Análisis de Componenetes Principales (ACP)

La Selección Natural (en Gorriones)

Estudio R

Octubre de 2017

Contenido

- Generalidades ACP

Contenido

- Generalidades ACP
- Objetivos del ACP

Contenido

- Generalidades ACP
- Objetivos del ACP
- Algunas aplicaciones

Contenido

- Generalidades ACP
- Objetivos del ACP
- Algunas aplicaciones
- Interpretación geométrica

Contenido

- Generalidades ACP
- Objetivos del ACP
- Algunas aplicaciones
- Interpretación geométrica
- Interpretación algebraica

Contenido

- Generalidades ACP
- Objetivos del ACP
- Algunas aplicaciones
- Interpretación geométrica
- Interpretación algebraica
- Ejemplo e implementación en R

Contenido

- Generalidades ACP
- Objetivos del ACP
- Algunas aplicaciones
- Interpretación geométrica
- Interpretación algebraica
- Ejemplo e implementación en R
- Datos de gorriones

Contenido

- Generalidades ACP
- Objetivos del ACP
- Algunas aplicaciones
- Interpretación geométrica
- Interpretación algebraica
- Ejemplo e implementación en R
- Datos de gorriones
- Matriz de correlación

Contenido

- Generalidades ACP
- Objetivos del ACP
- Algunas aplicaciones
- Interpretación geométrica
- Interpretación algebraica
- Ejemplo e implementación en R
- Datos de gorriones
- Matriz de correlación
- Estandarización

Contenido

- Generalidades ACP
- Objetivos del ACP
- Algunas aplicaciones
- Interpretación geométrica
- Interpretación algebraica
- Ejemplo e implementación en R
- Datos de gorriones
- Matriz de correlación
- Estandarización
- Graficación (Biplot)

Generalidades ACP

- El ACP se constituye como una técnica exploratoria, en etapas iniciales del análisis de los datos.

Generalidades ACP

- El ACP se constituye como una técnica exploratoria, en etapas iniciales del análisis de los datos.
- Su propósito fundamental es reducir la dimensionalidad de un conjunto de datos.

Generalidades ACP

- El ACP se constituye como una técnica exploratoria, en etapas iniciales del análisis de los datos.
- Su propósito fundamental es reducir la dimensionalidad de un conjunto de datos.
- Transforma las variables originales en un conjunto de variables más pequeñas, las cuales son combinaciones lineales de las variables originales y retienen la mayor parte de la variabilidad presente en las variables originales.

Generalidades ACP

- El ACP se constituye como una técnica exploratoria, en etapas iniciales del análisis de los datos.
- Su propósito fundamental es reducir la dimensionalidad de un conjunto de datos.
- Transforma las variables originales en un conjunto de variables más pequeñas, las cuales son combinaciones lineales de las variables originales y retienen la mayor parte de la variabilidad presente en las variables originales.
- Convierte un conjunto de variables posiblemente correlacionadas, en un conjunto de variables sin correlación lineal, denominadas *Componentes Principales (CP).*

Objetivos del ACP

- Obtener nuevas variables (componentes principales) que expresen la información contenida en los datos originales.

Objetivos del ACP

- Obtener nuevas variables (componentes principales) que expresen la información contenida en los datos originales.
- Reducir la dimensionalidad de los datos.

Objetivos del ACP

- Obtener nuevas variables (componentes principales) que expresen la información contenida en los datos originales.
- Reducir la dimensionalidad de los datos.
- Eliminar variables (si existe la posibilidad) que tengan poco aporte al objetivo del estudio.

Objetivos del ACP

- Obtener nuevas variables (componentes principales) que expresen la información contenida en los datos originales.
- Reducir la dimensionalidad de los datos.
- Eliminar variables (si existe la posibilidad) que tengan poco aporte al objetivo del estudio.
- Reconocer intuitivamente posibles patrones presentes en los datos.

Objetivos del ACP

- Obtener nuevas variables (componentes principales) que expresen la información contenida en los datos originales.
- Reducir la dimensionalidad de los datos.
- Eliminar variables (si existe la posibilidad) que tengan poco aporte al objetivo del estudio.
- Reconocer intuitivamente posibles patrones presentes en los datos.
- Contribuir a la interpretación de la información que poseen los datos.

