PCA

María Inés Fariello (basado en diapos de Matías Carrasco)

Instituto de Matemática y Estadística – Facultad de Ingeniería – UdelaR

27 de Abril, 2022”

Notación y terminología básica del PCA

La matriz de datos

El análisis de componentes principales (PCA en inglés) se aplica a *tablas de datos* donde las filas se consideran como individuos y las columnas como variables cuantitativas.

Denotamos x_{ik} el valor tomado por el individuo i para la variable k, donde i varía de 1 a I y k de 1 a K. De este modo la matriz de datos es:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & x_{ik} & \vdots \\ x_{I1} & \cdots & x_{IK} \end{pmatrix}$$

También vamos usar la media y el desvío estándar de cada variable:

$$\bar{x}_k = \frac{1}{I}\sum_{i=1}^{I} x_{ik}, \quad s_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \bar{x}_k)^2}$$

Ejemplo: caparazón de tortugas

Descripción del conjunto de datos

El conjunto de datos contiene mediciones del caparazón de 24 tortugas pintadas machos y 24 hembras (Chrysemys picta marginata).

El data frame `turtles` del paquete `Flury` contiene 48 observaciones de las siguientes 4 variables:

- Gender: a factor with levels Male/Female
- Length: carapace length
- Width: carapace width
- Height: carapace height

Fuente: Jolicoeur, P. and J.E. Mosimann (1960) "Size and Shape Variation in the Painted Turtle: A Principal Component Analysis' ', Growth, 24:339-354

Ejemplo: caparazón de tortugas

Fuente: Wikipedia

Dos puntos de vista

En este ejemplo la matriz de datos X consta de $K = 3$ columnas o variables e $I = 48$ individuos u observaciones. También disponemos de una cuarta variable categórica que indica el sexo de la tortuga.

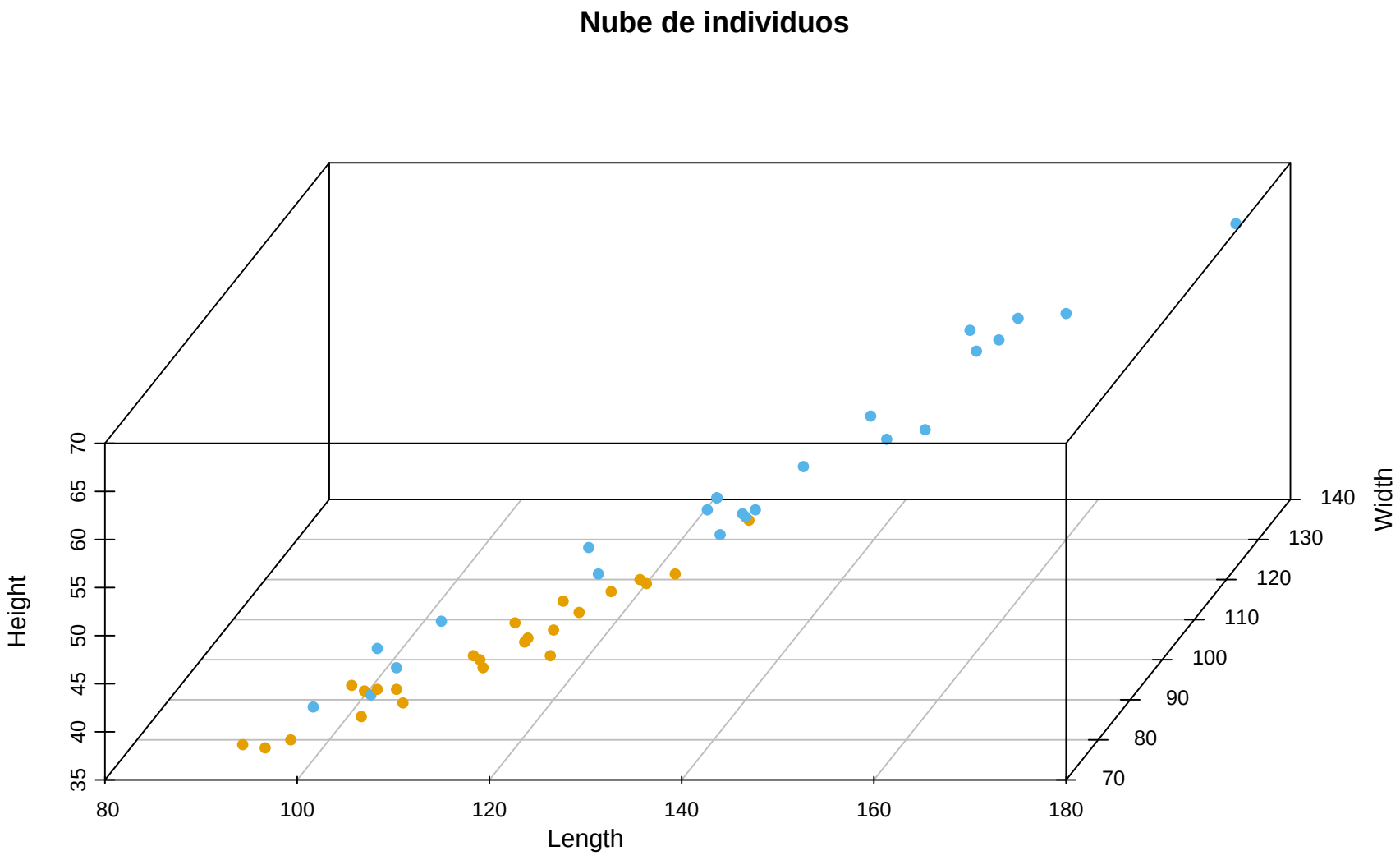
La nube de individuos

Cada fila de la matriz de datos es un vector de $\mathbb{R}^K$ que representa las mediciones realizadas sobre un mismo individuo. El conjunto de I puntos en el espacio $\mathbb{R}^K$ se conoce como *la nube de individuos*. El espacio $\mathbb{R}^K$ se llama el *espacio de individuos*.

La nube de variables

Cada columna de la matriz de datos es un vector de $\mathbb{R}^I$ que representa las mediciones de una sola variable sobre el conjunto de individuos. El conjunto de K puntos en el espacio $\mathbb{R}^I$ se conoce como *la nube de variables*. El espacio $\mathbb{R}^I$ se llama el *espacio de variables*.

El conjunto de datos en 3 dimensiones



La nube de individuos

La distancia euclídea entre individuos

La distancia euclídea entre los individuos i y l está dada por

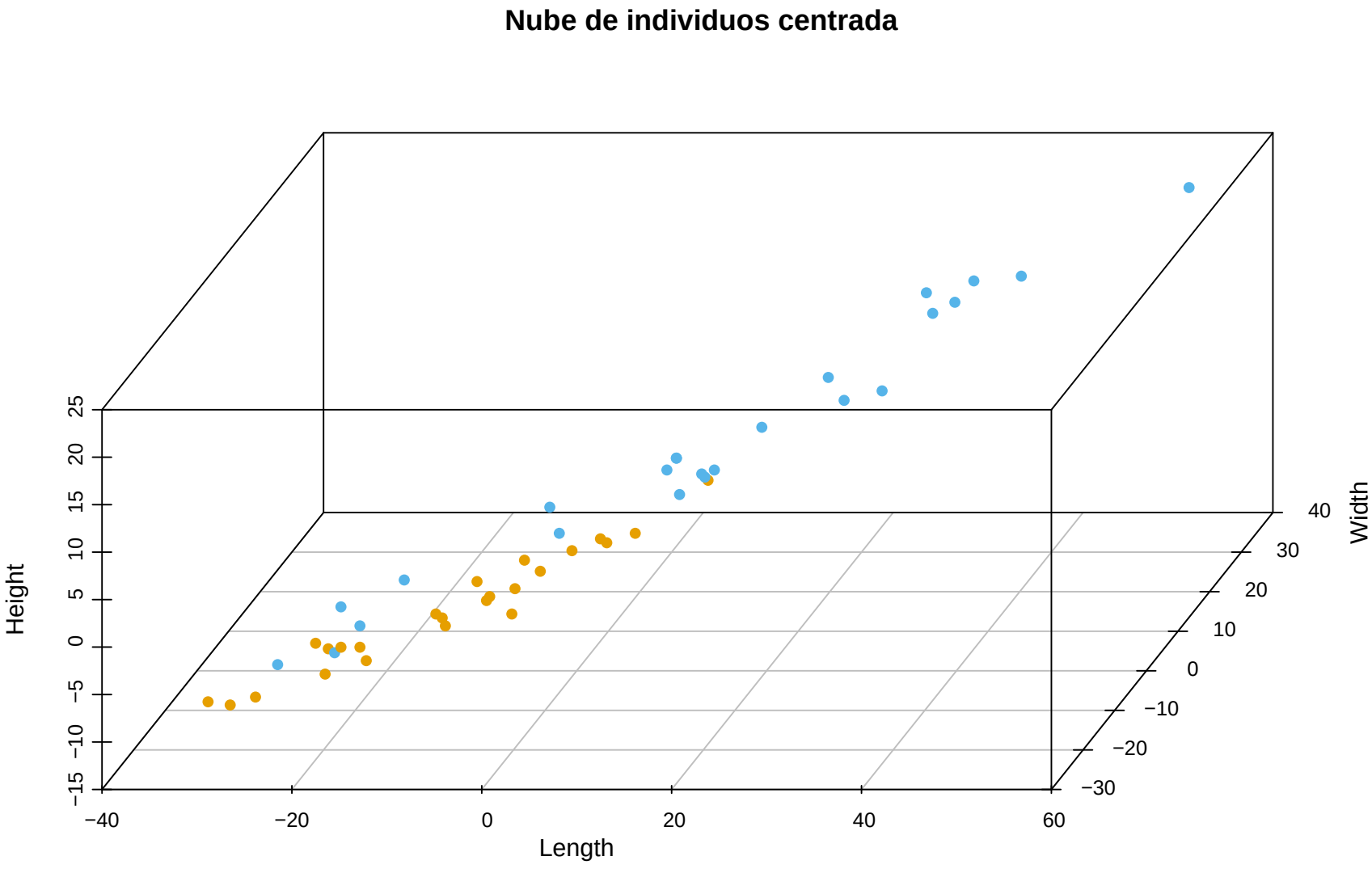
$$d(i,l) = \sqrt{\sum_{k=1}^{K} (x_{ik} - x_{lk})^2}$$

Si dos individuos tienen valores similares en las K variables de la tabla, también están cerca en el espacio $\mathbb{R}^K$.

