# Análisis de Componentes Principales (PCA)

José A. Perusquía Cortés

Análisis Multivariado Semestre 2024-I

# Intuición

- Motivación

  Visualizar y/o interpretar datos multivariados es complicado

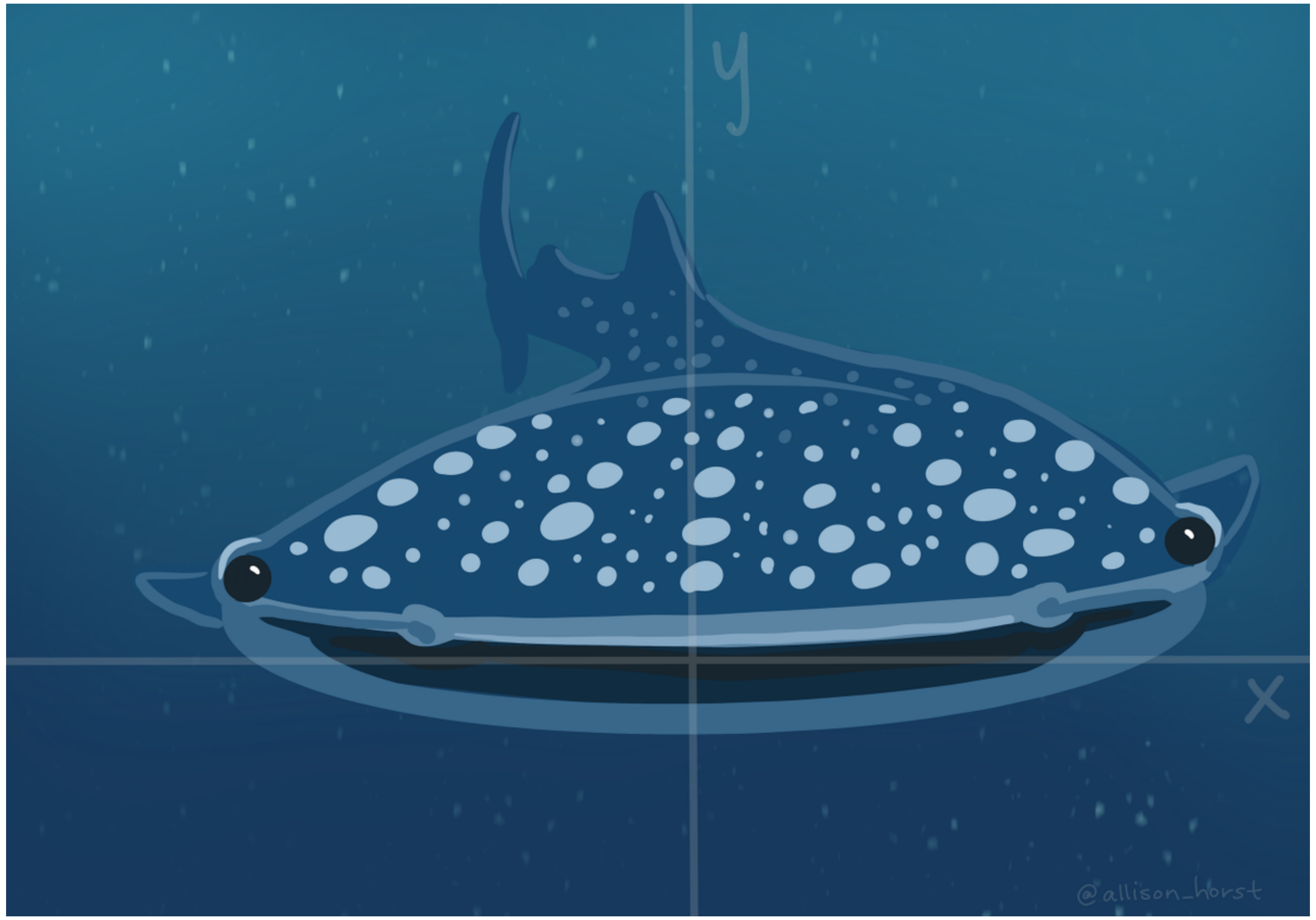
- A grandes rasgos PCA es un método estadístico que busca

  1. “Reducir” la dimensionalidad de los datos
  2. Retener la mayor cantidad de la variación original

- ¿Cómo?

  Crear un nuevo conjunto de variables no correlacionadas y ordenadas por varianza

- ¿Cómo debe girar la cabeza la ballena para comer la mayo cantidad de kril?





Fuente: Allison Horst (twitter)

# Construcción (Idea)

- Sea $\mathbf{x}_{p\times 1}$ un vector aleatorio real valuado

- El primer componente principal estará dado por

$$\alpha_1^T x = \alpha_{11}x_1 + \alpha 12 x_2 + \cdots + \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j$$

tal que sea la combinación lineal de **mayor varianza.**

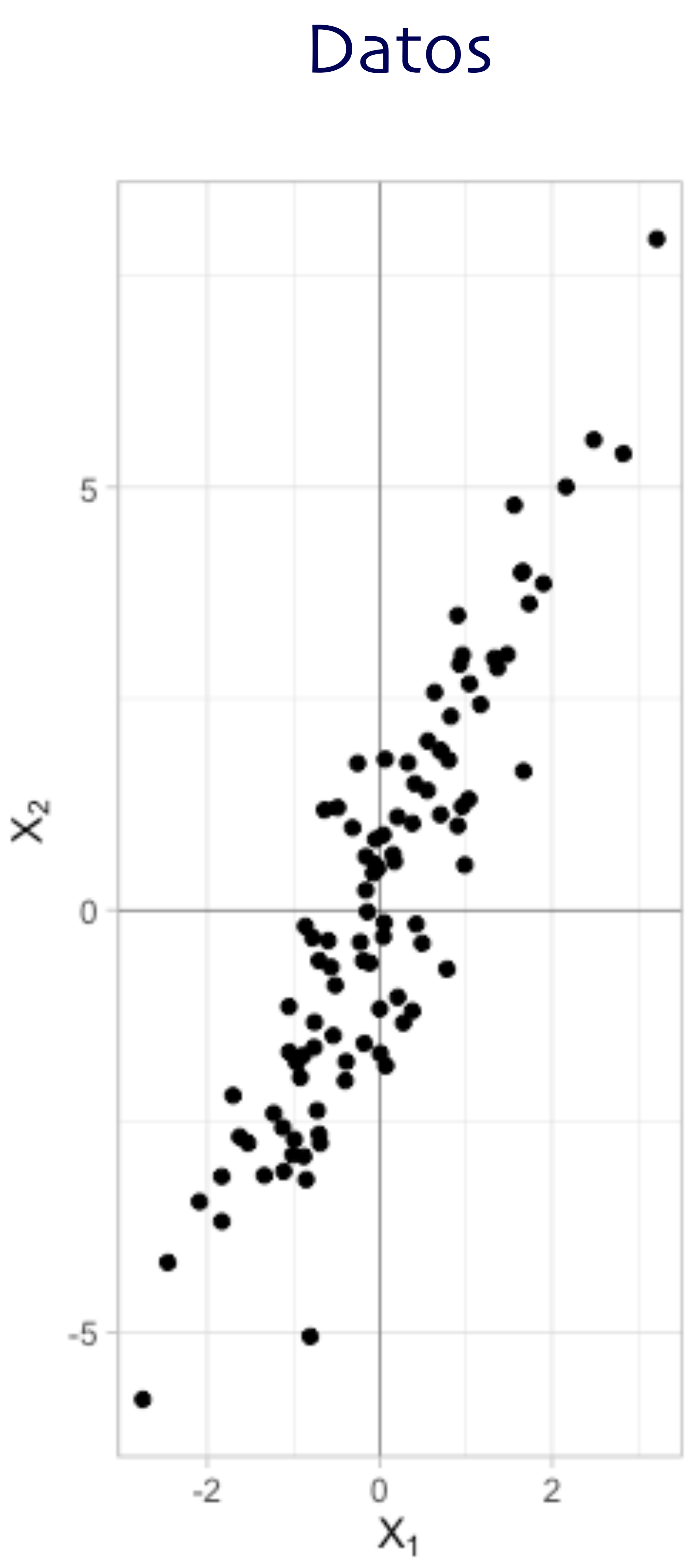
- El segundo componente principal está dado por

