   ..$ 1500: num [1:33] 269 273 263 285 257 ...
##  $ score: num [1:33] 8488 8399 8328 8306 8286 ...
```


Check data observations

```
head(olympic$tab,4)[,1:8]
```

```
##     100 long  poid haut   400   110  disq perc
## 1 11.25 7.43 15.48 2.27 48.90 15.13 49.28  4.7
## 2 10.87 7.45 14.97 1.97 47.71 14.46 44.36  5.1
## 3 11.18 7.44 14.20 1.97 48.29 14.81 43.66  5.2
## 4 10.62 7.38 15.02 2.03 49.06 14.72 44.80  4.9
```

```
tail(olympic$tab,4)[,2:10]
```

```
##    long  poid haut   400   110  disq perc  jave  1500
## 30 7.09 12.94 1.82 49.27 15.56 42.32  4.5 53.50 293.9
## 31 6.22 13.98 1.91 51.25 15.88 46.18  4.6 57.84 295.0
## 32 6.43 12.33 1.94 50.30 15.00 38.72  4.0 57.26 293.7
## 33 7.19 10.27 1.91 50.71 16.20 34.36  4.1 54.94 270.0
```

Partial data summary

```
colMeans(olympic$tab[,1:5]); colMeans(olympic$tab[,6:10])
```

```
##    100   long   poid   haut    400
## 11.196  7.133 13.976  1.983 49.277

##     110    disq    perc    jave    1500
##  15.049  42.354   4.739  59.439 276.038
```

```
summary(olympic$tab[,1:3])
```

```
##       100             long            poid
##  Min.   :10.6   Min.   :6.22   Min.   :10.3
##  1st Qu.:11.0   1st Qu.:7.00   1st Qu.:13.2
##  Median :11.2   Median :7.09   Median :14.1
##  Mean   :11.2   Mean   :7.13   Mean   :14.0
##  3rd Qu.:11.4   3rd Qu.:7.37   3rd Qu.:15.0
##  Max.   :11.6   Max.   :7.72   Max.   :16.6
```

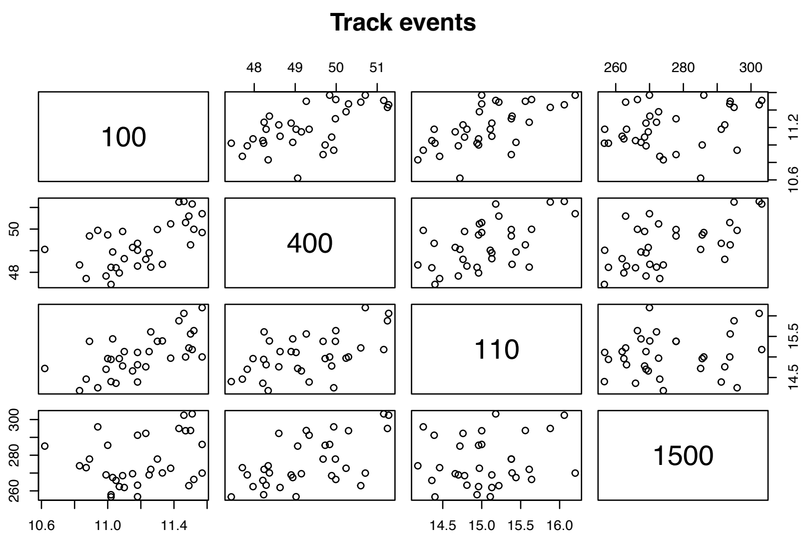

Correlation among variables