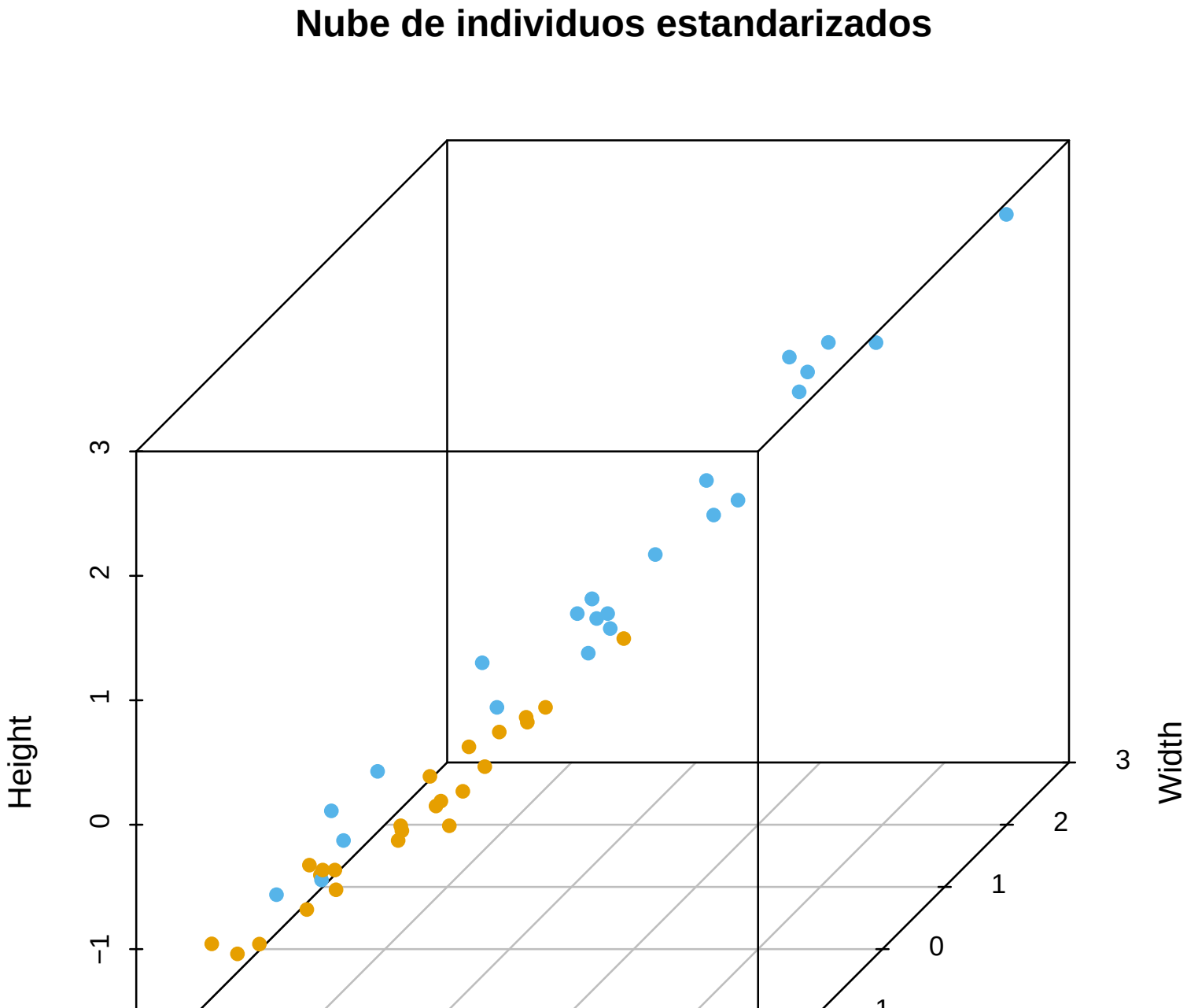
Centrar y estandarizar

La forma de la nube sigue siendo la misma incluso cuando se la traslada. En PCA siempre es conveniente trabajar con los datos centrados, lo que corresponde a considerar $x_{ik} - \bar{x}_k$ en lugar de x_{ik}. El PCA estandarizado es cuando trabajamos con $(x_{ik} - \bar{x}_k)/s_k$ que sí modifica la forma de la nube. Es la opción por defecto en la mayoría de los casos.

La nube de individuos centrada no pierde su forma



Individuos estandarizados: todas las variables "pesan" lo mismo



Ajustando la nube de individuos

El objetivo

El objetivo del PCA es representar la nube de puntos en un espacio de dimensiones reducidas de forma "óptima", es decir, distorsionando lo menos posible las distancias euclideas entre individuos.

Comencemos por la mejor representación en 1D

Para obtener esta representación en 1D, la nube se proyecta sobre una recta de $\mathbb{R}^K$ denotada r, elegida de tal manera que se minimice la distorsión de la nube de puntos, es decir, tal que las distancias entre los puntos proyectados sean lo más cercanas posible a las distancias entre los puntos iniciales.

Distorsión de distancias

Queremos minimizar

$$\min_r \left\{ \sum_{\{i,l\}} \left| d(i,l)^2 - d_r(i,l)^2 \right| \right\}$$

en donde $d_r(i,l)$ es la distancia entre los puntos proyectados.

Dado que, en la proyección ortogonal, las distancias solo pueden disminuir:

$$\min_r \left\{ \sum_{\{i,l\}} \left| d(i,l)^2 - d_r(i,l)^2 \right| \right\} = \min_r \left\{ \sum_{\{i,l\}} d(i,l)^2 - d_r(i,l)^2 \right\}$$

$$= \sum_{\{i,l\}} d(i,l)^2 - \max_r \left\{ \sum_{\{i,l\}} d_r(i,l)^2 \right\}$$

Distorsión de distancias

El problema es equivalente a maximizar

$$\max_r \left\{ \sum_{\{i,l\}} d_r(i,l)^2 \right\}$$

Llamemos r_i a la proyección ortogonal sobre r del individuo i.
Entonces

$$\begin{aligned}
\sum_{\{i,l\}} d_r(i,l)^2 &= \sum_{\{i,l\}} \|r_i - r_l\|^2 = \frac{1}{2} \sum_{i,l} \|r_i - r_l\|^2 \\
&= \frac{1}{2} \sum_{i,l} \|r_i\|^2 + \frac{1}{2} \sum_{i,l} \|r_l\|^2 - \sum_{i,l} r_i \cdot r_l \\
&= I \sum_i \|r_i\|^2 - \sum_{i,l} r_i \cdot r_l = I \sum_i \|r_i\|^2 - \left(\sum_i r_i \right) \cdot \left(\sum_l r_l \right) \\
&= I^2 \left(\frac{1}{I} \sum_i \|r_i\|^2 - \left\| \frac{1}{I} \sum_i r_i \right\|^2 \right) = I^2 \mathrm{Var}\left(\{r_i\} \right)
\end{aligned}$$

El punto de vista estadístico

La recta de mayor variabilidad

En conclusión, la recta que distorsiona lo menos posible las distancias entre los puntos proyectados coincide con la recta que maximiza la varianza de los puntos proyectados.

La mejor representación en nD

Para obtener la representación en nD (lo mas usual es $n = 2$), la nube se proyecta sobre un hiperplano n-dimensional de $\mathbb{R}^K$ denotado H, elegido de tal manera que se minimice la distorsión de la nube de puntos. Al igual que en 1D, esto equivale a buscar el hiperplano que maximiza la variabilidad:

$$\min_H \left\{ \sum_{\{i,l\}} \left| d(i,l)^2 - d_H(i,l)^2 \right| \right\} = I^2 \max_H \{ \mathrm{Var}\left(\{ p_H(i) \} \right) \}$$

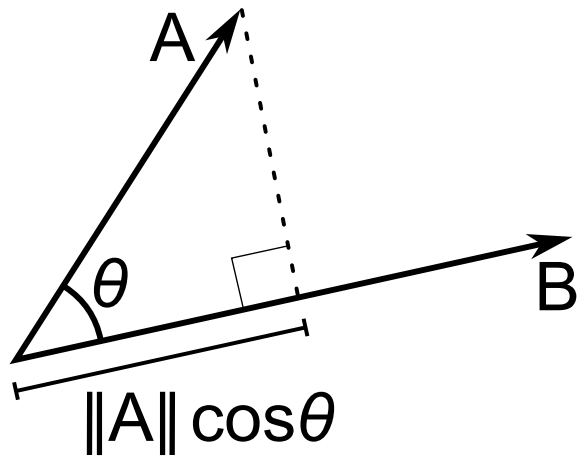
en donde p_H indica la proyección ortogonal sobre H.

Repaso: proyección ortogonal

Recordar que al proyectar ortogonalmente un vector A sobre un vector B, la longitud de la proyección está dada por $\|A\| \cos(\theta)$ en donde θ es el ángulo entre A y B.
Si el vector B tiene norma 1, esta longitud es igual al producto escalar entre A y B:

$$A \cdot B = \langle A, B \rangle = A^t B$$

si pensamos a los vectores como matrices de una sola columna.