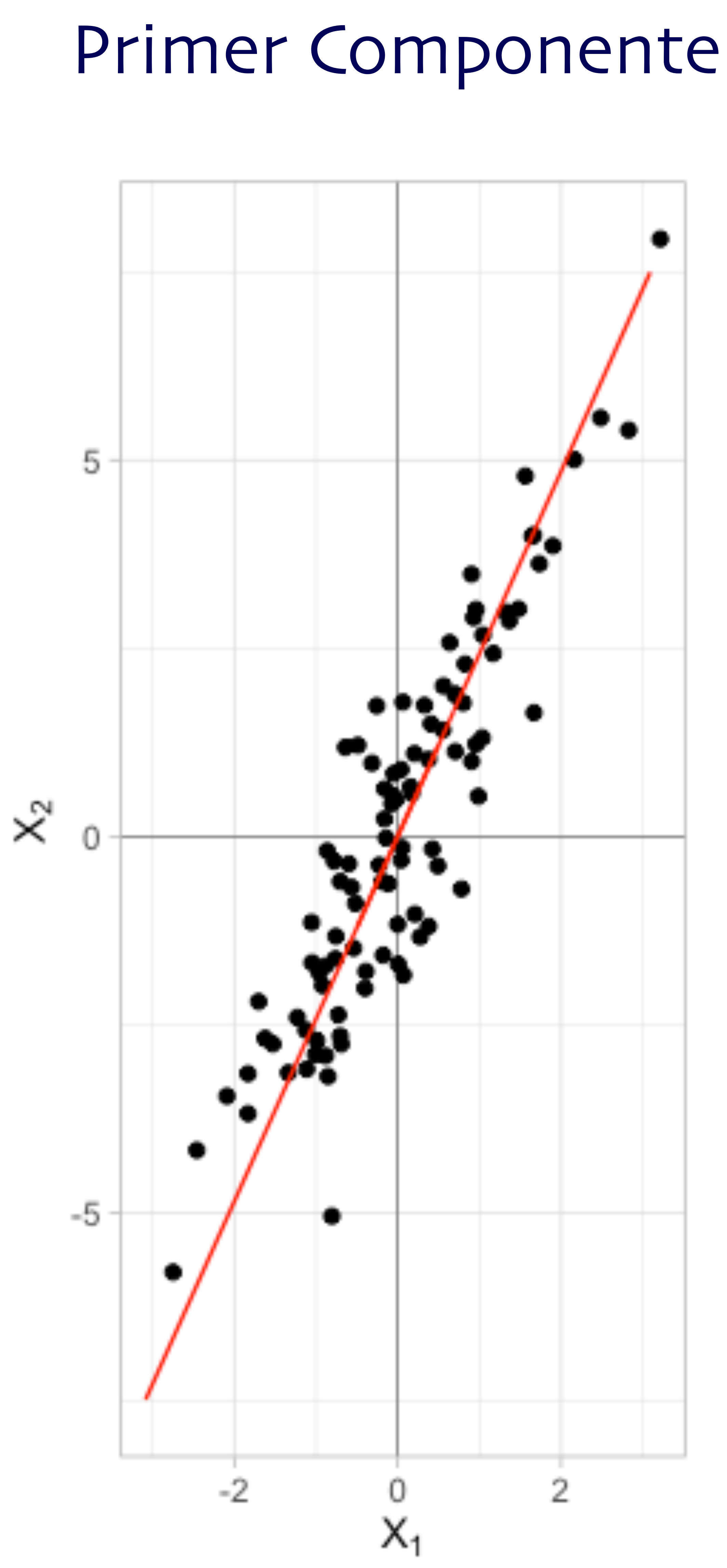
$$\alpha_2^T x = \alpha_{21}x_1 + \alpha 22 x_2 + \cdots + \alpha_{2p}x_p = \sum_{j=1}^{p} \alpha_{2j}x_j$$

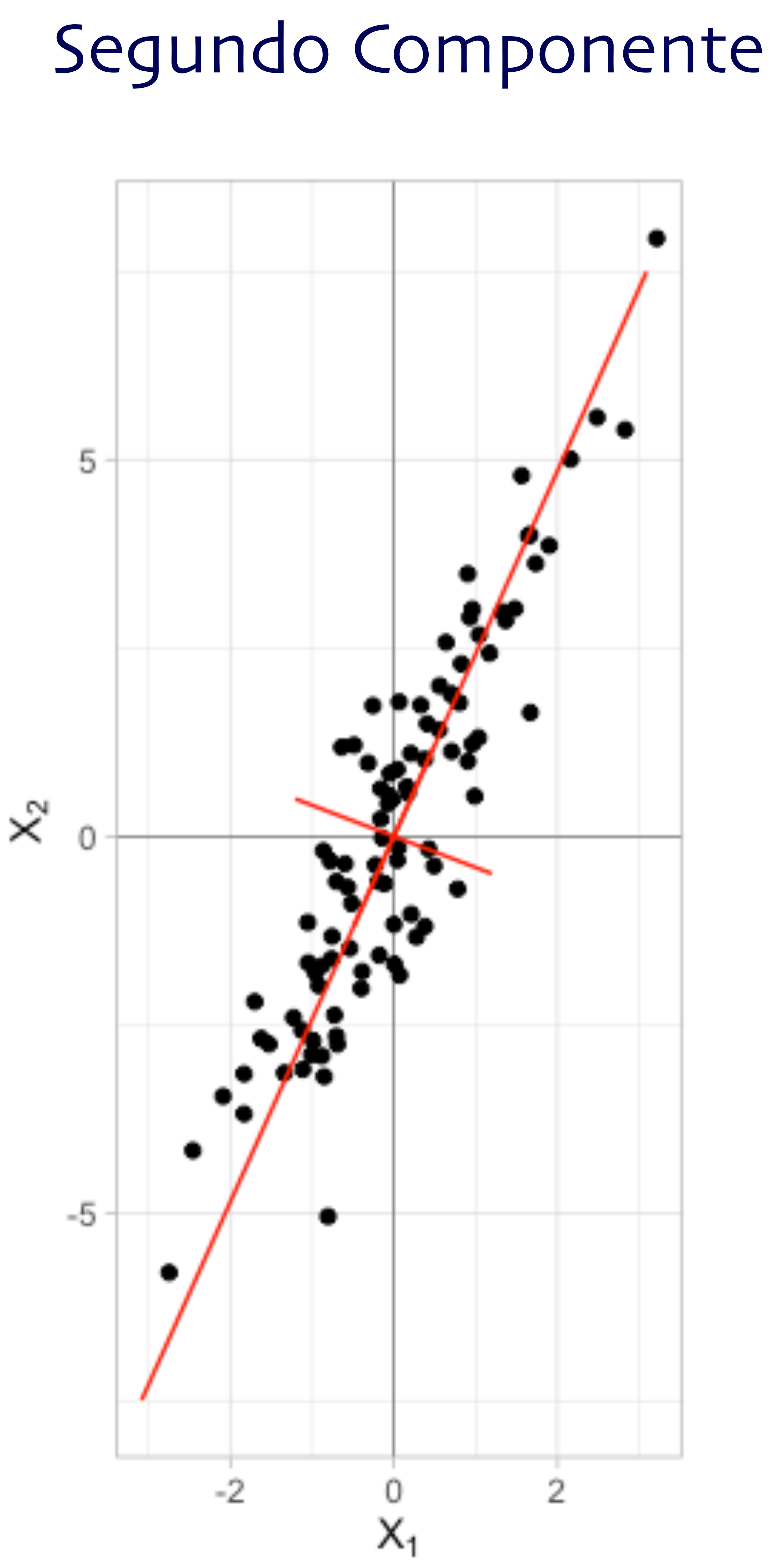
tal que sea la combinación lineal de **mayor varianza** y **no esté correlacionado** con el primero.