```
round(cor(olympic$tab[,1:10]),1)
```

```
##       100 long poid haut  400  110 disq perc jave 1500
## 100   1.0 -0.5 -0.2 -0.1  0.6  0.6  0.0 -0.4 -0.1  0.3
## long -0.5  1.0  0.1  0.3 -0.5 -0.5  0.0  0.3  0.2 -0.4
## poid -0.2  0.1  1.0  0.1  0.1 -0.3  0.8  0.5  0.6  0.3
## haut -0.1  0.3  0.1  1.0 -0.1 -0.3  0.1  0.2  0.1 -0.1
## 400   0.6 -0.5  0.1 -0.1  1.0  0.5  0.1 -0.3  0.1  0.6
## 110   0.6 -0.5 -0.3 -0.3  0.5  1.0 -0.1 -0.5 -0.1  0.1
## disq  0.0  0.0  0.8  0.1  0.1 -0.1  1.0  0.3  0.4  0.4
## perc -0.4  0.3  0.5  0.2 -0.3 -0.5  0.3  1.0  0.3  0.0
## jave -0.1  0.2  0.6  0.1  0.1 -0.1  0.4  0.3  1.0  0.1
## 1500  0.3 -0.4  0.3 -0.1  0.6  0.1  0.4  0.0  0.1  1.0
```

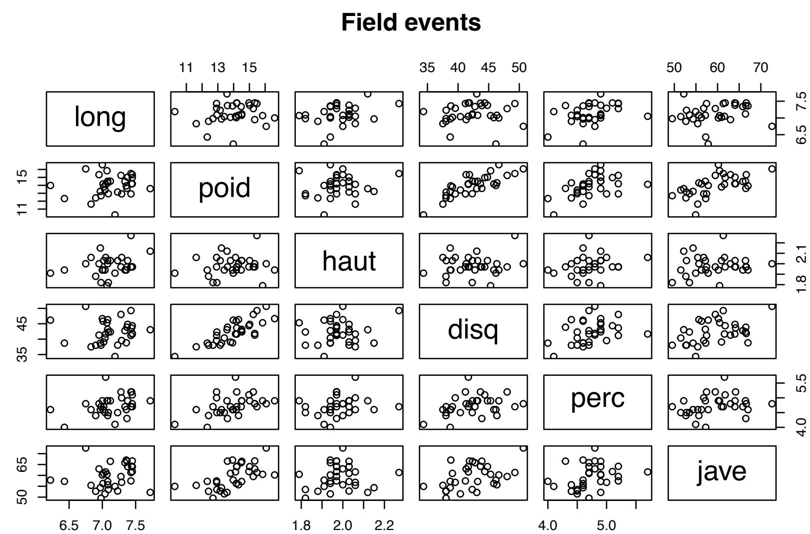

Track event records

```
pairs(olympic$tab[,c(1,5,6,10)],main="Track events")
```



Field event records

```
pairs(olympic$tab[,-c(1,5,6,10)],main="Field events")
```



Objective: Relation of performance in the two event groups

We are interested in the relationship between

- **Performance in track events**
- **Performance in field events**

Group the variables into two vectors:

$X = (X_1, X_2, X_3, X_4)$, a vector vector of track records

$Y = (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6)$, a vector of field records $\times(-1)$

```
X= olympic$tab[,c(1,5,6,10)]
Yold= olympic$tab[,c(2,3,4,7,8,9)]
Y=(-1)*Yold
```


Sample covariance of X (track)

$S_{11} = \hat{\Sigma}_{11}$, a $p \times p$ matrix, $p = 4$.

```
S11 = cov(X)
round(S11,2)
```

```
##       100  400  110   1500
## 100  0.06 0.16 0.08   0.87
## 400  0.16 1.14 0.30   8.58
## 110  0.08 0.30 0.26   0.99
## 1500 0.87 8.58 0.99 186.52
```


Sample covariance of Y (field)

$S_{22} = \hat{\Sigma}_{22}$, a $q \times q$ matrix, $q = 6$.