¿Cómo encontrar la recta óptima?

Proyección de individuos sobre una dirección

Un individuo viene representado en el espacio de individuos $\mathbb{R}^K$ por una fila de la matriz de datos X. Sea u un vector unitario (una dirección) en $\mathbb{R}^K$. La proyección ortogonal del individuo i sobre u está dada por el producto escalar (fila i de X) $\cdot\, u$. Entonces, la proyección de la nube de individuos entera viene dada por el vector Xu.

La varianza de la proyección

Cuando X está centrada, lo mismo ocurre con la proyección Xu. De este modo la varianza de la proyección viene dada por $\frac{1}{I}\|Xu\|^2$. Esto quiere decir que la recta de mayor variabilidad será aquella que maximiza

$$\max_{u:\|u\|=1} \|Xu\|^2$$

Aparece la matriz de covarianzas

Vamos a asumir de ahora en mas que X está centrada.

La varianza de la proyección

La varianza de la proyección está dada por

$$\|Xu\|^2 = (Xu)^t(Xu) = (u^tX^t)(Xu) = u^t(X^tX)u$$

La matriz de covarianzas

La entrada k, l de la matriz X^tX es (recordar que X está centrada):

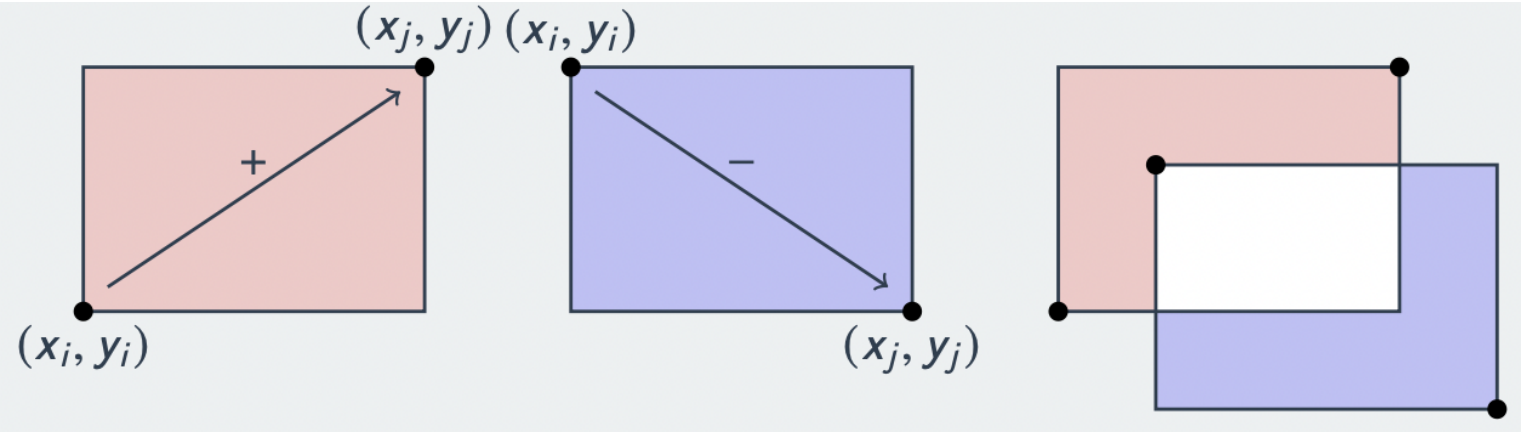
$$(X^tX)_{kl} = \sum_{i=1}^{l}(x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l) = l\mathrm{Cov}(k, l)$$

que es (a menos del factor l) la covarianza entre la variable k y la variable l.

Interpretación de la covarianza

Sean $\{x_1, \ldots, x_n\}$ e $\{y_1, \ldots, y_n\}$ muestras de tamaño n de dos variables cuantitativas. Para cada par de pares (x_i, y_i) y (x_j, y_j) podemos construir un rectángulo que los tenga como vértices.

Dos situaciones se pueden dar:



Fuente: Wikipedia

En el primer caso lo pintamos de rojo para indicar la asociación positiva, y en el segundo de azul para indicar la asociación negativa. Cuando dos rectángulos opuestos y de la misma intensidad se solapan, la intersección se representa en blanco como color neutro.

¿Qué representa cada rectángulo?}

La cantidad de color de un rectángulo es el área del mismo: $(x_i - x_j) \times (y_i - y_j)$ (rojo cuando es positivo, azul cuando es negativo). Cuando sumamos en todos los rectángulos, obtenemos la cantidad total de rojo menos la de azul, teniendo en cuenta que cuando un rectángulo rojo se solapa con uno azul, la intersección queda con el color correspondiente a la resta de las intensidades.

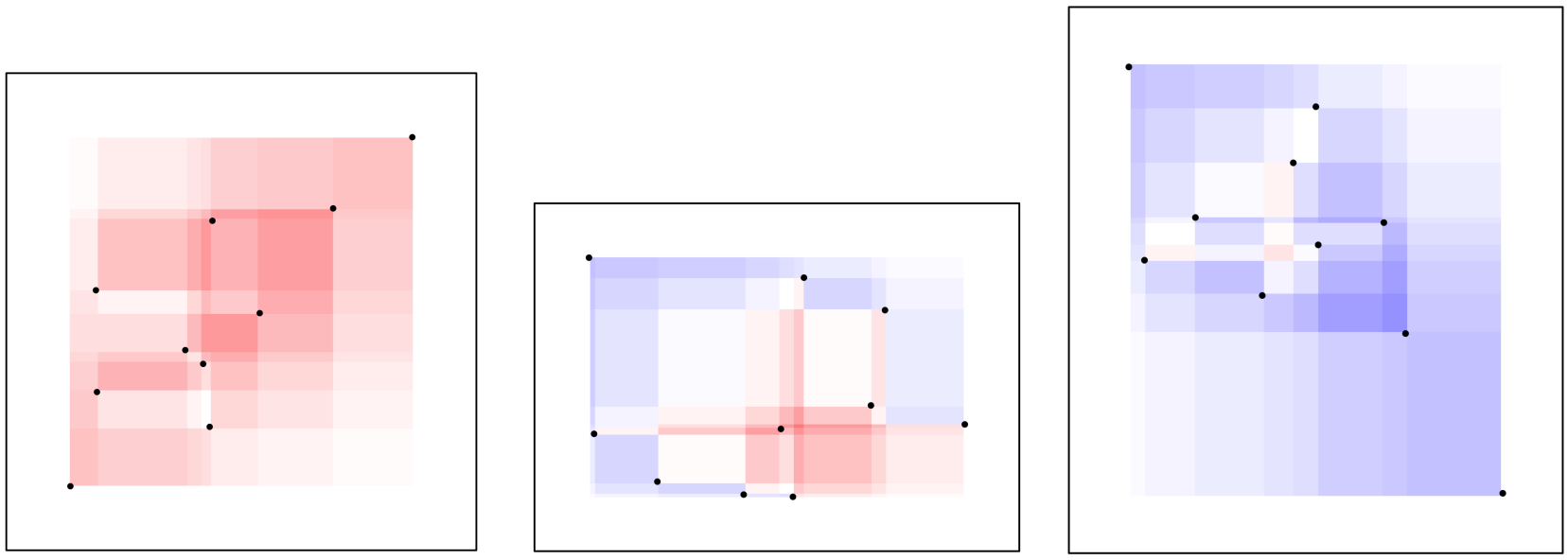
$$\sum_{\text{rectángulos}} (x_i - x_j)(y_i - y_j) = \frac{1}{2}\sum_{i,j}(x_i - x_j)(y_i - y_j) =$$

$$\frac{1}{2}\sum_{i,j}(x_i y_i - x_i y_j - x_j y_i + x_j y_j)$$

$$= \sum_{i,j} x_i y_i - \sum_{i,j} x_i y_j = n\sum_{i=1}^{n} x_i y_i - n^2 \bar{x}\,\bar{y}$$

$$= n^2\left(\overline{xy} - \bar{x}\,\bar{y}\right) = n^2 \mathrm{Cov}(x, y)$$

Interpretación visual de la covarianza



Izquierda a derecha: covarianza positiva; covarianza nula (o casi); covarianza negativa.

Interpretación geométrica de la covarianza

La dirección de la media

Llamamos *dirección de la media* en $\mathbb{R}^n$ al vector unitario

$$u = \frac{1}{\sqrt{n}}(1, \ldots, 1)$$

de coordenadas iguales.