Algunas aplicaciones

- Mecánica

Algunas aplicaciones

- Mecánica
- Genómica

Algunas aplicaciones

- Mecánica
- Genómica
- Inteligencia Artificial

Algunas aplicaciones

- Mecánica
- Genómica
- Inteligencia Artificial
- Economia

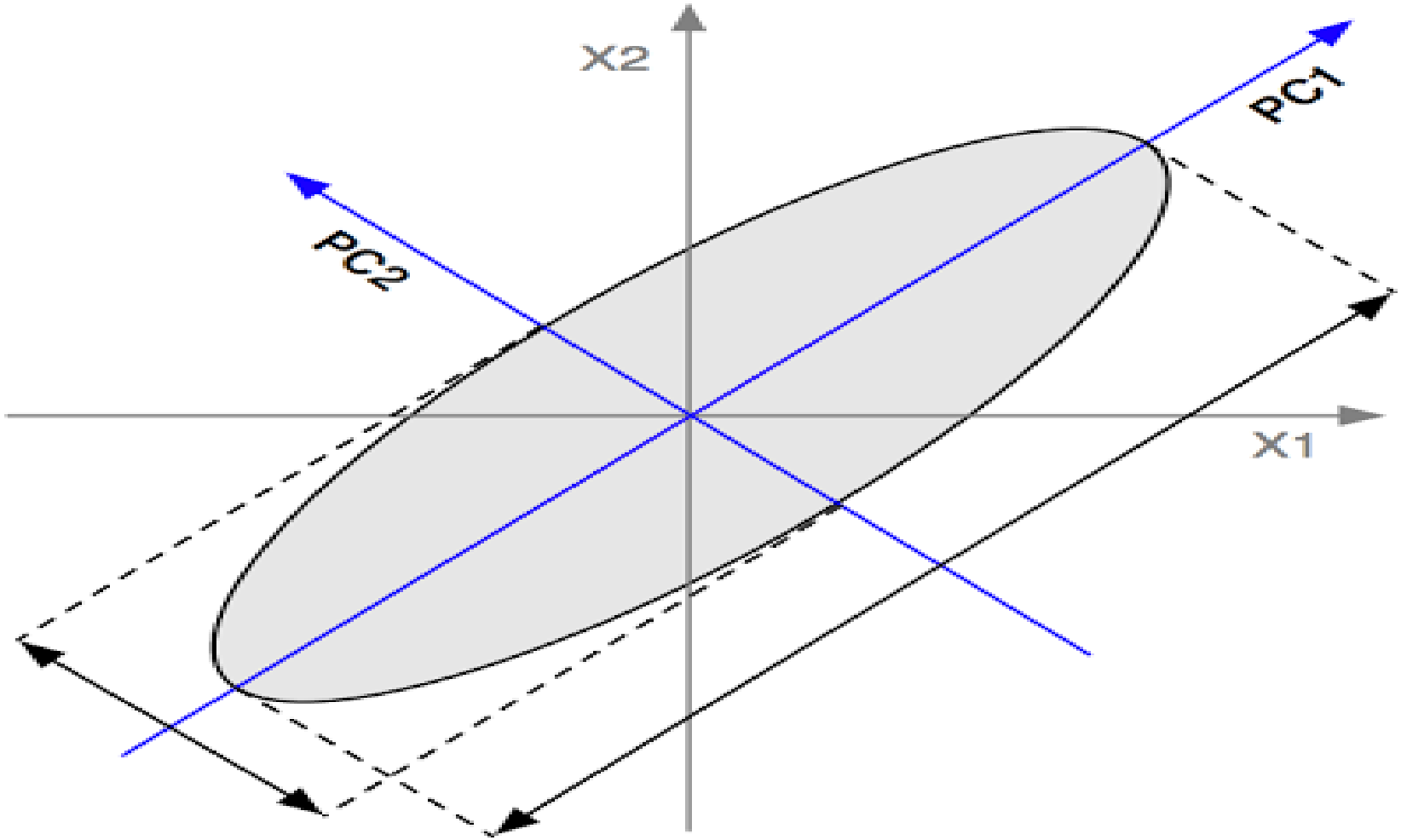
Algunas aplicaciones

- Mecánica
- Genómica
- Inteligencia Artificial
- Economia
- Estudios sociales

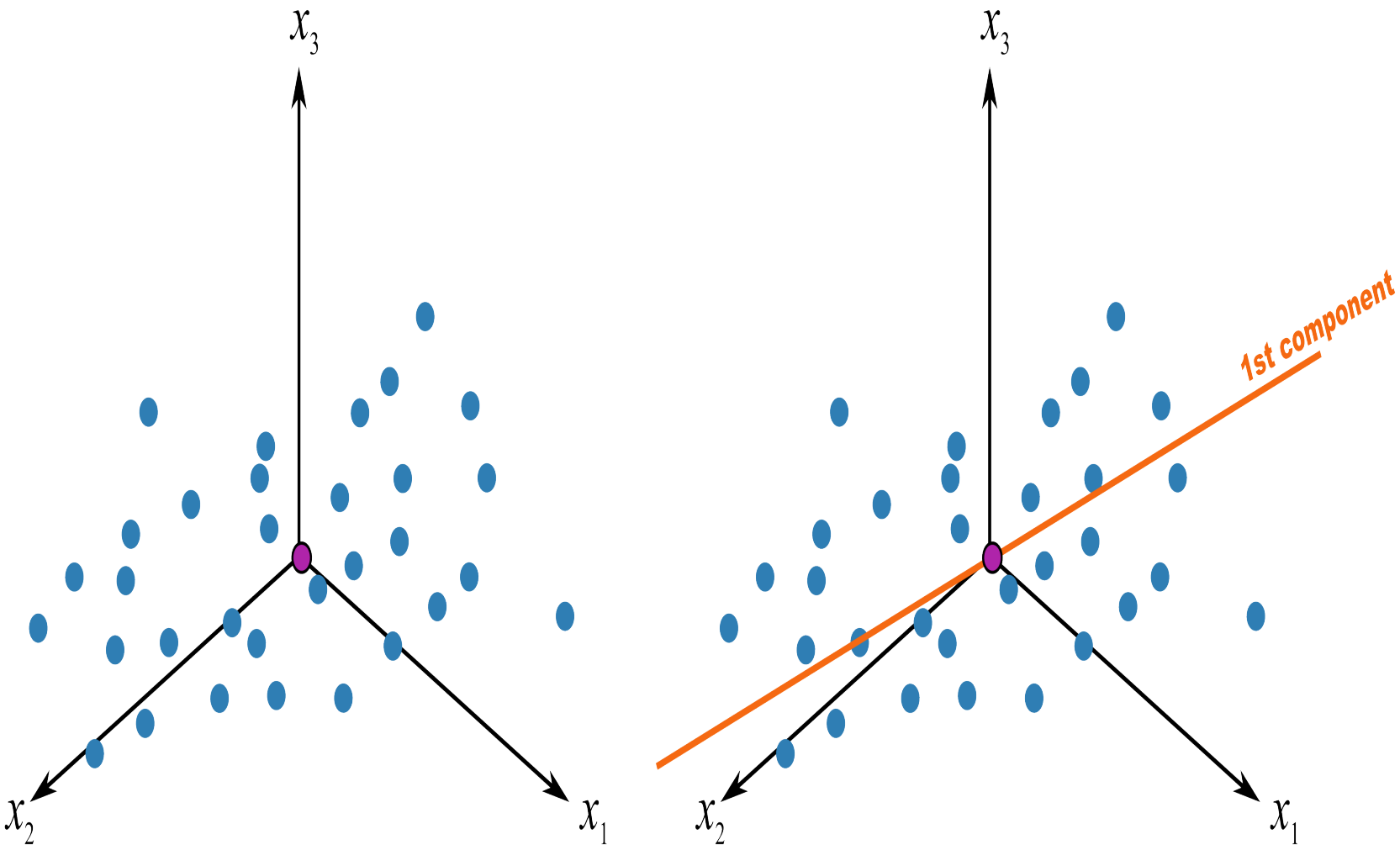
Algunas aplicaciones

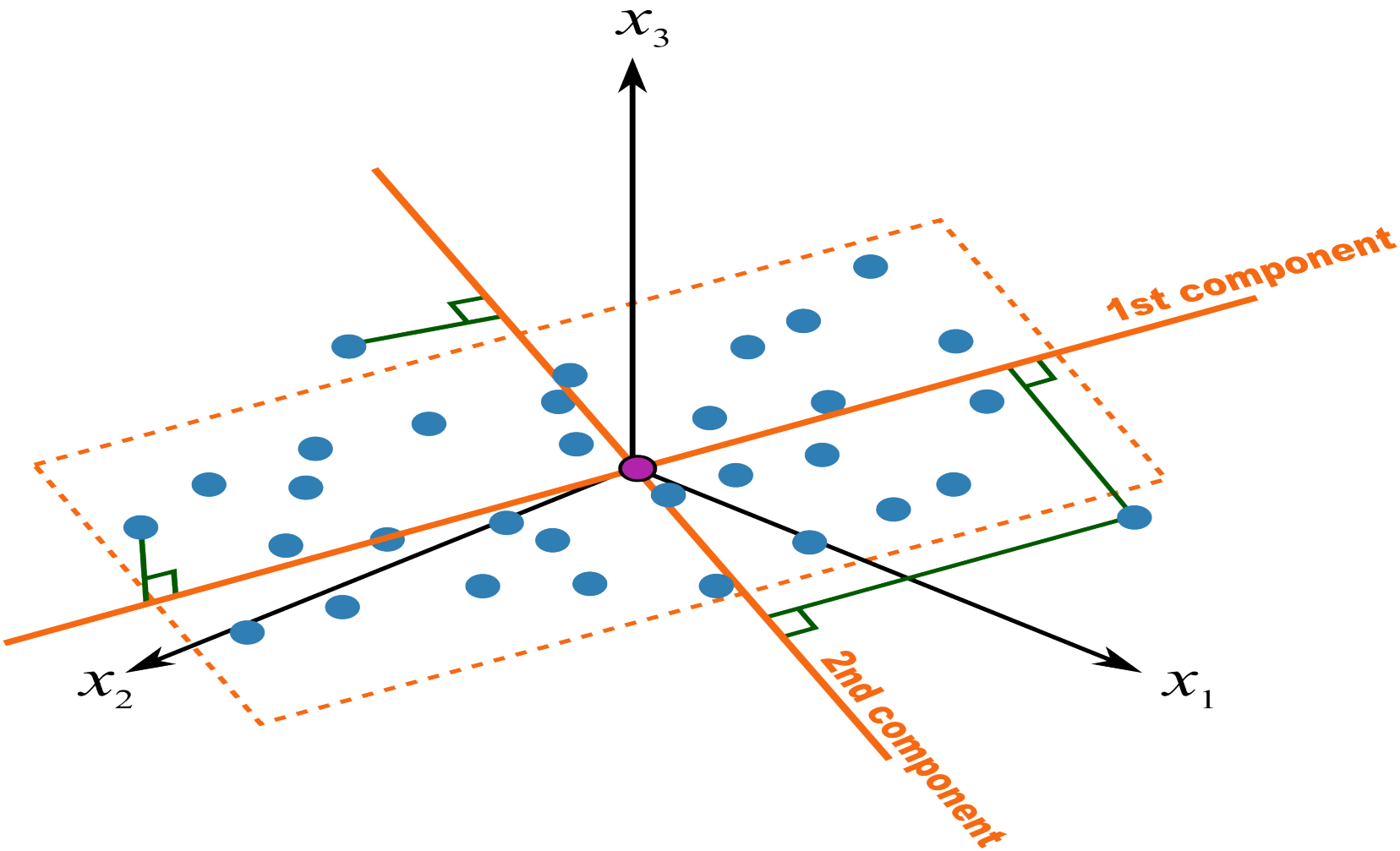
- Mecánica
- Genómica
- Inteligencia Artificial
- Economia
- Estudios sociales
- Análisis de imagen

Interpretación geométrica



- Nuevo sistema de coordenadas.
- Obtención de direcciones ortogonales (o componentes principales) con máxima variabilidad.
- Las nuevas direcciones proporcionan una dispersión simple y parsimoniosa de la estructura de covarianza de los datos.
- Las *componentes principales* dependen de la matriz de varianzas-covarianzas o de la matriz de correlación.