```
S22 = cov(Y)
round(S22,2)
```

```
##      long poid haut  disq perc  jave
## long 0.09 0.06 0.01  0.05 0.04  0.30
## poid 0.06 1.77 0.02  3.99 0.21  4.38
## haut 0.01 0.02 0.01  0.05 0.01  0.06
## disq 0.05 3.99 0.05 13.83 0.43  9.05
## perc 0.04 0.21 0.01  0.43 0.11  0.50
## jave 0.30 4.38 0.06  9.05 0.50 30.21
```


Sample covariance matrix between X and Y

$$S_{12} = \hat{\Sigma}_{12} \qquad (p \times q \ matrix)$$

```
S12=cov(X,Y)
round(S12,2)
```

```
##      long  poid haut   disq perc  jave
## 100  0.04  0.07 0.00   0.04 0.03  0.09
## 400  0.17 -0.13 0.01  -0.57 0.11 -0.71
## 110  0.07  0.20 0.01   0.21 0.09  0.17
## 1500 1.64 -4.89 0.15 -20.43 0.14 -7.23
```

Sample covariance matrix between Y and X

$$S_{21} = \hat{\Sigma}_{21} = \hat{\Sigma}'_{12} = S'_{12} \qquad (q \times p \ matrix)$$

```
round(cov(Y,X),2)
```

```
##       100   400  110   1500
## long 0.04  0.17 0.07   1.64
## poid 0.07 -0.13 0.20  -4.89
## haut 0.00  0.01 0.01   0.15
## disq 0.04 -0.57 0.21 -20.43
## perc 0.03  0.11 0.09   0.14
## jave 0.09 -0.71 0.17  -7.23
```


Overall sample covariance matrix and correlation matrix

$$\widehat{Cov}\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

```
round(cor(cbind(X,Y)),1)
```

```
##      100  400 110 1500 long poid haut disq perc jave
## 100  1.0  0.6 0.6  0.3  0.5  0.2  0.1  0.0  0.4  0.1
## 400  0.6  1.0 0.5  0.6  0.5 -0.1  0.1 -0.1  0.3 -0.1
## 110  0.6  0.5 1.0  0.1  0.5  0.3  0.3  0.1  0.5  0.1
## 1500 0.3  0.6 0.1  1.0  0.4 -0.3  0.1 -0.4  0.0 -0.1
## long 0.5  0.5 0.5  0.4  1.0  0.1  0.3  0.0  0.3  0.2
## poid 0.2 -0.1 0.3 -0.3  0.1  1.0  0.1  0.8  0.5  0.6
## haut 0.1  0.1 0.3  0.1  0.3  0.1  1.0  0.1  0.2  0.1
## disq 0.0 -0.1 0.1 -0.4  0.0  0.8  0.1  1.0  0.3  0.4
## perc 0.4  0.3 0.5  0.0  0.3  0.5  0.2  0.3  1.0  0.3
## jave 0.1 -0.1 0.1 -0.1  0.2  0.6  0.1  0.4  0.3  1.0
```


Canonical Correlation Analysis (population CCA)

Find

$$U_1 = a'_1 X, \quad V_1 = b'_1 Y$$

such that for all vectors $a \in \mathbb{R}^4$ and vectors $b \in \mathbb{R}^6$,

$$Corr(U_1, V_1) = \max_{a,b} Corr(a'X, b'Y)$$

Note: Multiples of a or b do not change the correlation.

$\implies$ Imposing necessary constraints (normalizations)

$$Var(U_1) = a'_1 S_{11} a_1 = 1, \qquad Var(V_1) = b'_1 S_{22} b_1 = 1.$$

Review population CCA derivation in class

a_1 is an e-vector of matrix $A = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ w.r.t. e-value ρ_1^{*2}.

b_1 is an e-vector of matrix $B = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ w.r.t. e-value ρ_1^{*2}.

$$Corr(U_1, V_1) = Corr(a'_1 X, b'_1 Y) = \rho_1^*$$

A and B share the same non-zero eigenvalues,

because for $C = \Sigma_{11}^{1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$,

$A = \Sigma_{11}^{-1/2}(CC')\Sigma_{11}^{1/2} \quad \implies \quad A$ and CC' share the same e-values.

$B = \Sigma_{22}^{-1/2}(C'C)\Sigma_{22}^{1/2} \quad \implies \quad A$ and CC' share the same e-values.

$C'C$ and CC' share the same non-zero eigenvalues ordered as

$$\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_r^{*2} \geq 0, \quad r = \min(p, q)$$

Properties of p-by-p matrix CC' and q-by-q matrix $C'C$

- Symmetric $(CC')' = CC'$, $(C'C)' = C'C$
- Positive semi-definite: For any $v \in \mathbb{R}^p$, $w \in \mathbb{R}^q$,

$$v'CC'v = \|C'v\|^2 \geq 0, \quad w'C'Cw = \|Cw\|^2 \geq 0$$

- Share the same non-zero eigenvalues:

$$\textit{If} \quad (CC')v = \lambda v \neq 0, \quad \textit{then} \quad (C'C)w = \lambda w \quad \textit{for} \;\; w = C'v \neq 0$$

$$\textit{If} \quad (C'C)w = \delta w \neq 0 \quad \textit{then} \quad (CC')v = \delta v \quad \textit{for} \;\; v = Cw \neq 0$$

- Have the same rank $\leq r = \min(p, q)$
- Their common non-zero eigenvalues can be ordered as

$$\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_r^{*2} \geq 0$$

- The above results are related to the Singular Value Decomposition of matrices.

Canonical correlation pairs (population CCA)

We may further derive more canonical variates

$$(U_i, V_i) = (a_i'X, b_i'Y)$$

with

$$Cor(U_i, V_i) = \rho_i^*$$

for $\quad i = 1, \cdots, r$, with the properties

- $Corr(U_i, U_j) = 0$,
- $Corr(V_i, V_j) = 0$
- $Corr(U_i, V_j) = 0, \quad \textit{if} \;\; i \neq j$.

Discussion: What should be the pattern of the covariance matrix of all canonical variates?

Verify eigenvalue-eigenvector structure of sample estimates

$$\hat{A} = S_{11}^{-1}S_{12}S_{22}^{-1}S_{21} \quad \text{and} \quad \hat{B} = S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$$