- Y así sucesivamente…

# Ejemplo

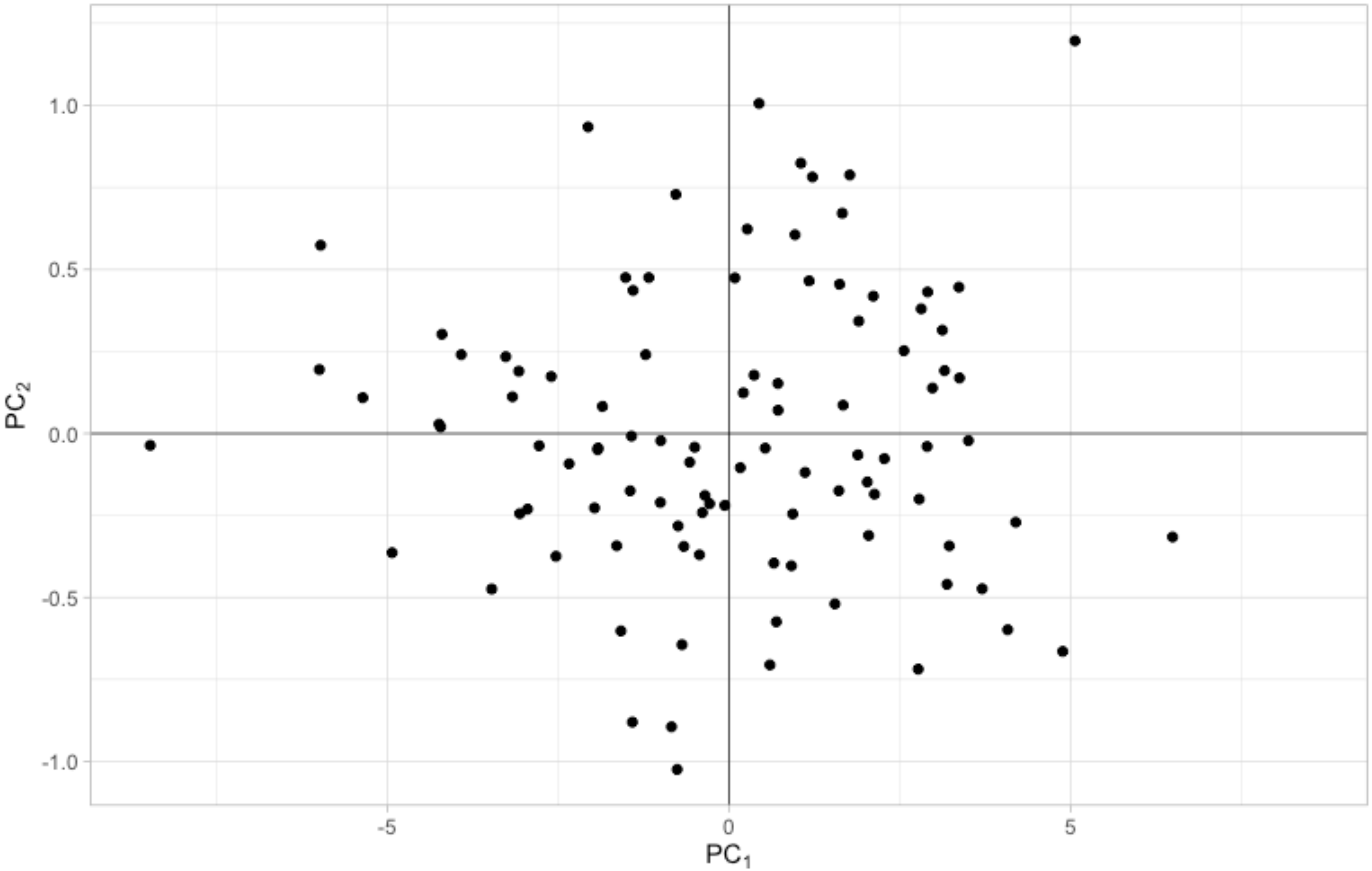

Datos
5
0
-5
-2
0
2
$X_1$
$X_2$


Primer Componente
5
0
-5
-2
0
2
$X_1$
$X_2$


Segundo Componente
5
0
-5
-2
0
2
$X_1$
$X_2$

Ejemplo

- Nuevas variables



- Sea $\mathbf{x}_{p\times 1}$ un vector aleatorio real valuado, con $\Sigma$ conocida

- (Formalmente) el primer componente principal se encuentra resolviendo

$$\max_{\alpha_1} \quad \text{var}\left(\alpha_1^T x\right) = \alpha_1^T \Sigma \alpha_1$$

$$\text{s.a.} \quad \alpha_1^T \alpha_1 = 1$$

- Dando como resultado que

  $\lambda$: eigenvalor más grande

  $\alpha_1$: eigenvector asociado

- Para el segundo componente resolvemos:

$$\begin{aligned} \max_{\alpha_2} \quad & \text{var}\left(\alpha_2^T x\right) = \alpha_2^T \Sigma \alpha_2 \\ \text{s.a.} \quad & \alpha_2^T \alpha_2 = 1 \\ & \text{cov}\left(\alpha_1^T \mathbf{x}, \alpha_2^T \mathbf{x}\right) = 0 \end{aligned}$$