x_3
x_2
x_1
x_3
1st component
x_2
x_1


x_3
1st component
x_2
2nd component
x_1

Interpretación algebraica

- Las componentes principales son combinaciones lineales particulares de $p-$ variables aleatorias $X1, X2, ...Xp$ originales.
- Valores propios
- Vectores propios

Ejemplo e implementación en R

Descripción del conjunto de datos (Gorriones.csv): " . . . el 1 de febrero del presente año (1898), cuando, después de una tormenta extraordinariamente severa de nieve, lluvia y aguanieve, varios gorriones ingleses fueron llevados al Laboratorio Anatómico de la Universidad de Brown. Setenta y dos de estas aves revivieron; sesenta y cuatro perecieron; . . . "." . . . la tormenta fue de larga duración, y las aves fueron recogidas, no en una localidad, sino en varias localidades; . . . ". **Este evento fue descrito por Hermon Bumpus (1898) como un ejemplo clásico de la selección natural en acción.**

Descripción de variables

- Sexo: Machos (m) y hembras (f)
- Edad: Adulto (a) y joven (y)
- Sobrevivió: Sí (SI) y no (NO)
- Longitud total (mm): `LongitudTotal` (desdel la punta del pico hasta la punta de la cola)
- Extensión de las alas (mm): `ExteAlas` (de punta a punta de las alas extendidas)
- Peso (gr): peso del ave
- Longitud del pico y la cabeza (mm): `LonPicoCabe`
- Longitud del húmero (pulgadas): `LonHumero`
- Longitud del fémur (pulgadas): `LonFemur`
- Longitud de tibia-tarso (pulgadas): `LonTibTarso`
- Ancho del cráneo (pulgadas): `AncCraneo`
- Longitud de la quilla (pulgadas): `LonQuilla`

Lectura de datos

```
datos <- read.csv(file = "Gorriones.csv", dec = ",")
head(datos, n = 3)
```

```
##   Sexo Edad Sobrevivio LongitudTotal ExteAlas
## 1    m    a          SI          154      241
## 2    m    a          NO          165      240
## 3    m    a          NO          160      245
##   Peso LonPicoCabe LonHumero LonFemur LonTibTarso
## 1 24.5        31.2     0.687    0.668       1.022
## 2 26.5        31.0     0.738    0.704       1.095
## 3 26.1        32.0     0.736    0.709       1.109
##   AncCraneo LonQuilla
## 1     0.587     0.830
## 2     0.606     0.847
## 3     0.611     0.842
```

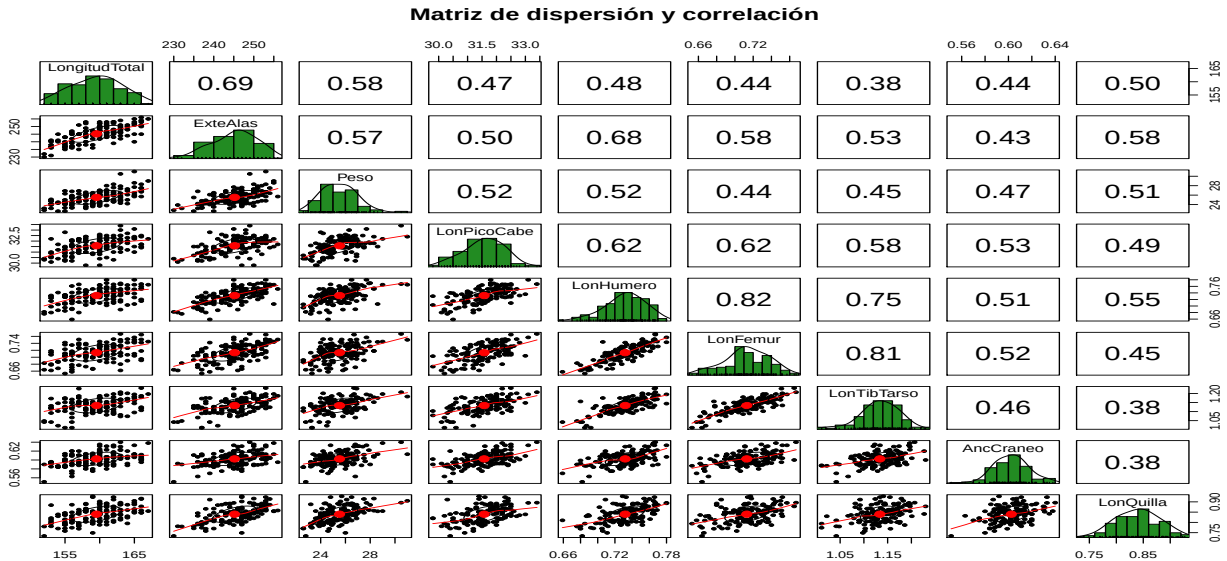
Resumen de datos

```
summary(datos[, 4:12])
```

```
##  LongitudTotal      ExteAlas          Peso
##  Min.   :152.0   Min.   :230.0   Min.   :22.60
##  1st Qu.:157.0   1st Qu.:242.0   1st Qu.:24.57
##  Median :160.0   Median :246.0   Median :25.55
##  Mean   :159.5   Mean   :245.2   Mean   :25.52
##  3rd Qu.:162.0   3rd Qu.:249.0   3rd Qu.:26.50
##  Max.   :167.0   Max.   :256.0   Max.   :31.00
##   LonPicoCabe      LonHumero
##  Min.   :29.80   Min.   :0.6590
##  1st Qu.:31.10   1st Qu.:0.7177
##  Median :31.60   Median :0.7330
##  Mean   :31.57   Mean   :0.7319
##  3rd Qu.:32.02   3rd Qu.:0.7482
##  Max.   :33.40   Max.   :0.7800
```