```
A=solve(S11)%*%S12%*%solve(S22)%*%t(S12)
eigen(A)
```

```
## eigen() decomposition
## $values
## [1] 0.54252 0.25775 0.18827 0.05094
##
## $vectors
##          [,1]     [,2]     [,3]      [,4]
## [1,] -0.72944 -0.26290  0.17699 -0.9549978
## [2,]  0.04499  0.33960 -0.83082  0.0676619
## [3,] -0.68235 -0.90274  0.52540  0.2887910
## [4,] -0.01695  0.02474  0.04868  0.0009303
```


What is the norm normalization of the eigenvectors?

Comparison: norm normalization use here vs CCA (later)

```
a1 = eigen(A)$vector[,1]
t(a1)%*%S11%*%a1
```

```
##        [,1]
## [1,] 0.2881
```

```
t(a1)%*%a1   # = sum((a1^2) = 1
```

```
##      [,1]
## [1,]    1
```

```
sqrt(round(eigen(A)$value,3))     # sqrt(lambda)=rho
```

```
## [1] 0.7369 0.5079 0.4336 0.2258
```

Do A and B have common > 0 eigenvalues?

```
B=solve(S22)%*%t(S12)%*%solve(S11)%*%S12
eigen(B); sqrt(round(eigen(B)$value,3)) #sqrt(lambda)=rho
```

```
## eigen() decomposition
## $values
## [1]  5.425e-01  2.578e-01  1.883e-01  5.094e-02  1.556e-
##
## $vectors
##            [,1]      [,2]      [,3]       [,4]     [,5]
## [1,]  0.67083  0.453859  0.141926  0.2871489 -0.02480  (
## [2,]  0.14297 -0.227601 -0.131795  0.0620849 -0.40727 -(
## [3,]  0.64547 -0.852181 -0.937627 -0.9137452  0.78403  (
## [4,] -0.05468 -0.004922  0.056165 -0.0001737  0.06927  (
## [5,]  0.33119 -0.125720  0.283033 -0.2805955  0.45193 -(
## [6,] -0.01544  0.012930 -0.008962  0.0051759  0.09887  (
```