El promedio como proyección

Consideremos una muestra $\{x_1, \ldots, x_n\}$ como un vector de $\mathbb{R}^n$: $x = (x_1, \ldots, x_n)$. Llamemos $\{e_1, \ldots, e_n\}$ a la base canónica de $\mathbb{R}^n$. Así $u = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i$, y la proyección de x sobre u es

$$(x \cdot u)u = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \right) u = (\bar{x}, \ldots, \bar{x}) \,.$$

Interpretación geométrica de la covarianza

La varianza como norma al cuadrado de una proyección

Además, el espacio V ortogonal a u es

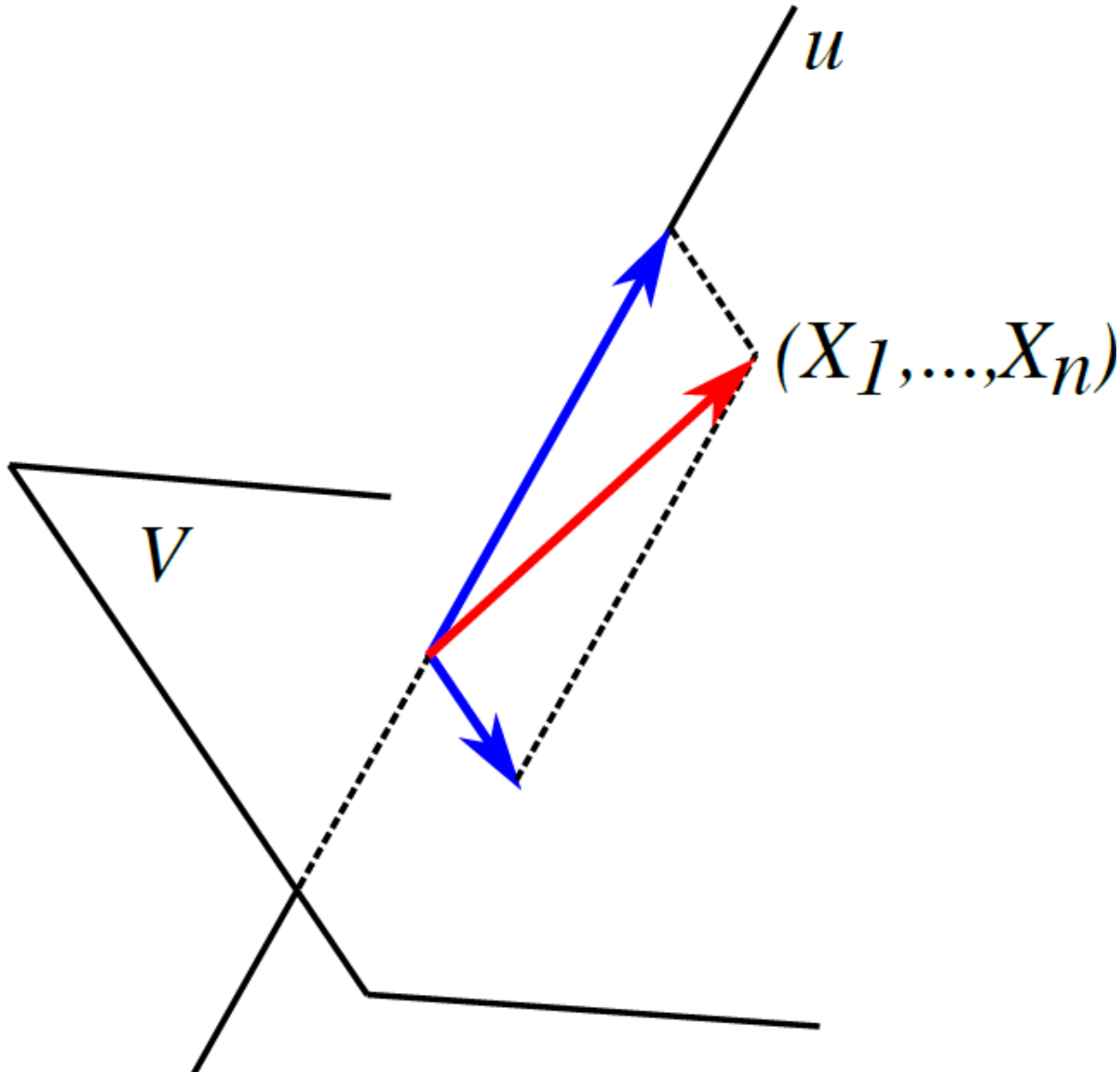
$$V = \{x \in \mathbb{R}^n : x \cdot u = 0\} = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 0 \right\},$$

y la proyección de x sobre V es

$$x_V = x - (\sqrt{n}\,\bar{x})u = \sum_{i=1}^{n} (x_i - \bar{x}) e_i$$

cuya norma al cuadrado es $\|x_V\|^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 = n\mathrm{Var}(x)$.

La varianza como norma al cuadrado de una proyección



Interpretación geométrica de la covarianza

Sean ahora $\{x_1, \ldots, x_n\}$ e $\{y_1, \ldots, y_n\}$ muestras de tamaño n de dos variables cuantitativas, que pensamos como vectores de $\mathbb{R}^n$:

$$x = (x_1, \ldots, x_n), \quad y = (y_1, \ldots, y_n).$$

Las proyecciones de x e y sobre u y sobre V son como antes:

- $x_u = (x \cdot u)u = (\bar{x}, \ldots, \bar{x}), \quad y_u = (y \cdot u)u = (\bar{y}, \ldots, \bar{y})$
- $x_V = x - x_u, \quad y_V = y - y_u, \quad \|x_V\|^2 = n\mathrm{Var}(x), \quad \|y_V\|^2 = n\mathrm{Var}(y)$.

La covarianza como producto escalar

Más aún, el producto escalar entre x_V e y_V es

$$x_V \cdot y_V = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = n\mathrm{Cov}(x, y).$$

El coeficiente de correlación de Pearson

Producto escalar y ángulo entre dos vectores

Recordar que el coseno del ángulo θ entre dos vectores A y B viene dado por

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

El coeficiente de correlación

Definimos el coeficiente de correlación entre x e y como el coseno ángulo θ de sus proyecciones x_V e y_V:

$$r = \mathrm{Cor}(x, y) = \cos(\theta) = \frac{x_V \cdot y_V}{\|x_V\|\|y_V\|} = \frac{\mathrm{Cov}(x, y)}{\mathrm{sd}(x)\mathrm{sd}(y)}$$

Propiedades del coeficiente de correlación

El coeficiente de correlación tiene las siguientes propiedades:

- $-1 \leq r \leq 1$
- $r = 0$ solamente cuando x_V e y_V son ortogonales ($\theta = \pi/2$)
- $r = 1$ solamente cuando el ángulo entre x_V e y_V es $\theta = 0$, lo cual implica que $y_V = ax_V$ para cierta constante $a > 0$. A su vez, esto implica que $y = ax + b$ en donde b es una constante que compensa la relación entre los promedios.
- $r = -1$ solamente cuando el ángulo entre x_V e y_V es $\theta = \pi$, lo cual implica que $y_V = ax_V$ para cierta constante $a < 0$. A su vez, esto implica que $y = ax + b$ en donde b es una constante que compensa la relación entre los promedios.

Volviendo a la matriz de covarianzas

La matriz de covarianzas y la dirección óptima

Dada X una matriz de datos, definimos la matriz de covarianzas S cuya entrada kl es

$$S_{kl} = \mathrm{Cov}(k, l).$$

En el caso en que trabajemos con la matriz de datos centrada, la matriz S puede calcularse como

$$S = \frac{1}{l} X^t X$$

Recordar que estamos buscando la dirección $u \in \mathbb{R}^K$ que maximiza $\frac{1}{l}\|Xu\|^2$, y por lo que vimos antes esto equivale a maximizar

$$\max_{u:\|u\|=1} u^t S u$$

ya que $\frac{1}{l}\|Xu\|^2 = u^t S u$.

Covarianza y correlación para los datos de las tortugas

```
(S=cov(turtles[,2:4]))
```

```
##             Length    Width    Height
## Length 419.4960 253.9907 165.82979
## Width  253.9907 160.6769 102.19149
## Height 165.8298 102.1915  70.43972
```

```
(R=cor(turtles[,2:4]))
```

```
##             Length     Width    Height
## Length 1.0000000 0.9783116 0.9646946
## Width  0.9783116 1.0000000 0.9605705
## Height 0.9646946 0.9605705 1.0000000
```

Observaciones sobre matrices de covarianza y correlación

- La diagonal de la covarianza S aparecen las respectivas varianzas de las tres variables .
- Ambas matrices son simétricas
- La matriz de correlaciones R es la matriz de covarianzas de la matriz de datos estandarizada.

Cálculo de la dirección óptima

El caso de 2 variables

Por simplicidad, supongamos $K = 2$ variables. Así la matriz de covarianzas es una matriz 2×2:

$$S = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

y queremos maximizar $\frac{1}{I}\|Xu\|^2 = u^t S u$:

$$F(u) = u^t S u = (u_1, u_2) \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = s_x^2 u_1^2 + 2 s_{xy} u_1 u_2 + s_y^2 u_2^2$$

con la restricción $N(u) = u_1^2 + u_2^2 = 1$.

Cálculo de la dirección óptima usando multiplicadores de Lagrange

Recordar que el método de los multiplicadores de Lagrange consiste en introducir un multiplicador λ y resolver el sistema

$$\nabla F - \lambda \nabla N = 0$$

Calculamos los gradientes:

$$\nabla F = 2\left(s_x^2 u_1 + s_{xy} u_2, s_{xy} u_1 + s_y^2 u_2\right) = 2 \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 2Su$$

$$\nabla N = 2(u_1, u_2) = 2u$$

de donde obtenemos la ecuación $Su = \lambda u$. Esto quiere decir que la dirección óptima está dada por el vector propio de valor propio maximal.

Los valores y vectores propios de S

Diagonalización

Toda matriz simétrica es diagonalizable en una base ortonormal. Además, si la matriz es definida no negativa, sus valores propios son no negativos. Esto implica que existe una base ortonormal $\{pc_1, \ldots, pc_K\}$ de vectores propios de $\mathbb{R}^K$ (vendría a ser como rotar los ejes coordenados) con valores propios

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_K \geq 0$$

tales que $S\, pc_k = \lambda_k\, pc_k$.