- Dando como resultado que

  $\lambda$ : segundo eigenvalor más grande

  $\alpha_2$ : eigenvector asociado

- Y así sucesivamente…

- Los componente principales corresponden a una transformación ortogonal de $\mathbf{x}$

$$\mathbf{z} = \mathbf{A}\mathbf{x}$$

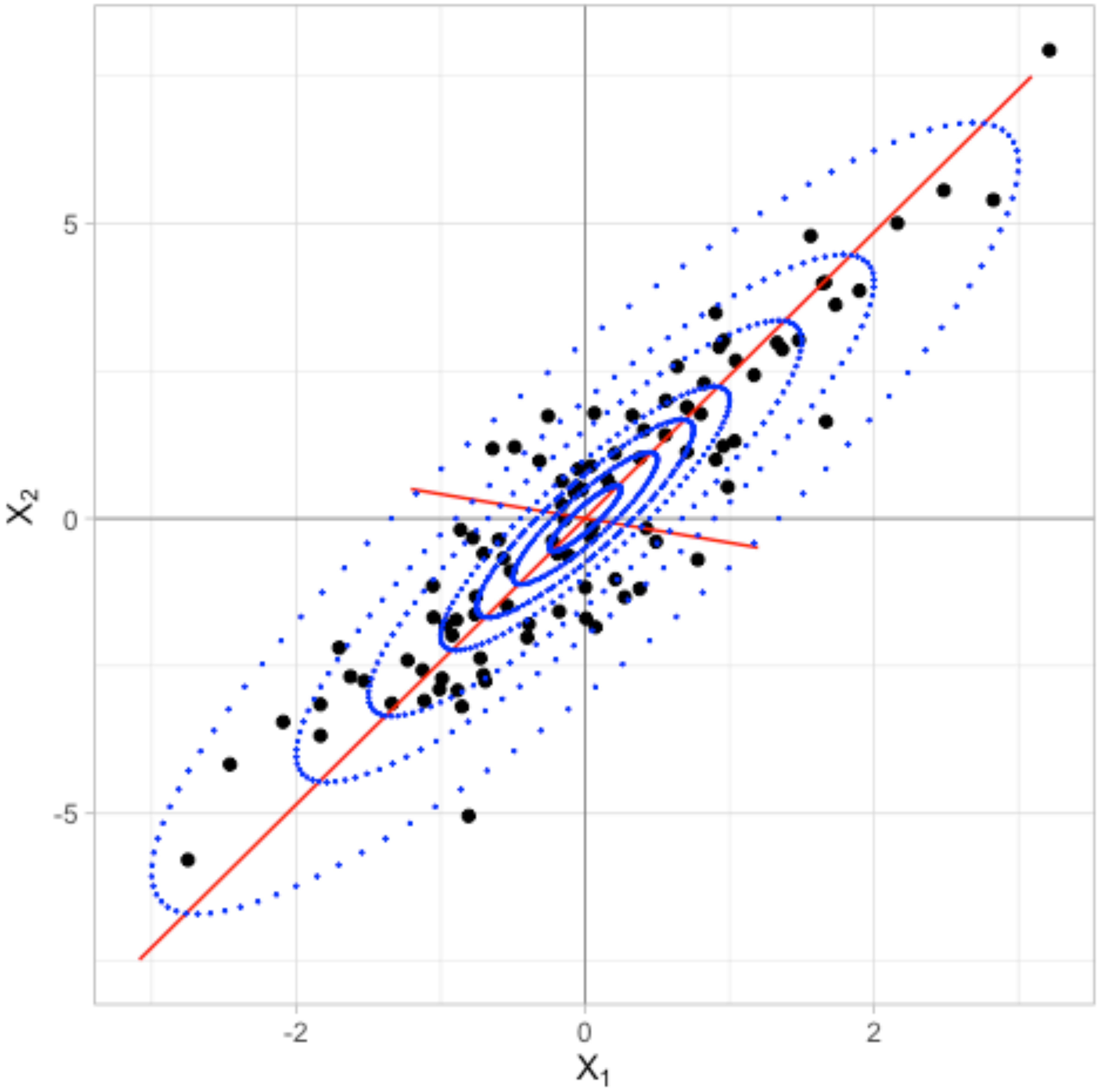
- Donde $\mathbf{A}$ es la matriz de eigenvectores

- Así, $\mathrm{var}\left(z_k\right) = \lambda_k$

**Proposición**

Sea la familia de elipsoides $\mathbf{x}^T\Sigma^{-1}\mathbf{x} = c$. Entonces los componentes principales definen los ejes principales.

# Interpretación geométrica



**Proposición (A1)**

Sea la transformación ortogonal $\mathbf{y} = \mathbf{B}^T\mathbf{x}$. Donde $\mathbf{B}_{q\times p}$ y $\Sigma_y = \mathbf{B}^T\Sigma\mathbf{B}$ entonces,

1. $\mathrm{tr}(\Sigma_y)$ y $|\Sigma_y|$ se maximizan cuando $\mathbf{B} = \mathbf{A}_q$ (las primeras q columnas)
2. $\mathrm{tr}(\Sigma_y)$ se minimiza cuando $\mathbf{B} = \mathbf{A}_q^*$ (las últimas q columnas)

**Proposición (A2)**

La descomposición espectral de $\Sigma$ está dada por $\Sigma = \sum_{i=1}^{p} \lambda_i \alpha_i \alpha_i^T$.

**Proposición (A3)**

Si $\sigma_j^2$ es la varianza residual de predecir $x_j$ en términos de $\mathbf{y}$, entonces $\sum \sigma_i^2$ se minimiza cuando $\mathbf{B} = \mathbf{A}_q$.

# Componentes vía matriz de correlación

- En la práctica es más común definir a los componentes como

$$\mathbf{z} = \mathbf{A}\mathbf{x}^*$$

donde $\mathbf{x}^*$ son las variables estandarizadas y $\mathbf{A}$ es la matriz de eigenvectores de la matriz de correlación

**Observaciones**

1. Todas las propiedades anteriores siguen siendo válidas
2. Se pueden mezclar variables en diferentes escalas
3. Los componentes no están dominados por una posible variable de mayor varianza

- Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra aleatoria (centrados) con matriz de varianzas y covarianzas

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

1. El primer componente principal es el eigenvector $\mathbf{a}_1$ asociado al eigenvalor más grande
2. Se tienen $n$ nuevas variables $z_{i1} = \mathbf{a}_1^T\mathbf{x}_i$
3. Y sucesivamente para los otros componentes

**Observaciones**

1. Las vectores $\mathbf{z}_i$ se les conoce como scores
2. Los eigenvectores $\mathbf{a}_i$ se les conoce como loadings

- En muchas ocasiones es preferible usar la descomposición en valores singulares (SVD) para encontrar los componentes principales

$$\mathbf{S} = \frac{1}{n-1}\mathbf{W}^T\mathbf{W} \qquad \mathbf{W} = \mathbf{UDV}^T \qquad \mathbf{S} = \mathbf{V}\Lambda\mathbf{V}^T$$

1. Numéricamente más estable.
2. Permite considerar el caso $p > n$
3. Puede ser más rápido

**Proposición**

Sea $\mathbf{Z} = \mathbf{HXV}$ la matriz de cargas, i.e., $\mathbf{z}_i = \mathbf{V}^T(\mathbf{x}_i - \bar{\mathbf{x}})$ entonces se cumple

1. La media muestral es el vector de ceros
2. La matriz de varianza y covarianza es $\Lambda$
3. $\mathbf{v_1}^T\mathbf{S}\mathbf{v_1} > \mathbf{v_2}^T\mathbf{S}\mathbf{v_2} > \cdots > \mathbf{v_p}^T\mathbf{S}\mathbf{v_p}$ y si $\text{ran}(S) = q < p$ se tiene que $\mathbf{v_s}^T\mathbf{S}\mathbf{v_s} = 0$ para $s = q+1,\ldots,p$
4. $\sum_{i=1}^{p} \mathbf{v_i}^T\mathbf{S}\mathbf{v_i} = \sum_{i=1}^{p} \lambda_i = \text{tr}(S)$
5. $\prod_{i=1}^{p} \mathbf{v_i}^T\mathbf{S}\mathbf{v_i} = \prod_{i=1}^{p} \lambda_i = |S|$