```
## [1] 0.7369 0.5079 0.4336 0.2258 0.0000 0.0000
```


CCA by R

```
attributes(cancor(X,Y))
```

```
## $names
## [1] "cor"     "xcoef"    "ycoef"    "xcenter" "ycenter"
```

```
#"cor" "xcoef" "ycoef" "xcenter" "ycenter"
```

```
cancor(X,Y)$cor  # comp. w/ A, B root e-values
```

```
## [1] 0.7366 0.5077 0.4339 0.2257
```


The canonical correlations

$$
\begin{aligned}
\Rightarrow \quad \rho_1^* &= Corr(U_1, V_1) = 0.74,\\
\rho_2^* &= Corr(U_2, V_2) = 0.51,\\
\rho_3^* &= Corr(U_3, V_3) = 0.44,\\
\rho_4^* &= Corr(U_4, V_4) = 0.23.
\end{aligned}
$$

ρ_i^{*2} are the eigenvalues of matrices

$\hat{A} = S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}$ and $\hat{B} = S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$

and CC' and $C'C$

Coefficient vectors b_i of the canonical variates

```
cancor(X,Y)$ycoef
```

```
##            [,1]      [,2]      [,3]       [,4]      [,5]
## long -0.374504 -0.246936 -0.152525  0.3628867 -0.245003
## poid -0.079817  0.123834  0.141638  0.0784603  0.045620
## haut -0.360350  0.463656  1.007652 -1.1547527 -1.091883
## disq  0.030528  0.002678 -0.060359 -0.0002195 -0.038856
## perc -0.184896  0.068402 -0.304171 -0.3546048  0.357047
## jave  0.008621 -0.007035  0.009631  0.0065411  0.008231
```

$$
\Rightarrow \quad b_1 = \begin{bmatrix} -0.374504 \\ -0.079817 \\ -0.360350 \\ 0.030528 \\ -0.184896 \\ 0.008621 \end{bmatrix}, \quad b_2 = \cdots, \quad b_3 = \cdots, \quad b_4 = \cdots
$$

Coefficient vectors a_i of the canonical variates, their norm

```
cancor(X,Y)$xcoef
```

```
##             [,1]       [,2]     [,3]       [,4]
## 100  -0.240222  0.075806  0.05036  0.9973029
## 400   0.014818 -0.097923 -0.23637 -0.0706592
## 110  -0.224716  0.260302  0.14948 -0.3015841
## 1500 -0.005583 -0.007135  0.01385 -0.0009715
```

```
a1 = cancor(X,Y)$xcoef[,1]   # sum(a1^2) #.1085
round((t(a1)%*%cov(X)%*%a1)*length(X[,1]),1)  #1.031
```

```
##      [,1]
## [1,]    1
```

$$a_1 = \begin{bmatrix} -0.240222 \\ 0.014818 \\ -0.224716 \\ -0.005583 \end{bmatrix}, \quad a_2 = \begin{bmatrix} 0.075806 \\ -0.097923 \\ 0.260302 \\ -0.007135 \end{bmatrix}, a_3 = \cdots, a_4 = \cdots$$

Centering and correlation