Matriz de correlación (gráfico)

```
color <- c("darkblue", "darkred")
library(psych)
pairs.panels(datos[,4:12], bg = color[datos$Sexo],
             hist.col = "forestgreen", density = TRUE,
             main = "Matriz de dispersión y correlación")
```



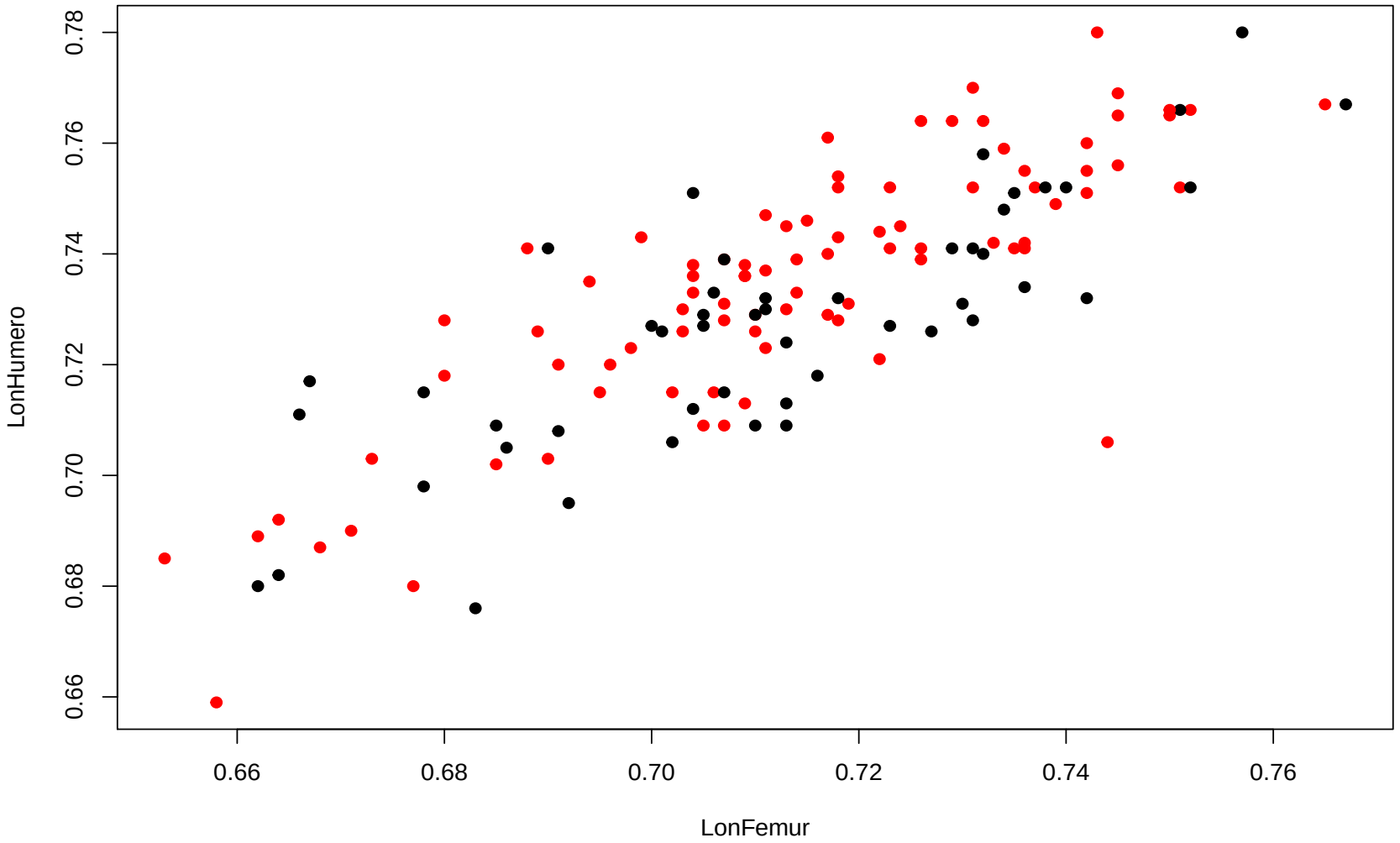
Matriz de correlaciones

```
cor(datos[, 4:12])
```

```
##                 LongitudTotal  ExteAlas      Peso
## LongitudTotal       1.0000000 0.6909709 0.5838648
## ExteAlas            0.6909709 1.0000000 0.5686500
## Peso                0.5838648 0.5686500 1.0000000
## LonPicoCabe         0.4694466 0.4990738 0.5192088
## LonHumero           0.4846190 0.6779536 0.5188943
## LonFemur            0.4447051 0.5782836 0.4441451
## LonTibTarso         0.3776146 0.5316798 0.4544589
## AncCraneo           0.4355363 0.4338913 0.4714846
## LonQuilla           0.5008898 0.5801525 0.5126353
##                 LonPicoCabe LonHumero  LonFemur
## LongitudTotal     0.4694466 0.4846190 0.4447051
## ExteAlas          0.4990738 0.6779536 0.5782836
## Peso              0.5192088 0.5188943 0.4441451
```

Longitud Húmero vs Longitud Fémur

```
with(datos, plot(LonFemur, LonHumero, pch = 19, col = Sexo))
```