# Ejemplo 2: Calificaciones

- 88 calificaciones de 5 exámenes a libro abierto o cerrado

| **Lineal (C)** | **Estadística (C)** | **Probabilidad(A)** | **Finanzas (A)** | **Cálculo (A)** |
|---|---|---|---|---|
| 97 | 92 | 77 | 72 | 96 |
| 83 | 88 | 90 | 75 | 96 |
| 95 | 83 | 81 | 71 | 96 |
| 75 | 82 | 73 | 75 | 83 |
| 83 | 73 | 75 | 75 | 78 |
| 73 | 71 | 82 | 69 | 88 |
| 71 | 77 | 75 | 70 | 83 |

- En R usamos prcomp() con $\widehat{\Sigma} = \mathbf{S}$

- Los eigenvalores resultantes son

$$\lambda_1 = 689.6583 > \lambda_2 = 200.9016 > \lambda_3 = 103.5280 > \lambda_4 = 83.3404 > \lambda_5 = 32.2476$$

- Los vectores de cargas:

| | | | | | |
|---|---|---|---|---|---|
| **Lineal** | -0.502 | -0.759 | 0.289 | -0.284 | -0.080 |
| **Estadística** | -0.371 | -0.188 | -0.417 | 0.785 | -0.186 |
| **Probabilidad** | -0.345 | 0.077 | -0.144 | -0.002 | 0.923 |
| **Finanzas** | -0.450 | 0.299 | -0.591 | -0.523 | -0.287 |
| **Cálculo** | -0.535 | 0.541 | 0.609 | 0.164 | -0.149 |

# Ejemplo 2: Calificaciones

- El primer componente es un "promedio"

$$-0.502 \cdot \text{Lineal} - 0.371 \cdot \text{Estadística} - 0.345 \cdot \text{Probabilidad} - 0.450 \cdot \text{Finanzas} - 0.535 \cdot \text{Cálculo}$$

- El segundo componente es una comparación entre libro abierto y cerrado

$$-0.759 \cdot \text{Lineal} - 0.188 \cdot \text{Estadística} + 0.077 \cdot \text{Probabilidad} + 0.299 \cdot \text{Finanzas} + 0.541 \cdot \text{Cálculo}$$

- El tercer componente es una comparación entre matemáticas "puras y aplicadas"

$$0.289 \cdot \text{Lineal} - 0.417 \cdot \text{Estadística} - 0.144 \cdot \text{Probabilidad} - 0.591 \cdot \text{Finanzas} + 0.609 \cdot \text{Cálculo}$$

- La interpretación requiere conocimiento del problema

- Algunos componentes pueden interpretarse como un promedio ponderado

- Algunos componentes pueden discriminar entre grupos de variables

- **¿Cuántos componentes elegir?**

# Número de componentes

- Seleccionar los componentes que expliquen un cierto porcentaje de la variación (por ejemplo, 70%, 80%, 90%, etc.)

- Usar la regla de codo

- Otros (e.g. pruebas de hipótesis)

# Ejemplo 2: Calificaciones

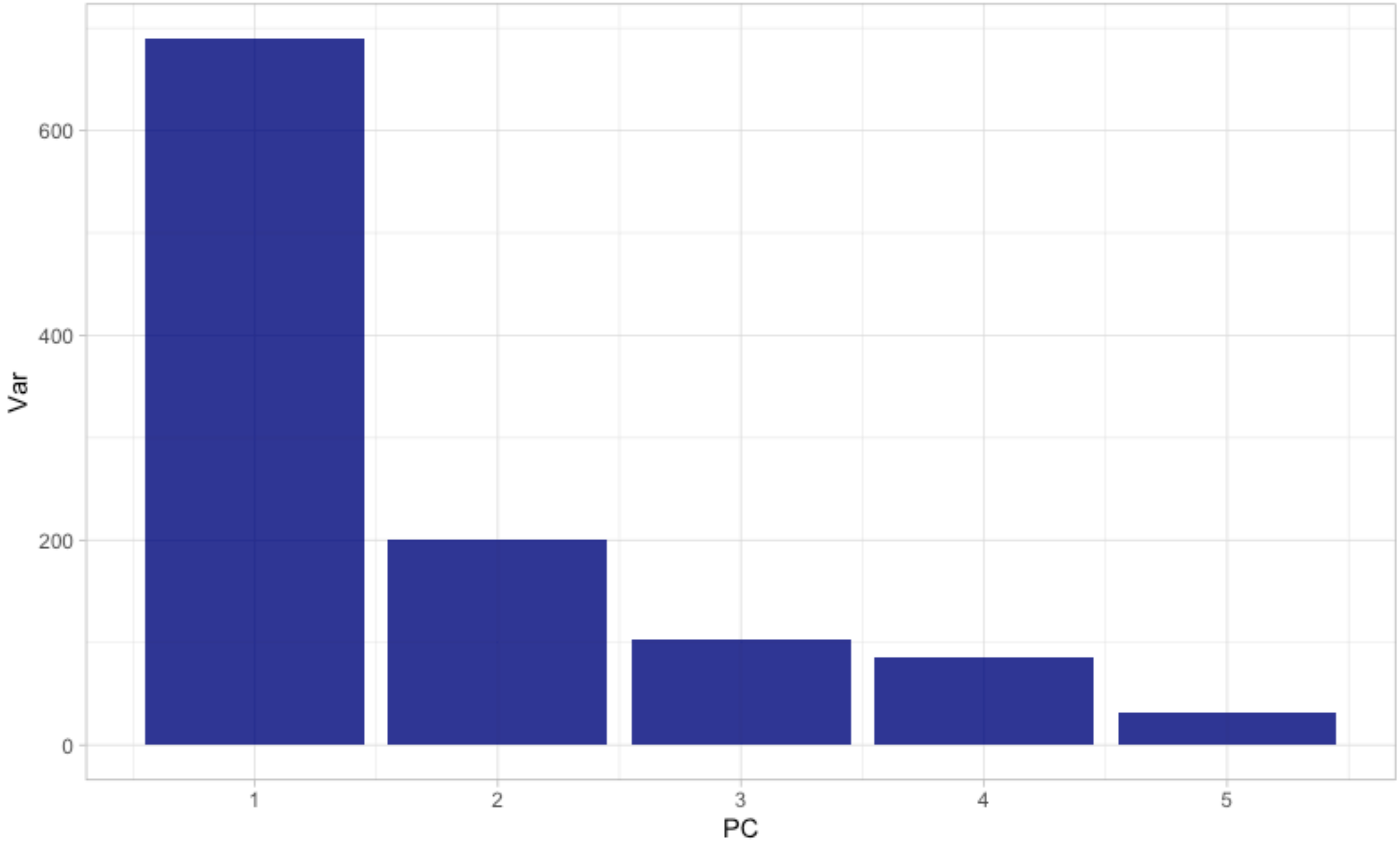
- La variación explicada por los componentes

  61.91%   18.21%   9.35%   7.63%   2.90%

- Nos quedamos con los primeros dos para tener arriba del 80% de la variación total (80.12%)