```
cancor(X,Y)$xcenter
```

```
##    100    400    110   1500
##  11.20  49.28  15.05 276.04
```

```
cancor(X,Y)$ycenter
```

```
##    long    poid    haut    disq    perc    jave
##  -7.133 -13.976  -1.983 -42.354  -4.739 -59.439
```

Discussion:

The role of center; the relation of correlation and centering.

Pairs of canonical variables

$U_1 = -0.240(100m) - 0.015(400m) - 0.225(110h) - 0.006(1500m)$

$V_1 = -0.375(long) - 0.080(poid) - 0.360(haut) + 0.031(disc)$
$-0.185(perc) + 0.009(jave)$

$U_2 = 0.076(100m) - 0.098(400m) - 0.260(110h) - 0.007(1500m)$

$V_2 = -0.247(long) + ...$

$U_3 = ...$

$V_3 = ...$

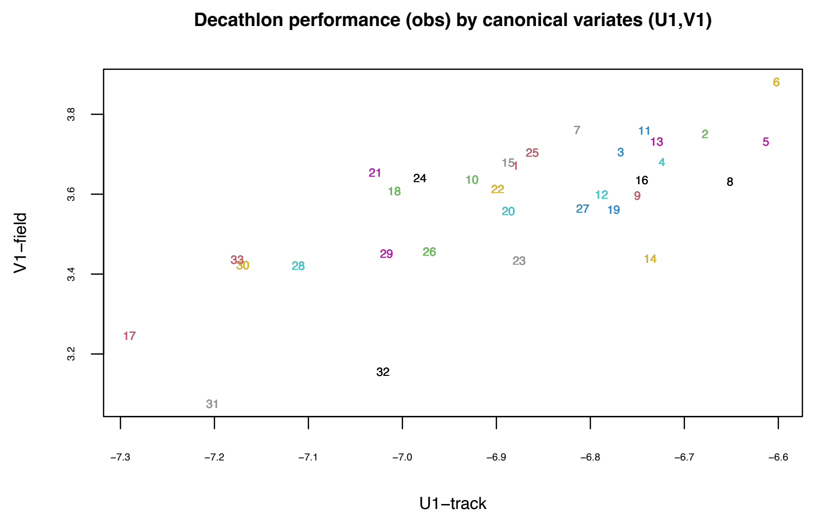
$U_4 = ...$

$V_4 = ...$

Observations in canonical variate coordinates (U1,V1)

```
U=as.matrix(X)%*%cancor(X,Y)$xcoef
V=as.matrix(Y)%*%cancor(X,Y)$ycoef
plot(U[,1],V[,1],type="n",xlab="U1-track",
     ylab="V1-field",cex.lab=.8,cex.axis=.5)
text(U[,1],V[,1],labels=row.names(X),cex=.6)
text(U[,1],V[,1],labels=row.names(X),cex=.6,col=2:34)
title(cex.main=.9, main=
"Decathlon performance (obs) by canonical variates (U1,V1)'
```

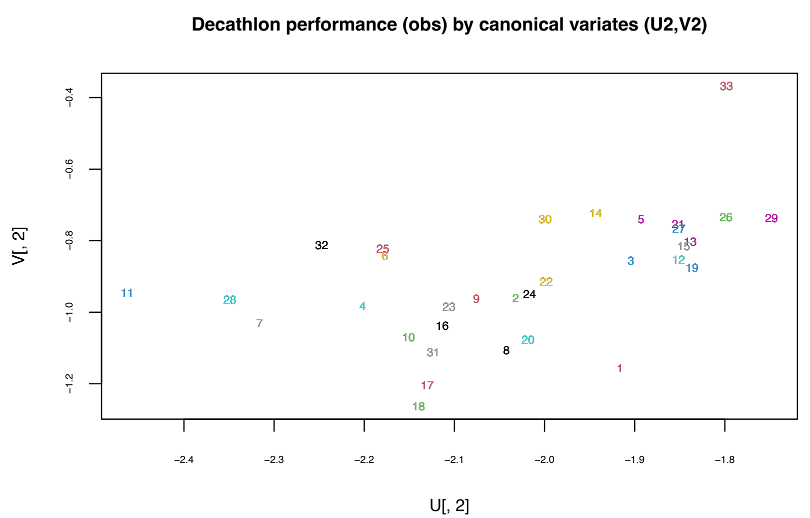
Plot observations in canonical variate coordinates (U1,V1)



Observations in canonical variate coordinates (U2,V2)