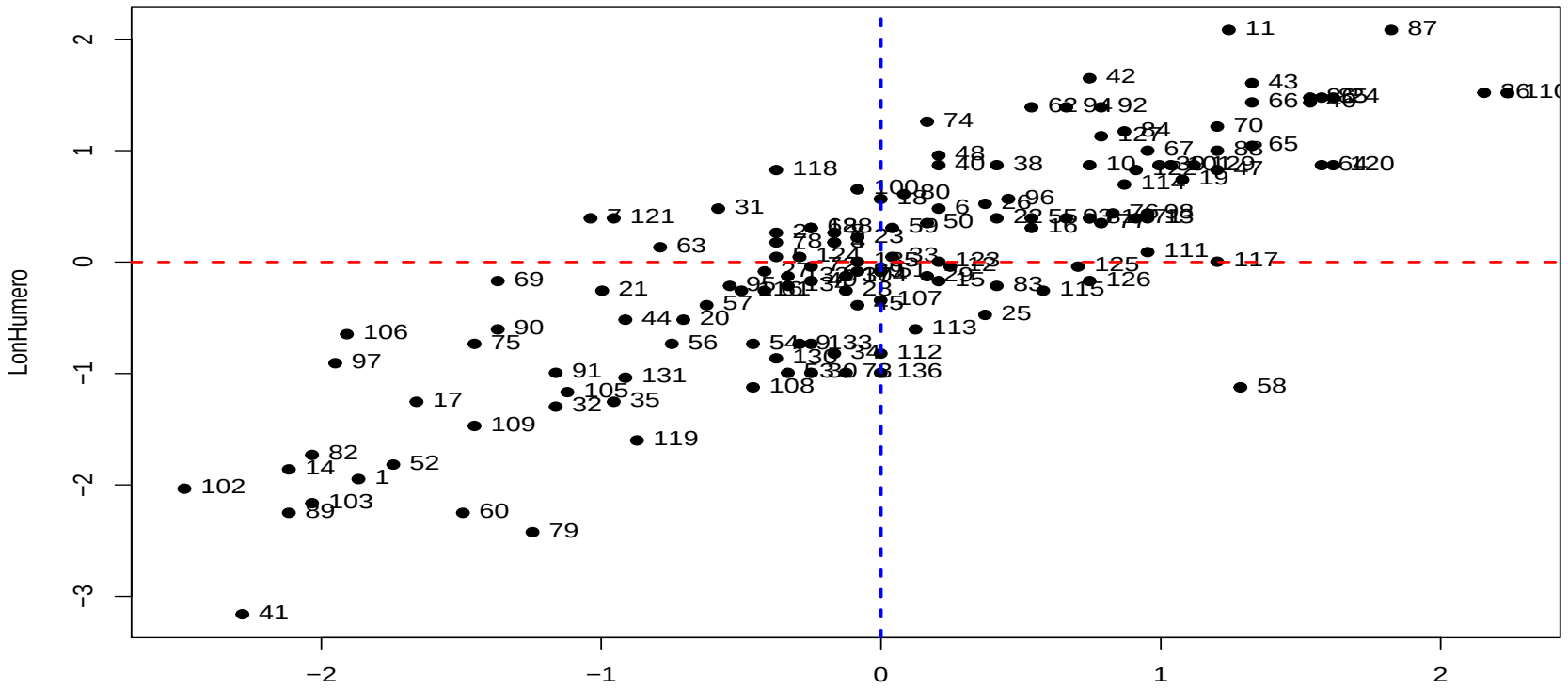
Estandarización de variables

```
datosstd <- data.frame(scale(datos[, 4:12]))
head(datosstd, n = 3)
```

```
##   LongitudTotal   ExteAlas       Peso LonPicoCabe
## 1    -1.5569728 -0.7591304 -0.6948140  -0.5328548
## 2     1.5321934 -0.9402562  0.6609206  -0.8176024
## 3     0.1280269 -0.0346270  0.3897737   0.6061354
##    LonHumero   LonFemur LonTibTarso  AncCraneo
## 1 -1.9473421 -1.8674102  -2.7381900 -1.0321527
## 2  0.2625343 -0.3744578  -0.9465370  0.2348699
## 3  0.1758725 -0.1671033  -0.6029324  0.5682969
##     LonQuilla
## 1 -0.25054259
## 2  0.17821720
## 3  0.05211138
```