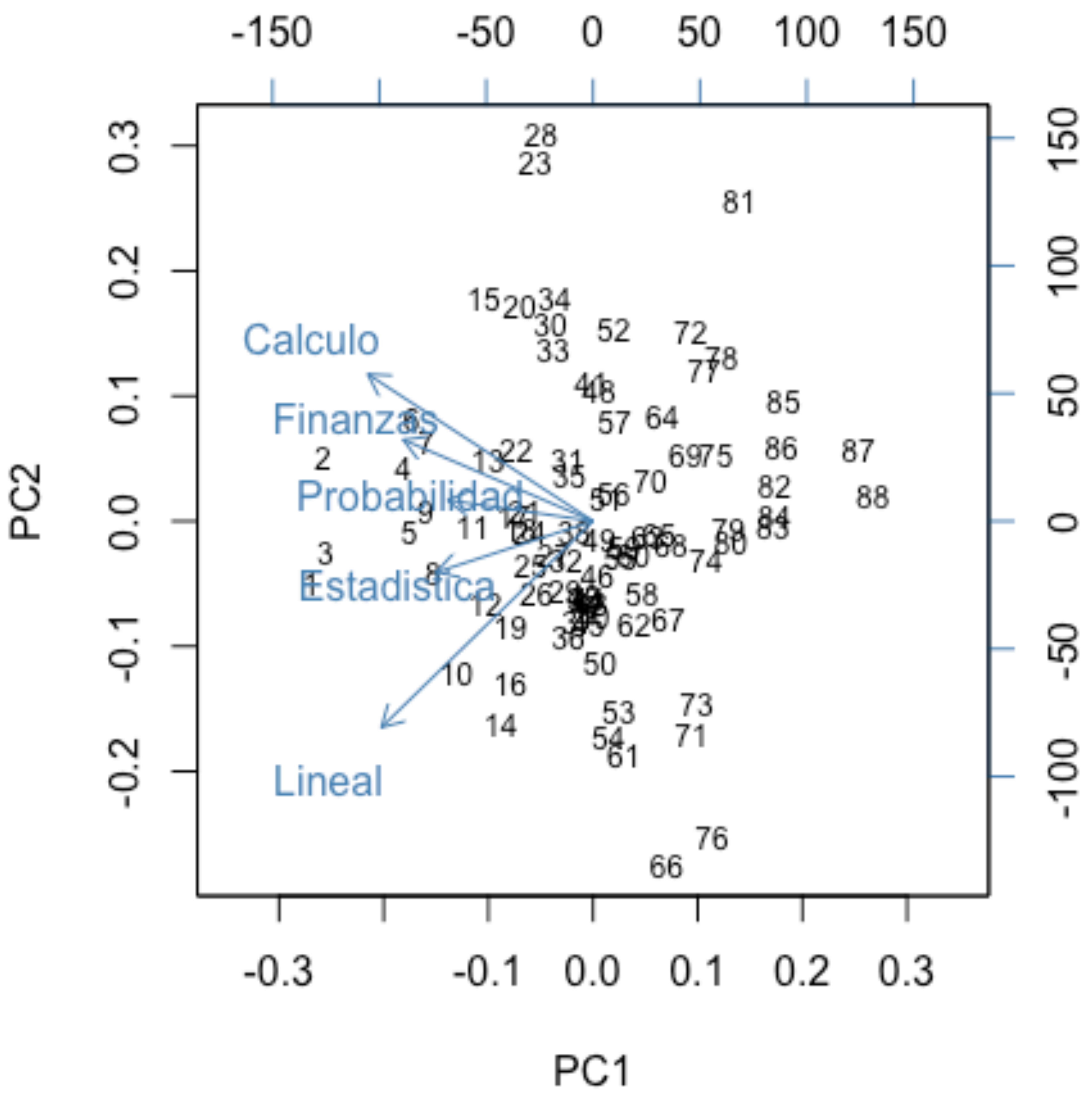
- Nos quedamos con los primeros tres para tener casi 90% de la variación total (89.47%)

- Regla de codo: graficar las varianzas (en **R** función screeplot)



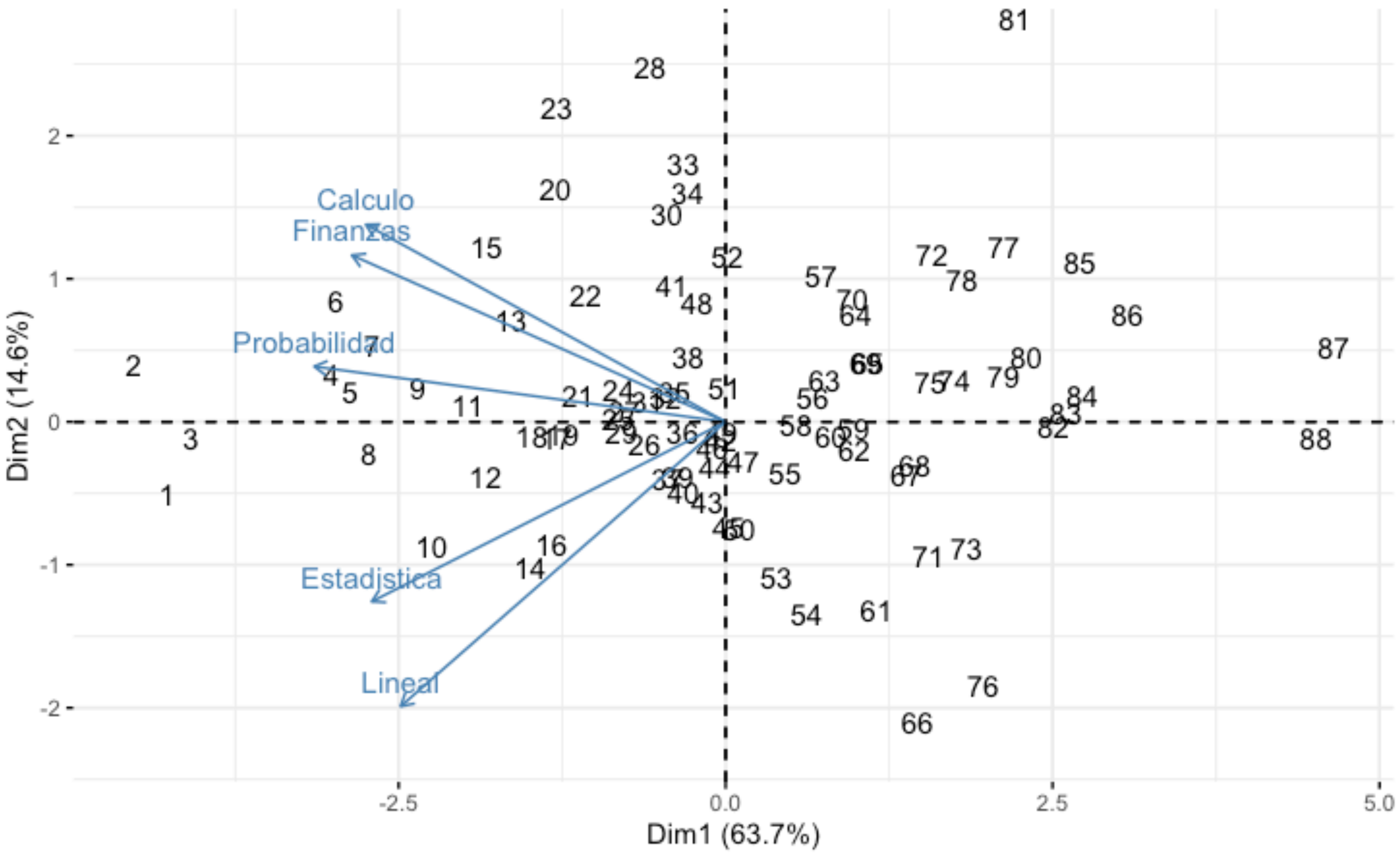
# Ejemplo 2: Calificaciones

- Si nos quedamos con dos componentes podemos graficarlos usando biplot()