```
plot(U[,2],V[,2],type="n",cex.lab=.8,cex.axis=.5)
text(U[,2],V[,2],labels=row.names(X),cex=.6)
text(U[,2],V[,2],labels=row.names(X),cex=.6,col=2:34)
title(,cex.main=.9, main=
"Decathlon performance (obs) by canonical variates (U2,V2)'
```


Plot observations in canonical variate coordinates (U2,V2)



Original variables and cononical variates

What are the relative positions of the original variables under the (new) canonical variables?

In the (U_1, U_2) plane, the track variables

$X_1(100m) = (-0.24, 0.076)$,

$X_2(400m) = (-0.015, -0.098), \quad \cdots$

In the (V_1, V_2) plane, the field variables

$Y_1(long) = (-0.375, -0.247)$,

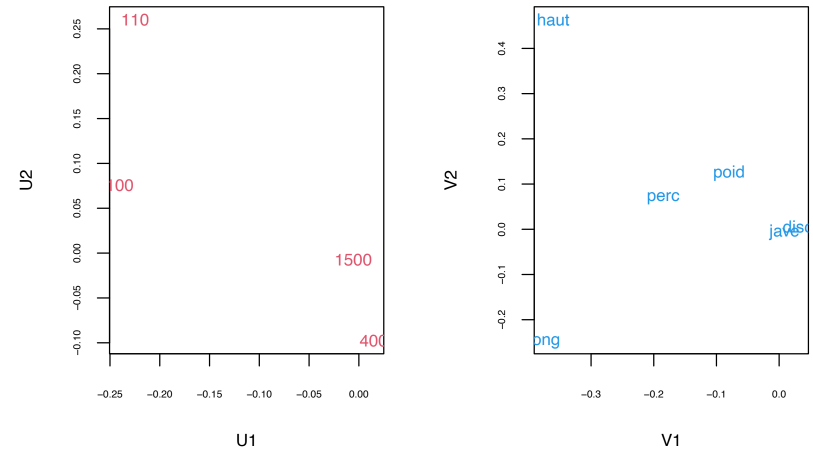
$\cdots$,

$Y_6(jave) = \cdots$

Variables in canonical var coordinates

```
par(mfrow=c(1,2))
plot(cancor(X,Y)$xcoef[,1:2],xlab="U1",
     ylab="U2", type="n",cex.lab=.8,cex.axis=.5)
text(cancor(X,Y)$xcoef[,1:2],
     labels=colnames(X),cex=.8,col=2)
plot(cancor(X,Y)$ycoef[,1:2],xlab="V1", ylab="V2",
     type="n",cex.lab=.8,cex.axis=.5)
text(cancor(X,Y)$ycoef[,1:2],
     labels=colnames(Y),cex=.8,col=4)
```


Plot variables in canonical var coordinates



Correlation matrix of canonical variables

```
round( cor(cbind(U,V)),2)
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,] 1.00 0.00 0.00 0.00 0.74 0.00 0.00 0.00    0     (
##  [2,] 0.00 1.00 0.00 0.00 0.00 0.51 0.00 0.00    0     (
##  [3,] 0.00 0.00 1.00 0.00 0.00 0.00 0.43 0.00    0     (
##  [4,] 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.23    0     (
##  [5,] 0.74 0.00 0.00 0.00 1.00 0.00 0.00 0.00    0     (
##  [6,] 0.00 0.51 0.00 0.00 0.00 1.00 0.00 0.00    0     (
##  [7,] 0.00 0.00 0.43 0.00 0.00 0.00 1.00 0.00    0     (
##  [8,] 0.00 0.00 0.00 0.23 0.00 0.00 0.00 1.00    0     (
##  [9,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00    1     (
## [10,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00    0     1
```


Normalization in our CCA derivation

In our derivation of canonical correlation variable pairs

$$(U_i, V_i) = (a_i' X, b_i' Y),$$

we imposed the constraints

$$a_i' cov(X) a_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

$$b_i' cov(Y) b_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

R command eigen(A) normalization

The command `eigen(A)` normalize the eigenvectors $\tilde{a}_i$ of A by

$$\tilde{a}_i' \tilde{a}_i = 1$$