La varianza proyectada en la base diagonal

Si trabajamos en esta base, la varianza proyectada en la dirección u queda

$$u^t S u = \lambda_1 u_1^2 + \cdots + \lambda_K u_K^2$$

La dirección óptima es PC1

Notar por un lado que

$$u^t S u = \lambda_1 u_1^2 + \cdots + \lambda_K u_K^2 \geq \lambda_1 u_1^2$$

y tomando máximo en ambos lados

$$\max_{u:\|u\|=1} u^t S u \geq \max_{u:\|u\|=1} \lambda_1 u_1^2 \geq \lambda_1.$$

Por otro lado

$$u^t S u = \lambda_1 u_1^2 + \cdots + \lambda_K u_K^2 \leq \lambda_1 (u_1^2 + \cdots + u_K^2) = \lambda_1,$$

y tomando máximo en ambos lados

$$\max_{u:\|u\|=1} u^t S u \leq \lambda_1.$$

Es decir,

$$\boxed{\max_{u:\|u\|=1} u^t S u = \lambda_1 \text{ y la dirección óptima es } u = \pm pc_1.}$$

En resumen

Para encontrar la dirección de mejor representación, o lo que es equivalente de mayor variabilidad, debemos:

1. A partir de la matriz de datos calculamos la matriz de covarianzas S. En caso de que X esté centrada, el cálculo es simplemente $S = \frac{1}{I} X^t X$.
2. Hallamos la base de vectores propios $\{pc_1, \ldots, pc_K\}$ de S y los valores propios

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_K \geq 0$$

3. La dirección óptima es $u = \pm pc_1$.
4. Los valores proyectados en esta dirección se obtienen haciendo $X\, pc_1$.

Ejemplo Tortugas

Calculamos los vectores y valores propios

```
(PC <- eigen(S,symmetric = T))
```

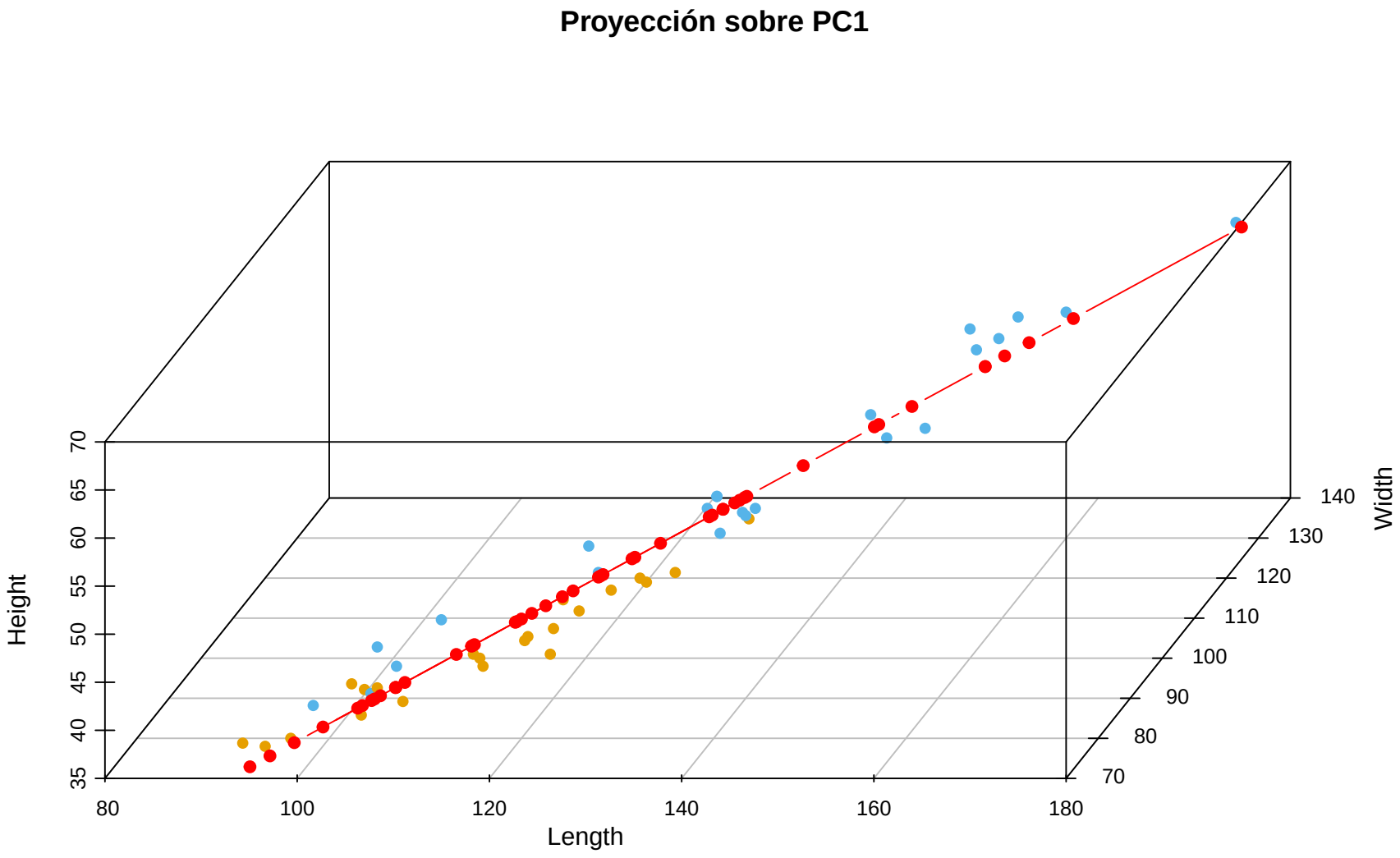
```
## eigen() decomposition
## $values
## [1] 641.577533   5.204385   3.830670
##
## $vectors
##            [,1]       [,2]        [,3]
## [1,] 0.8068646  0.5865919 -0.06985282
## [2,] 0.4947448 -0.7356259 -0.46269006
## [3,] 0.3227958 -0.3387689  0.88376382
```

```
vect_pc1 <- PC$vectors[,1]
```

Proyectamos sobre pc1 para calcular el 1er factor

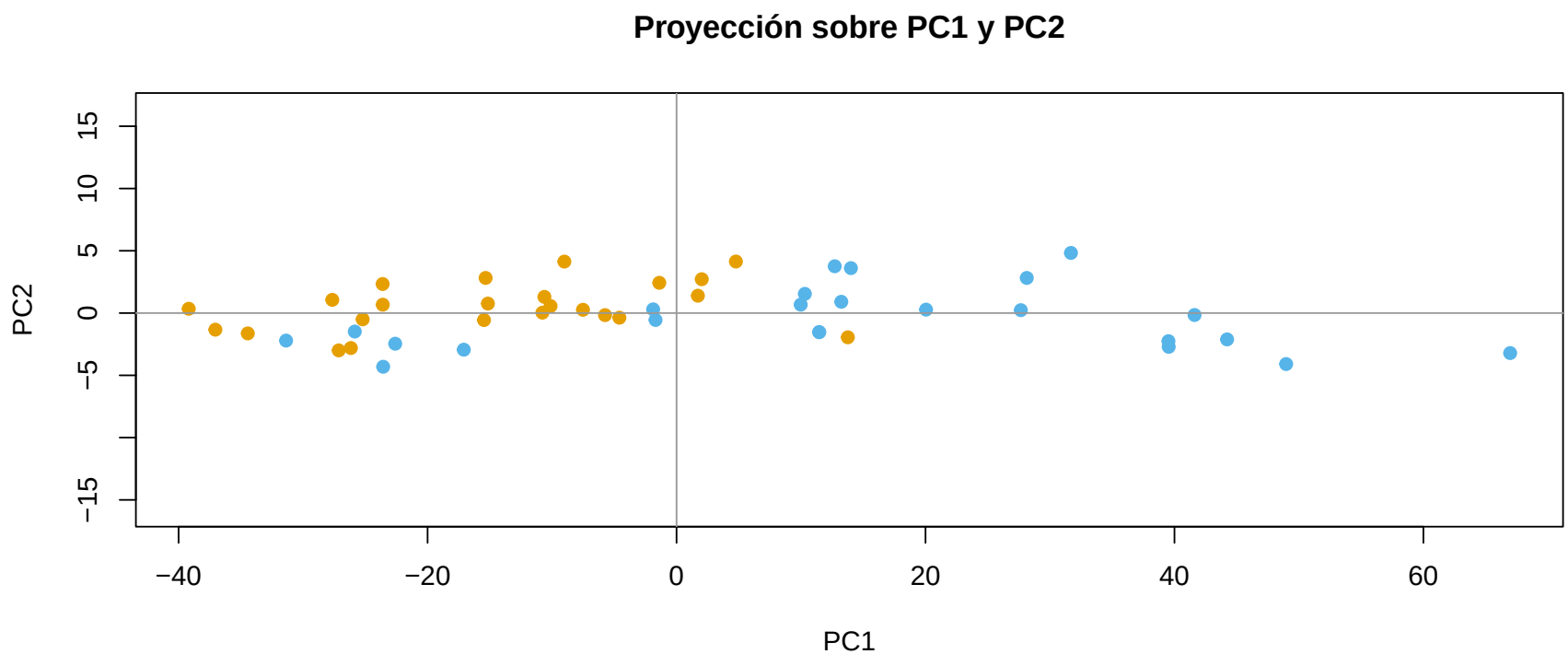
```
(f1 <- as.numeric(X_c%*%vect_pc1))
```