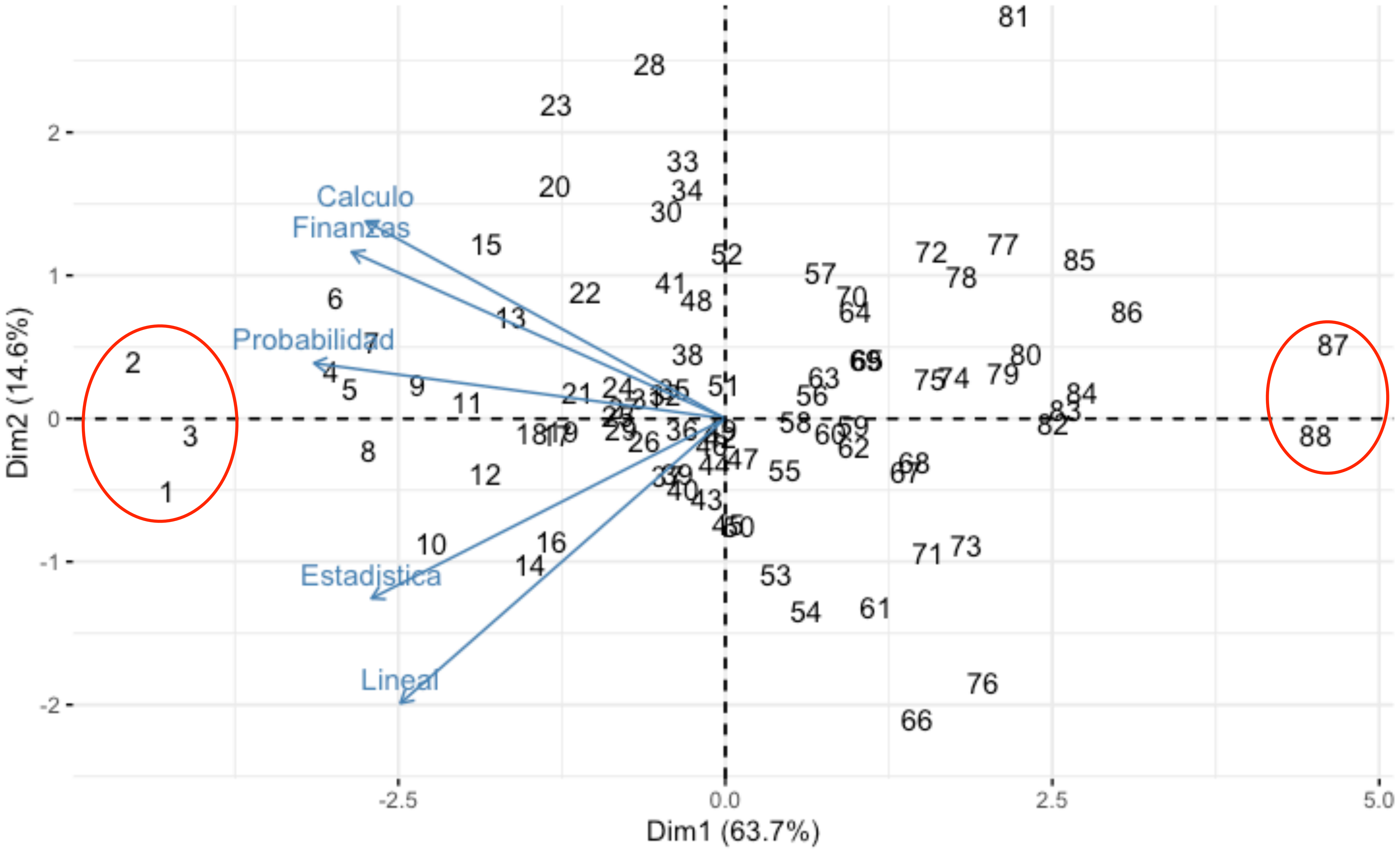
# Ejemplo 2: Calificaciones

- La librería factoextra proporciona una alternativa utilizando ggplot



# Ejemplo 2: Calificaciones

- Con el PC1, se pueden identificar los mejores y peores promedios

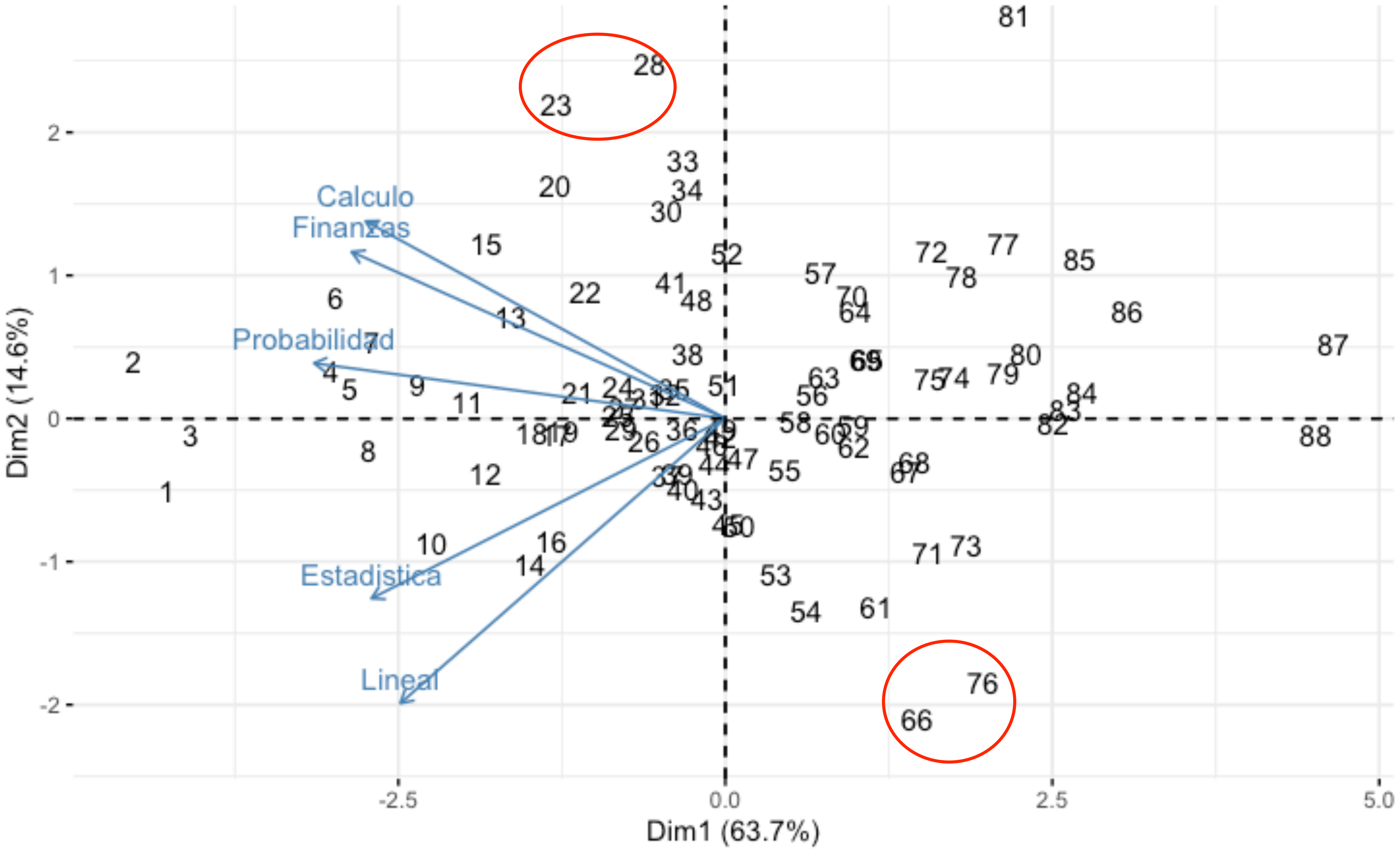



| Alumno | Lineal | Est. | Proba. | Finanzas | Cálculo |
|---|---|---|---|---|---|
| **1** | 97 | 92 | 77 | 72 | 96 |
| **2** | 83 | 88 | 90 | 75 | 96 |
| **3** | 95 | 83 | 81 | 71 | 96 |

| Alumno | Lineal | Est. | Proba. | Finanzas | Cálculo |
|---|---|---|---|---|---|
| **87** | 25 | 36 | 25 | 25 | 35 |
| **88** | 20 | 50 | 31 | 14 | 29 |