```
t(eigen(A)$vector)%*%eigen(A)$vector
```

```
##          [,1]     [,2]     [,3]     [,4]
## [1,]  1.0000  0.82262 -0.52582  0.50258
## [2,]  0.8226  1.00000 -0.80177  0.01337
## [3,] -0.5258 -0.80177  1.00000 -0.07347
## [4,]  0.5026  0.01337 -0.07347  1.00000
```


R command cancor(X,Y) normalization

R command `cancor(X,Y)` normalized canonical variates a_i^* by

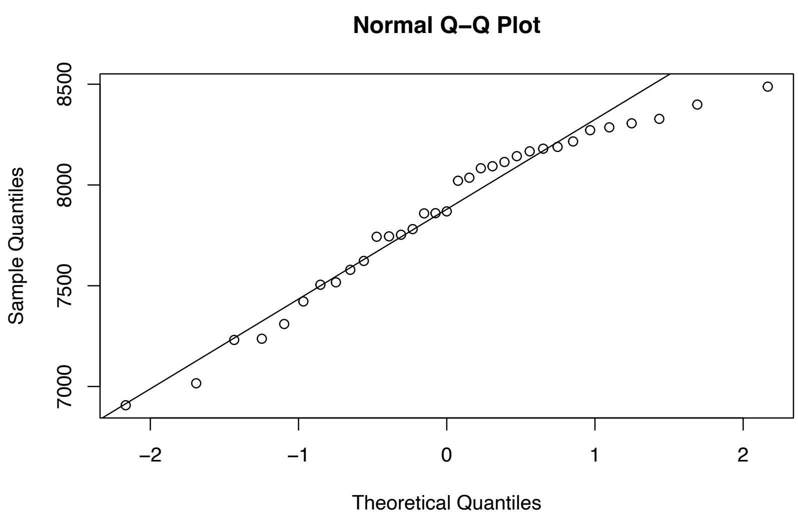
$$n(a_i^{*\prime} cov(X) a_j^*) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

```
(t(cancor(X,Y)$xcoef)%*%cov(X)%*%
   cancor(X,Y)$xcoef)*length(X[,1])
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,]  1.031e+00 -1.717e-16 -2.290e-16  4.293e-17
## [2,] -2.290e-16  1.031e+00  2.290e-16 -1.717e-16
## [3,]  2.290e-16  5.725e-17  1.031e+00 -6.798e-16
## [4,]  0.000e+00  5.009e-17 -3.435e-16  1.031e+00
```

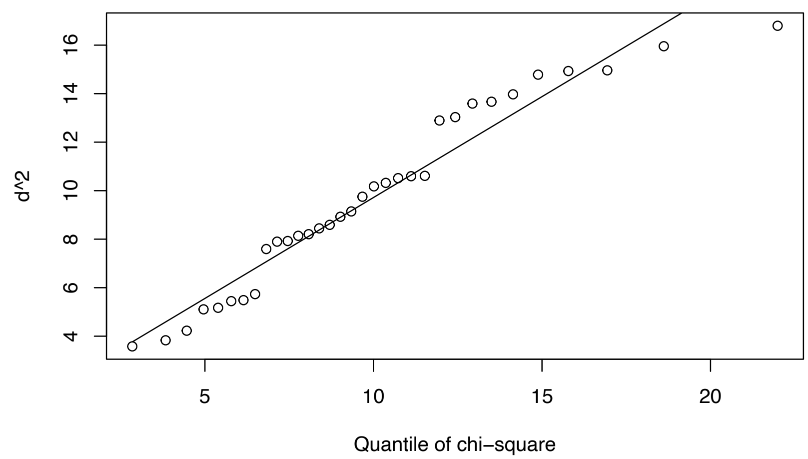

Univariate QQ plot normality check

```
qqnorm(olympic$score); qqline(olympic$score)
```



Multivariate χ^2 plot normality check

```
source("qqchi2.R"); qqchi2(olympic$tab)  #corr coeff=0.97
```




```
## [1] "correlation coefficient:"
## [1] 0.9727
```