Graficamos la proyección sobre PC1 de los caparazones



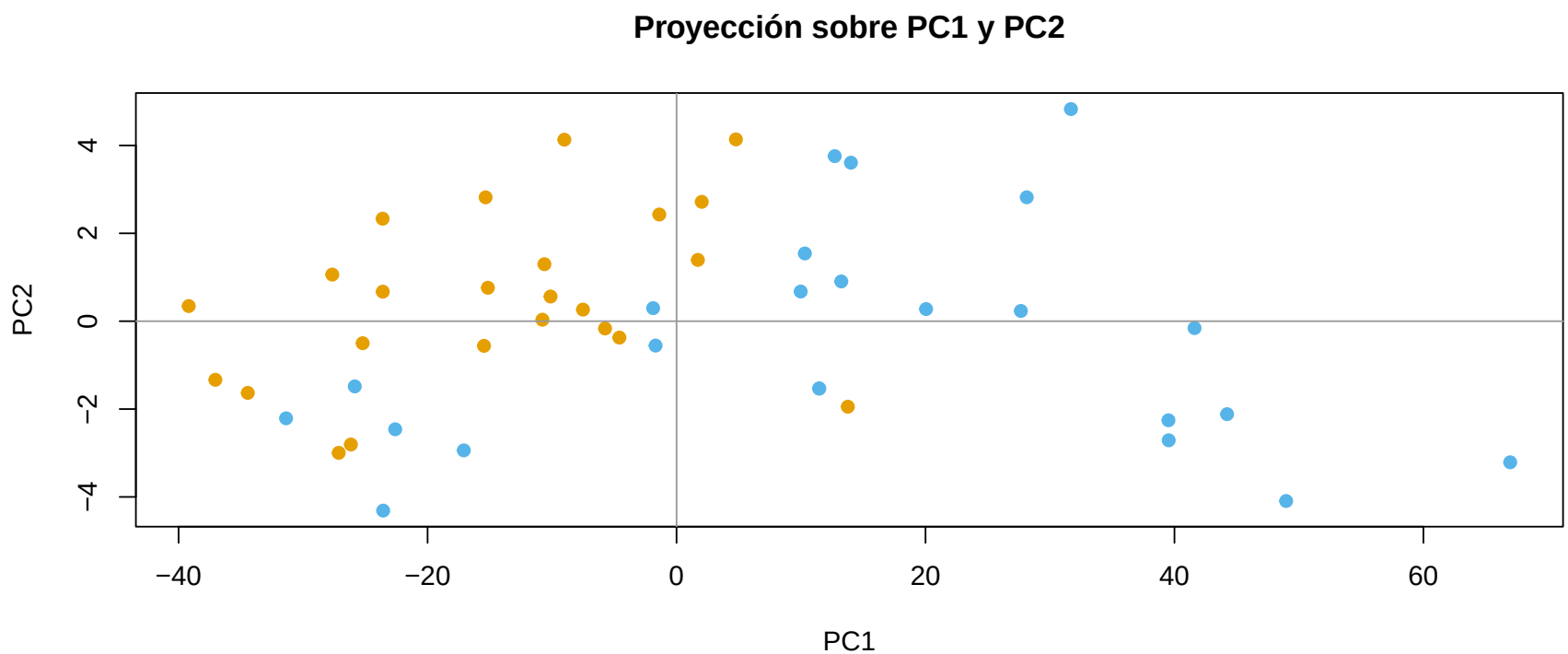
Proyección de los caparazones en el plano

```
vect_pc2 <- PC$vectors[,2]
f2 <- as.numeric(X_c%*%vect_pc2)
plot(f1,f2,
     pch=19,col=vector_colores,
     main = "Proyección sobre PC1 y PC2",
     xlab = "PC1",ylab = "PC2", asp=1)
abline(h = 0, v = 0, col = "gray60")
```



Proyección de los caparazones en el plano

```
vect_pc2 <- PC$vectors[,2]
f2 <- as.numeric(X_c%*%vect_pc2)
plot(f1,f2,
     pch=19,col=vector_colores,
     main = "Proyección sobre PC1 y PC2",
     xlab = "PC1",ylab = "PC2")
abline(h = 0, v = 0, col = "gray60")
```



Calidad de la representación

Variabilidad total y calidad de la representación

Una medida usual para la variabilidad total de la nube de puntos es la suma de las varianzas de cada variable:

$$\text{Variabilidad total} = \mathrm{Var}(1) + \cdots + \mathrm{Var}(K)$$

Este número es igual a la traza de la matriz de covarianzas S y se puede calcular como:

$$\mathrm{tr}(S) = \lambda_1 + \cdots + \lambda_K$$

Si usamos una representadión con $n \leq K$ componentes principales, el coeficiente

$$\frac{\lambda_1 + \cdots + \lambda_n}{\mathrm{tr}(S)} \times 100$$

es un indicador de la calidad de la representación.

Calidad de la representación en el ejemplo de las tortugas

```
X.princomp <- princomp(X)
cumsum(round(X.princomp$sdev^2/(sum(X.princomp$sdev^2))*100
```

```
## Comp.1 Comp.2 Comp.3
##   98.6   99.4  100.0
```

Observación:

```
X.princomp$sdev
```

```
##    Comp.1    Comp.2    Comp.3
## 25.064144  2.257423  1.936715
```

```
sqrt(PC$values)
```

```
## [1] 25.329381  2.281312  1.957210
```

Interpretación de las componentes

Los valores proyectados como variable

Consideremos la matriz de datos centrada X y la componente principal pc_s, para $1 \leq s \leq K$. Si proyectamos la nube de individuos sobre pc_s obtenemos un vector de I valores

$$f_s = Xpc_s = (f_s(1), \cdots, f_s(I))$$

en donde $f_s(i)$ es el valor de la proyección en el individuo i. Podemos pensar a f_s como una nueva variable que es una combinación lineal de las variables originales. Se la suele llamar *componente principal de rango s*.

La correlación como coordenadas

Podemos calcular la correlación de la variable k con la componente s como indicador de la influencia que ésta tiene sobre la componente. Si calculamos la correlación de la variable k en f_1 y f_2 podemos usarlas como coordenadas para representarla en un plano.

Interpretación de las componentes

Correlación nula entre componentes

Observar que la correlación entre f_s y f_t es nula:
$f_s \cdot f_t = (Xpc_s) \cdot (Xpc_t) = (Xpc_s)^t (Xpc_t) = (pc_s)^t (X^t X)(pc_t)$

$$f_s \cdot f_t = (I\lambda_t)(pc_s)^t (pc_t) = 0$$

Varianza de una componente

Observar también que la varianza de la componente f_s es igual a λ_s.

Correlación entre k y f_1 y f_2

Para cada variable k calculamos los coeficientes de correlación

$$r(k, f_1) \quad \text{y} \quad r(k, f_2)$$

para representar k como un vector en el plano.

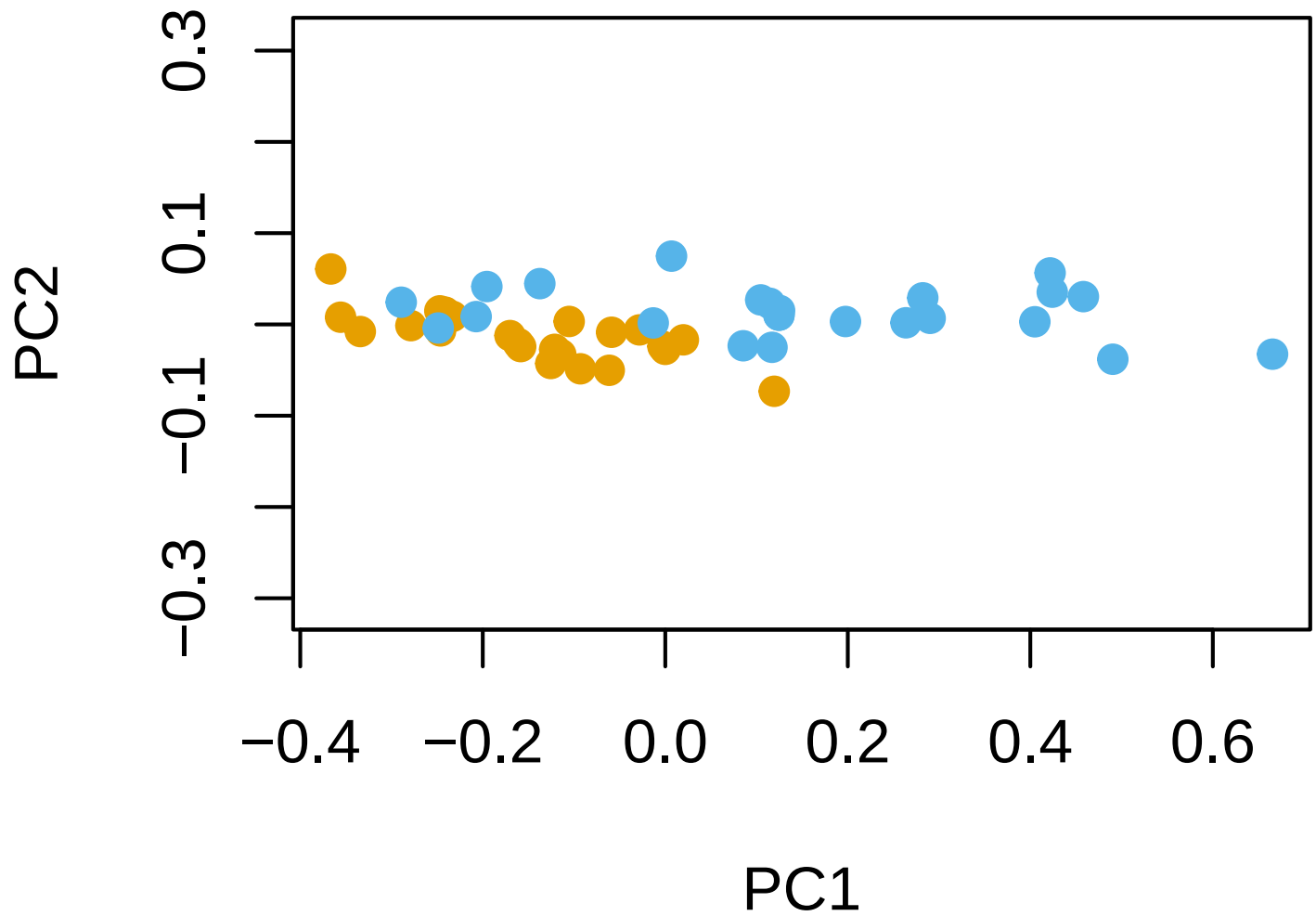
Calculemos las componentes

```
S_e <- t(X_e)%*%X_e
PC_e <- eigen(S_e,symmetric = T)
vect_pc1 <- PC_e$vectors[,1]
vect_pc2 <- PC_e$vectors[,2]

f1 <- as.numeric(X_e%*%vect_pc1)
f2 <- as.numeric(X_e%*%vect_pc2)
```