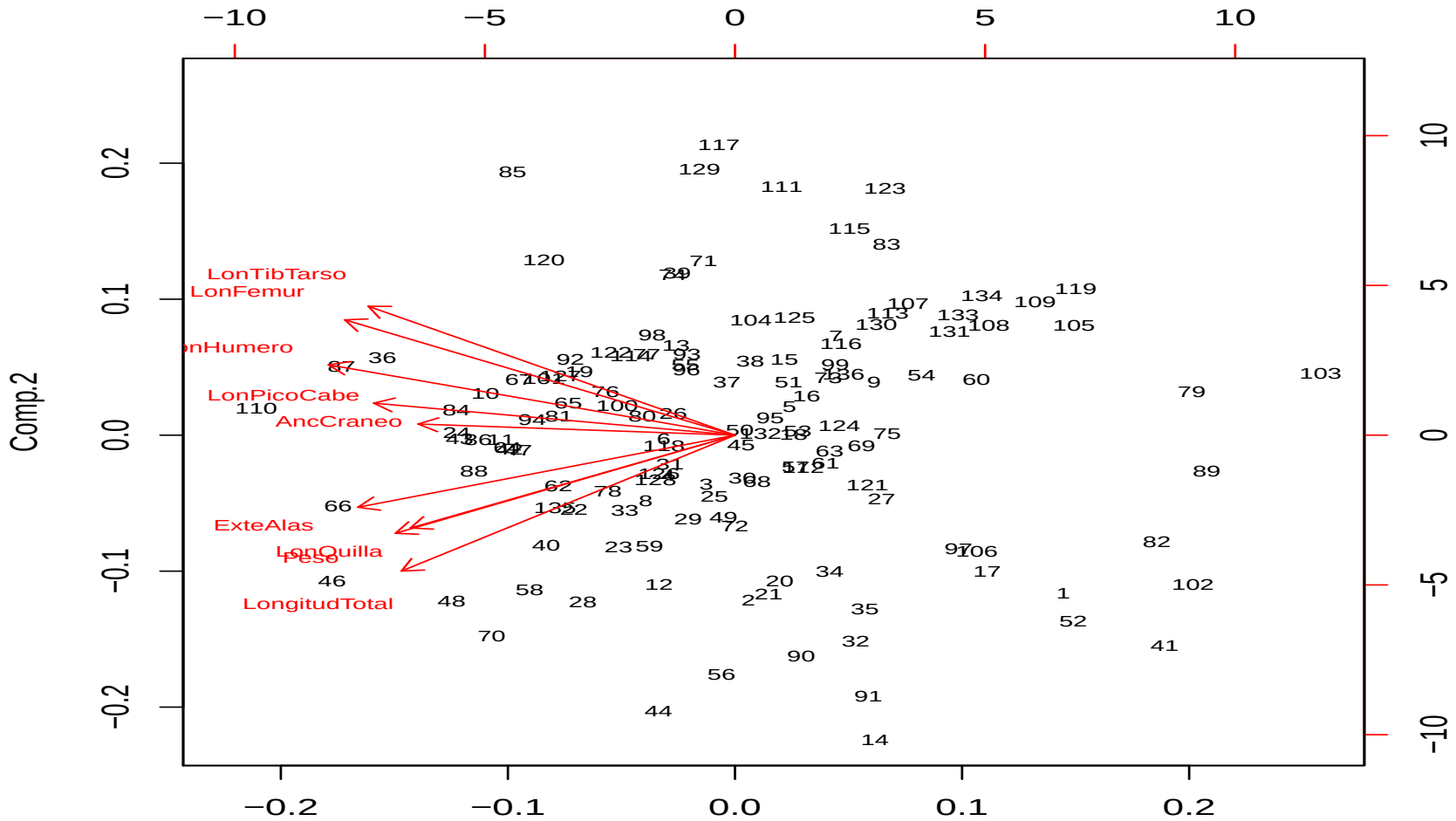
Longitud Húmero vs Longitud Fémur (estandarizadas)

```
with(datosstd, plot(LonFemur, LonHumero, pch = 19))
with(datosstd, text(LonFemur, LonHumero,
                    rownames(datosstd), pos = 4))
abline(h = 0, col = "red", lty = 2, lwd = 2)
abline(v = 0, col = "blue", lty = 2, lwd = 2)
```



Cálculo de las *componentes principales*

```
acp <- princomp(datosstd, cor = TRUE)
biplot(acp, cex = 0.7)
```

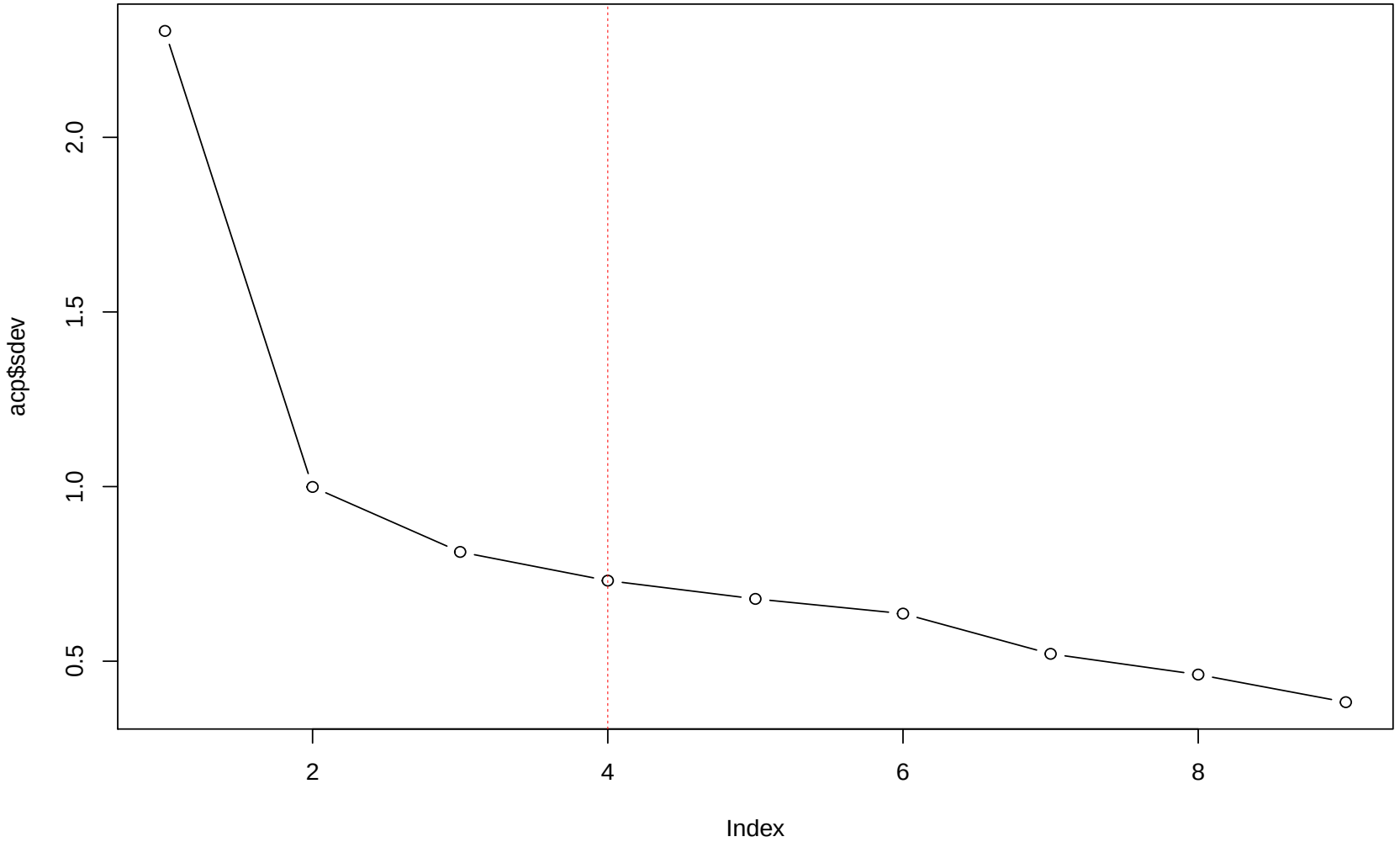


Valores propios

```
summary(acp)
```

```
## Importance of components:
##                           Comp.1    Comp.2
## Standard deviation     2.3046882 0.9988978
## Proportion of Variance 0.5901764 0.1108663
## Cumulative Proportion  0.5901764 0.7010427
##                           Comp.3     Comp.4
## Standard deviation     0.81280426 0.73068317
## Proportion of Variance 0.07340564 0.05932199
## Cumulative Proportion  0.77444838 0.83377037
##                           Comp.5     Comp.6
## Standard deviation     0.67837980 0.63624854
## Proportion of Variance 0.05113324 0.04497913
## Cumulative Proportion  0.88490361 0.92988274
##                           Comp.7     Comp.8
```

```r
plot(acp$sdev, type="b")
abline(v = 4, col = "red", lty = 2, lwd = 0.5)
```



Vectores propios

```
loadings(acp)[,1:9]
```

```
##                    Comp.1      Comp.2      Comp.3
## LongitudTotal -0.3100651 -0.48676787 -0.05636978
## ExteAlas      -0.3505712 -0.25833261 -0.33786910
## Peso          -0.3155630 -0.35142457  0.19791968
## LonPicoCabe   -0.3358877  0.11469994  0.29286266
## LonHumero     -0.3779134  0.25196559 -0.22209147
## LonFemur      -0.3628170  0.41308546 -0.14264156
## LonTibTarso   -0.3408942  0.46175891 -0.15231419
## AncCraneo     -0.2946541  0.03996533  0.78349062
## LonQuilla     -0.3017853 -0.33274799 -0.22582687
##                    Comp.4      Comp.5      Comp.6
## LongitudTotal  0.45377621  0.1326996  0.36031440
## ExteAlas       0.25610634  0.2419548  0.02384397
## Peso           0.07935593 -0.7038798 -0.46084211
```

Puntajes

```
head(acp$scores, n = 3)
```

```
##           Comp.1     Comp.2    Comp.3     Comp.4
## [1,]  3.8880757 -1.3522616 0.4160354 -1.1326393
## [2,]  0.1575123 -1.4096787 0.4072278  0.4664938
## [3,] -0.3388752 -0.4201252 0.7720830 -0.2187090
##            Comp.5     Comp.6     Comp.7    Comp.8
## [1,] -0.09606579  0.3027823 -0.9560481 0.1354796
## [2,]  0.09095160 -0.3538219  1.1263744 1.2205056
## [3,] -0.07081691  0.2364849 -0.3398357 0.4256793
##           Comp.9
## [1,]  0.61889581
## [2,] -0.53167046
## [3,] -0.03965116
```

Concatenando resultados (por CP1)

```
datos2 <- data.frame(datos, acp$scores[, c(1, 2, 3)])
a <- datos2[order(acp$scores[, 1], decreasing = TRUE), ]
head(a, n = 3)
```

```
##     Sexo Edad Sobrevivio LongitudTotal ExteAlas
## 103    f              NO           152      230
## 89     f              NO           153      231
## 102    m    y         SI           155      237
##     Peso LonPicoCabe LonHumero LonFemur
## 103 22.8        30.4     0.682    0.664
## 89  23.9        30.1     0.680    0.662
## 102 23.3        30.2     0.685    0.653
##     LonTibTarso AncCraneo LonQuilla   Comp.1
## 103       1.042     0.551     0.734 6.930919
## 89        1.042     0.592     0.781 5.583038
## 102       1.011     0.587     0.794 5.420028
```

CP1 vs CP2 (código_1...)

```
color <- c("magenta4",  "#66A61E")
simbolos <- c(15, 17)
with(datos2, plot(Comp.1, Comp.2, col = color[Sexo],
                  pch = simbolos[Sobrevivio],
                  cex = 1.5,  xlab = "CP1",
                  ylab = "CP2",
    main = "Grupo de aves sobre las componentes principales"))
```

CP1 vs CP2 (código_2...)

```
legend("topleft",
       legend = c("Hembra", "Macho"),
       col = color,
       cex = 1,
       lty =1,
       lwd =2)
legend("topright",
       legend = c("Murió", "Sobrevió"),
       pch = simbolos,
       col = "black",
       cex = 1)
abline(h = 0, col = "red")
abline(v = 0, col = "red")
```

CP1 vs CP2 (código_3)

```
arrows(0, 0,
       acp$loadings[, 1]*5,
       acp$loadings[, 2]*5,
       col = "red",
       lwd = 2.5)
text(acp$loadings[, 1]*5.2,
     acp$loadings[, 2]*5.2,
     row.names(acp$loadings),
     cex = 1.3)
```

Aves proyectadas sobre las componentes principales 1 y 2