Nube de individuos en $\mathbb{R}^K$ (Recordamos)

```
plot(f1,f2, pch=19,col=vector_colores,
     xlab = "PC1",ylab = "PC2",asp = 1)
```

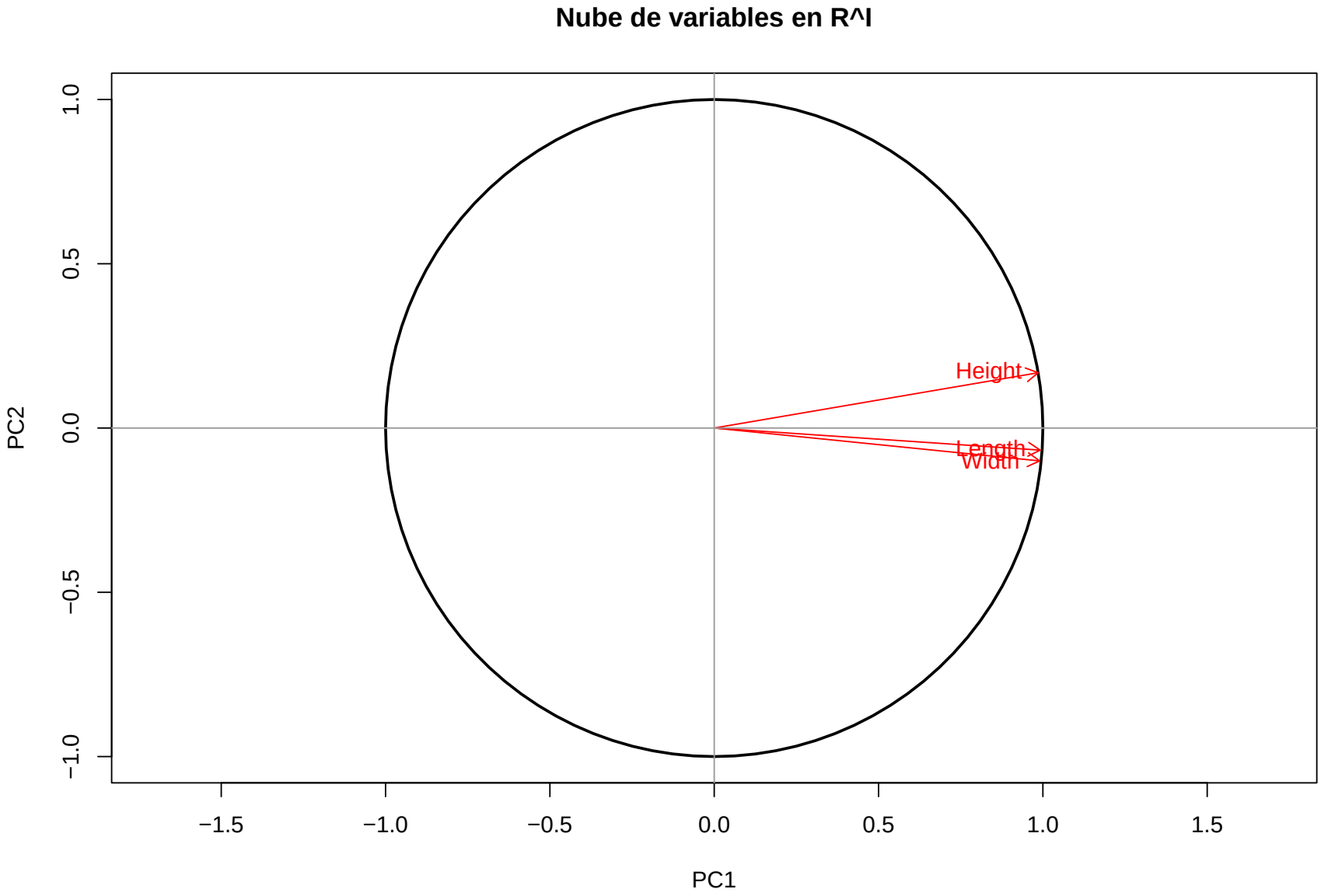


Las coordenadas de la nube de variables

```
library(plotrix)
pc12 <- X_e%*%PC_e$vectors[,1:2]
(R <- cor(x=X_e,y=pc12))
```

```
##                [,1]         [,2]
## Length 0.9917124 -0.06727662
## Width  0.9903175 -0.10007227
## Height 0.9856549  0.16823574
```

Nube de variables



La nube de variables

Las columnas de la matriz de datos

Recordar que cada variable es una columna de la matriz X, y que la consideramos como un vector del espacio $\mathbb{R}^I$ de variables. Vamos a asumir que X está centrada.

El producto escalar entre dos variables k y l:

$k \cdot l = \sum_{i=1}^{I}(x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l) = \|k\|\|l\| \cos(\theta_{kl}) = I s_k s_l r(k,l)$ en donde $\theta_{kl} \in [0, \pi]$ es el ángulo entre k y l. Recordar que el coeficiente de correlación lo identificamos con el coseno del ángulo:

$$r(k,l) = \cos(\theta_{kl}).$$

Los dos puntos de vista

Los cometidos de la nube de puntos

1. Evaluar la estructura de los individuos (distinguir diferentes grupos, si los hay).
2. Reconocer las direcciones de mayor variabilidad.

Los cometidos de la nube de variables

1. Evaluar o distinguir diferentes grupos de variables
2. Visualizar la matriz de correlaciones a través de los ángulos.

Ajustando la nube de variables

Estandarización

Para mejor visualizar los ángulos en la nube de variables, es conveniente estandarizar la matriz de datos de forma que todas las variables tengan norma 1. Esto equivale a considerar la entrada ik de X como siendo:

$$X_{ik} = \frac{x_{ik} - \bar{x}_k}{\sqrt{I} s_k}$$

Con esta estandarización el producto escalar entre dos variables es igual a su correlación:

$$k \cdot l = r(i, l) = cos(\theta_{kl})$$

Criterio de optimización

Un vector de $\mathbb{R}^I$ puede pensarse como una variable. El criterio consiste en buscar la dirección v en $\mathbb{R}^I$ (o sea $\|v\| = 1$) que maximiza la correlación promedio:

$$\max_{v:\|v\|=1} \sum_{k=1}^{K} r(k, v)^2$$

Es decir, buscamos la variable v que está más correlacionada con las K variables originales (en promedio).

Proyección de una variable sobre otra

Como asumimos X estandarizada, las variables tienen norma 1, por lo que proyectar k sobre v equivale a considerar el producto escalar

$$k \cdot v = \cos(\theta_{kv}) = r(k, v)$$

Podemos entonces pensar que el criterio consiste en buscar la dirección para la cual los puntos proyectados se encuentran lo más alejados del origen que sea posible (en promedio).

La proyección de la nube entera

Proyectar la nube entera equivale al producto de matrices $X^t v$, que resulta un vector columna en donde cada entrada es igual a $k \cdot v$. El criterio expresado en forma matricial resulta ser

$$\max_{v:\|v\|=1} \|X^t v\|^2.$$

Ajustando la nube de variables

Resulta familiar

Observar que $\|X^t v\|^2 = (X^t v)^t (X^t v) = v^t (XX^t) v$. La matriz XX^t es una matriz simétrica de tamaño $I \times I$ y el problema de optimización para la nube de variables resulta ser análogo que el de la nube de individuos, salvo que cambiando $X^t X$ por XX^t.

La solución

Al igual que antes, la solución está dada por los vectores propios asociados a los valores propios de XX^t ordenados de mayor a menor: $\mu_1 \geq \cdots \geq \mu_I$.

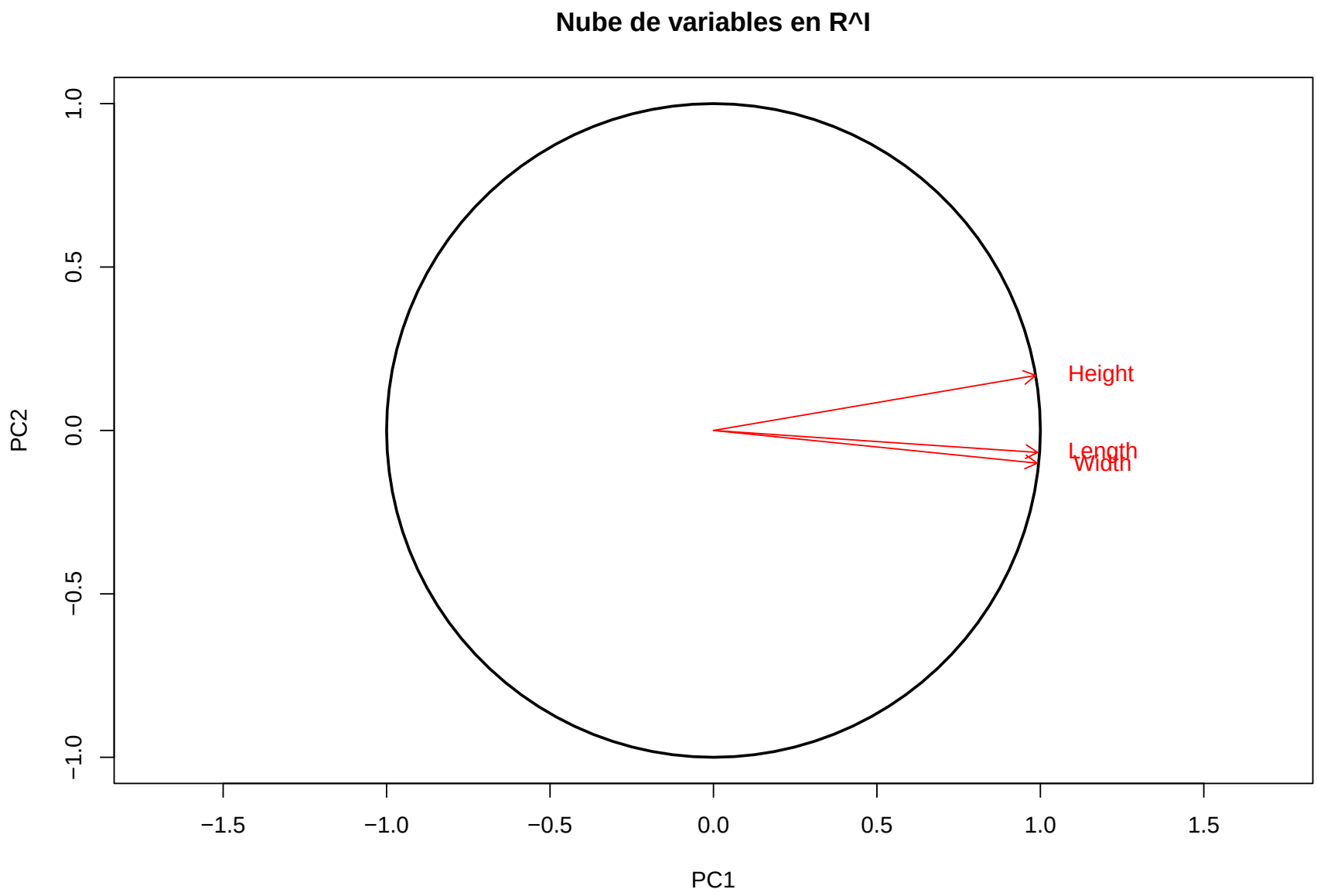
En el ejemplo de las tortugas

Si observamos las coordenadas de las variables en el plano generado por v_1 y v_2, los dos primeros vectores propios de XX^t, obtenemos:

```
A <- X_e%*%t(X_e)
PCV <- eigen(A,symmetric = T)
(coords<-t(X_e)%*%PCV$vectors[,1:2])
```

```
##                [,1]         [,2]
## Length 0.9917124 -0.06727662
## Width  0.9903175 -0.10007227
## Height 0.9856549  0.16823574
```

Notar las semejanzas con lo anterior.



La relación entre ambos puntos de vista

Volvamos a hacer las correlaciones con las componentes f_1 y f_2 pero haciendo el cálculo con la matriz de datos estandarizada del mismo modo que en las últimas diapositivas.

```
t(X_e)%*%PCV$vectors[,1:2]
```

```
##                [,1]         [,2]
## Length 0.9917124 -0.06727662
## Width  0.9903175 -0.10007227
## Height 0.9856549  0.16823574
```

```
cor(x=X_e,y=pc12)
```

```
##                [,1]         [,2]
## Length 0.9917124 -0.06727662
## Width  0.9903175 -0.10007227
## Height 0.9856549  0.16823574
```

Puede haber diferencia de signos

La diferencia de signo es arbitraria, pues depende de la elección de dirección de los vectores propios. Eligiendo las direcciones de forma correcta ambas matrices son iguales.

De una nube a la otra

Supongamos de ahora en más que estandarizamos X de forma tal que

$$X_{ik} = \frac{x_{ik} - \bar{x}_k}{\sqrt{I} s_k}$$

Esto es para que las variables tengan norma 1 y como consecuencia los productos escalares sean iguales a las correlaciones.

Consideremos una componente principal pc_s y su variable asociada $f_s = X\, pc_s$. Notar que
$(XX^t)f_s = (XX^t)(X\, pc_s) = X(X^t X\, pc_s) = X(\lambda_s pc_s) = \lambda_s (X\, pc_s)$

$(XX^t)f_s = \lambda_s f_s$ Esto quiere decir que λ_s es también un valor propio y que f_s es un vector propio de XX^t.

La norma de f_s es $\sqrt{\lambda_s}$, por lo que $\frac{1}{\sqrt{\lambda_s}} f_s$ debe ser igual a uno de los v_t. Como están ordenados por tamaño:

$$v_1 = \frac{1}{\sqrt{\lambda_1}} f_1, \quad v_2 = \frac{1}{\sqrt{\lambda_2}} f_2, \quad \ldots$$

De una nube a la otra

Al calcular las correlaciones de la variable k con las f_s tenemos

$$r(k, f_s) = \frac{k \cdot f_s}{\sqrt{\lambda_s}} = k \cdot v_s = (\text{fila } k \text{ de } X^t) \cdot v_s$$

Cuando lo hacemos para la nube entera de K variables el resultado es $X^t v_s$.

Correlaciones de las variables con las componentes principales

Por ello, al hacerlo con f_1 y f_2, si $[v_1, v_2]$ denota la matriz de $I \times 2$ cuyas dos columnas son v_1 y v_2 obtenemos que las correlaciones de las K variables con f_1 y f_2 son exactamente las coordenadas de las variables proyectadas sobre el plano óptimo generado por v_1 y v_2:

$$(\text{correlaciones de todas las variables con } f_1 \text{ y } f_2) = X^t[v_1, v_2]$$

De una nube a la otra

Llamemos $r_s(k) = (\text{fila } k \text{ de } X^t) \cdot v_s$ a la coordenada de la variable k en la componente s. El vector $r_s = X^t v_s$ de $\mathbb{R}^K$ representa entonces las coordenadas de las K variables en la componente s.

Relaciones de transición

La relación anterior, que describe cómo pasar de las coordenadas f_s a las r_s, se escribe

$$r_s = \frac{1}{\sqrt{\lambda_s}} X^t f_s$$

Del mismo modo se muestra que el pasaje de r_s a f_s es

$$f_s = \frac{1}{\sqrt{\lambda_s}} X r_s$$

Correlación y distancia entre variables

La distancia entre las variables k y l está dada por

$$d(k,l)^2 = \|k-l\|^2 = \sum_{i=1}^{I}(x_{ik}-x_{il})^2$$

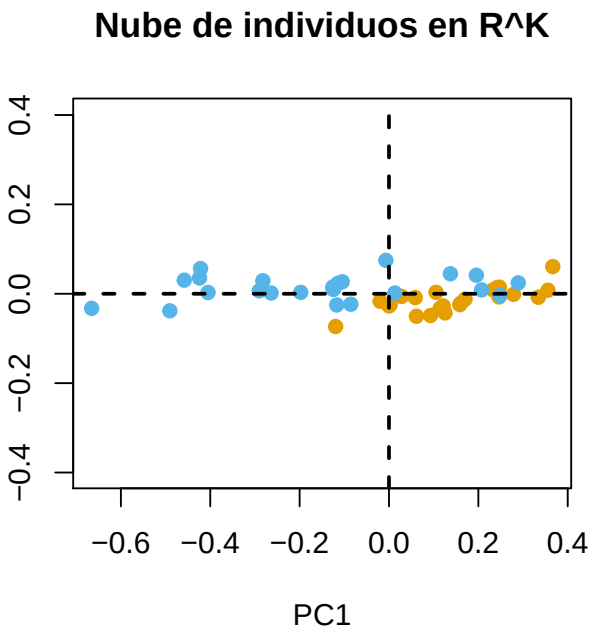
Desarrollando el cuadrado y reordenando los términos obtenemos

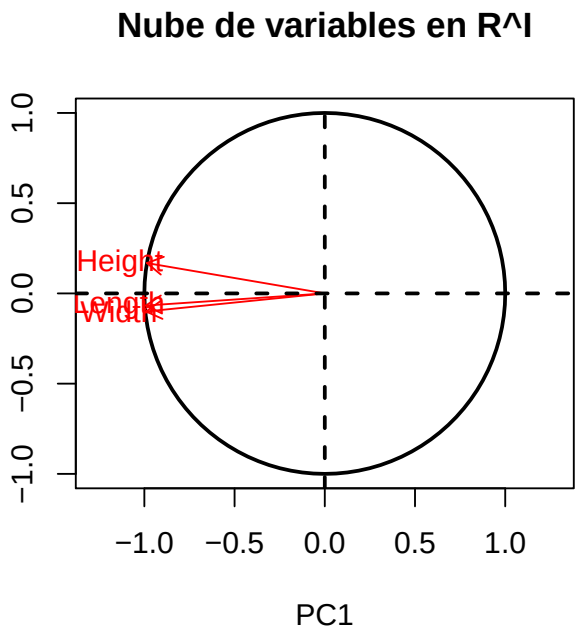
$$d(k,l)^2 = \|k\|^2 + \|l\|^2 - 2(k\cdot l) = \sum_{i=1}^{I} x_{ik}^2 + \sum_{i=1}^{I} x_{il}^2 - 2\sum_{i=1}^{I} x_{ik}x_{il}$$

Recordar que bajo la estandarización que convenimos al principio $\|k\| = \|l\| = 1$ y que el producto escalar $k\cdot l$ es igual a la correlación $r(k,l)$. Entonces

$$d(k,l)^2 = 2(1-r(k,l))$$

Representaciones separadas





El análisis de la nube de I individuos en $\mathbb{R}^K$ se hace en base al sistema de coordenadas $\{pc_1, pc_2\}$. La representación provee la mejor visualización aproximada de las distancias entre los individuos.

El análisis de la nube de K variables en $\mathbb{R}^I$ se hace en base al sistema de coordenadas $\{v_1, v_2\}$. La representación provee la mejor visualización aproximada de la matriz de correlaciones.