# Ejemplo 2: Calificaciones

- Con el PC2, se pueden identificar las mejores y peores calificaciones en examen abierto y cerrado




| Alumno | Lineal | Est. | Proba. | Finanzas | Cálculo |
|---|---|---|---|---|---|
| 66 | 79 | 63 | 47 | 27 | 34 |
| 76 | 69 | 60 | 48 | 28 | 24 |

| Alumno | Lineal | Est. | Proba. | Finanzas | Cálculo |
|---|---|---|---|---|---|
| 23 | 38 | 54 | 60 | 62 | 96 |
| 28 | 32 | 68 | 72 | 68 | 82 |

- Los eigenvalores resultantes con la matriz de correlación:

$$\lambda_1 = 1.7849 > \lambda_2 = 0.8536 > \lambda_3 = 0.6688 > \lambda_4 = 0.62582 > \lambda_5 = 0.4961$$

- Los vectores de cargas:

| | | | | | |
|---|---|---|---|---|---|
| **Lineal** | -0.397 | -0.664 | 0.612 | -0.091 | -0.131 |
| **Estadística** | -0.432 | -0.420 | -0.740 | 0.234 | -0.179 |
| **Probabilidad** | -0.502 | 0.129 | -0.021 | -0.116 | 0.846 |
| **Finanzas** | -0.456 | 0.389 | -0.064 | -0.674 | -0.425 |
| **Cálculo** | -0.439 | 0.461 | 0.268 | 0.684 | -0.230 |

# Número de componentes

- Seleccionar los componentes que expliquen un cierto porcentaje de la variación (por ejemplo, to%, 80%, 90%, etc.)

- Usar la regla de codo

- Otros (e.g. pruebas de hipótesis)

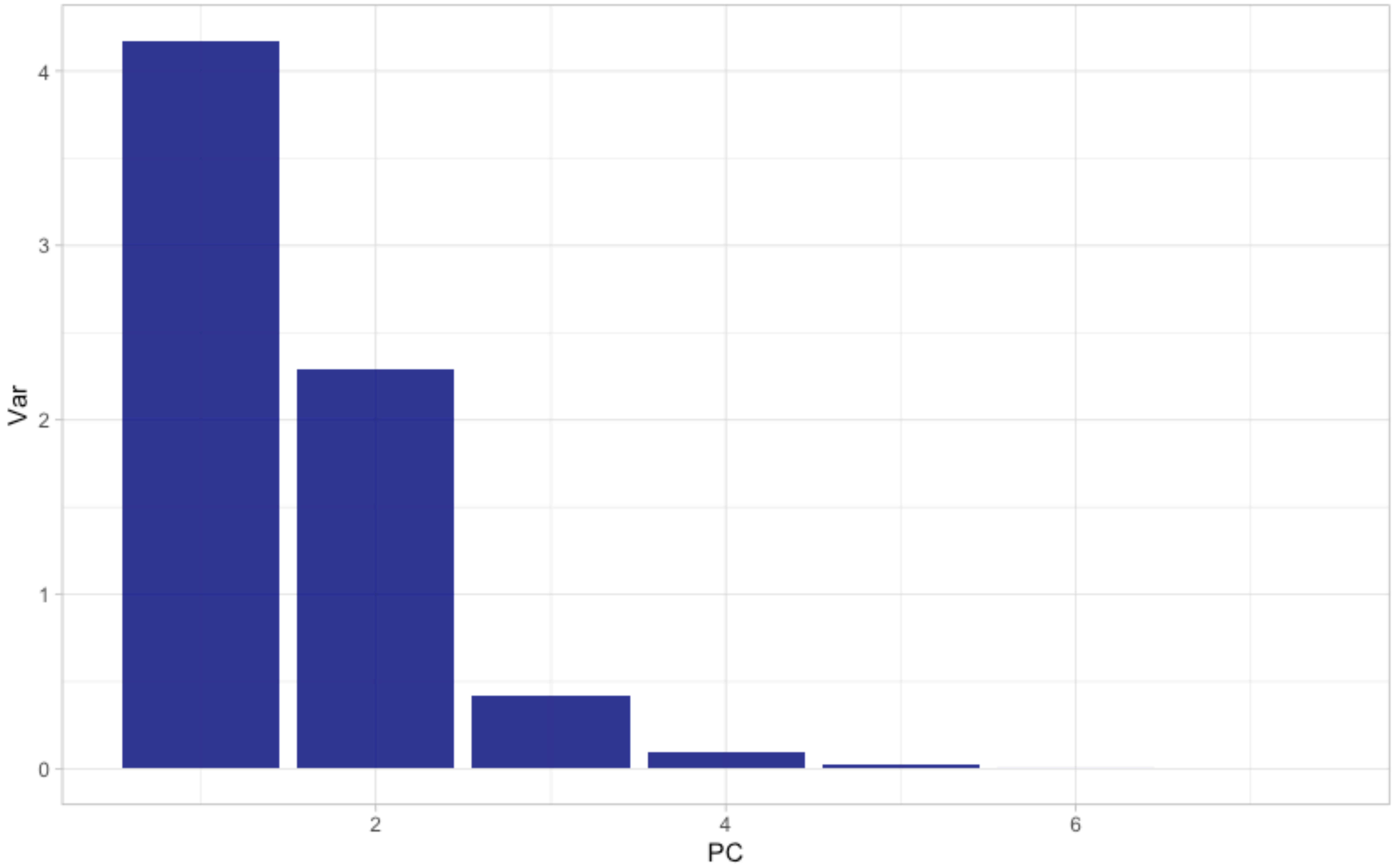
- **Regla de Kaiser.** Retener los componentes con varianza mayor a cierto valor (e.g. >.7)

# Ejemplo 3: Pizza

- 300 observaciones de 7 valores nutricionales en 10 marcas de pizza diferentes
  1. **Mois:** Cantidad de agua por cada 100g
  2. **Prot**: Cantidad de proteína por cada 100g
  3. **Fat**: Cantidad de grasa por cada 100g
  4. **Ash**: Cantidad de ceniza por cada 100g
  5. **Sodium**: Cantidad de sodio por cada 100g
  6. **Carb**: Cantidad de carbohidratos por cada 100g
  7. **Cal**: Cantidad de calorías por cada 100g

- Obtenemos los componentes principales con matriz de correlación, `prcomp(...,scale=T)`

- ¿Cuántos componentes?



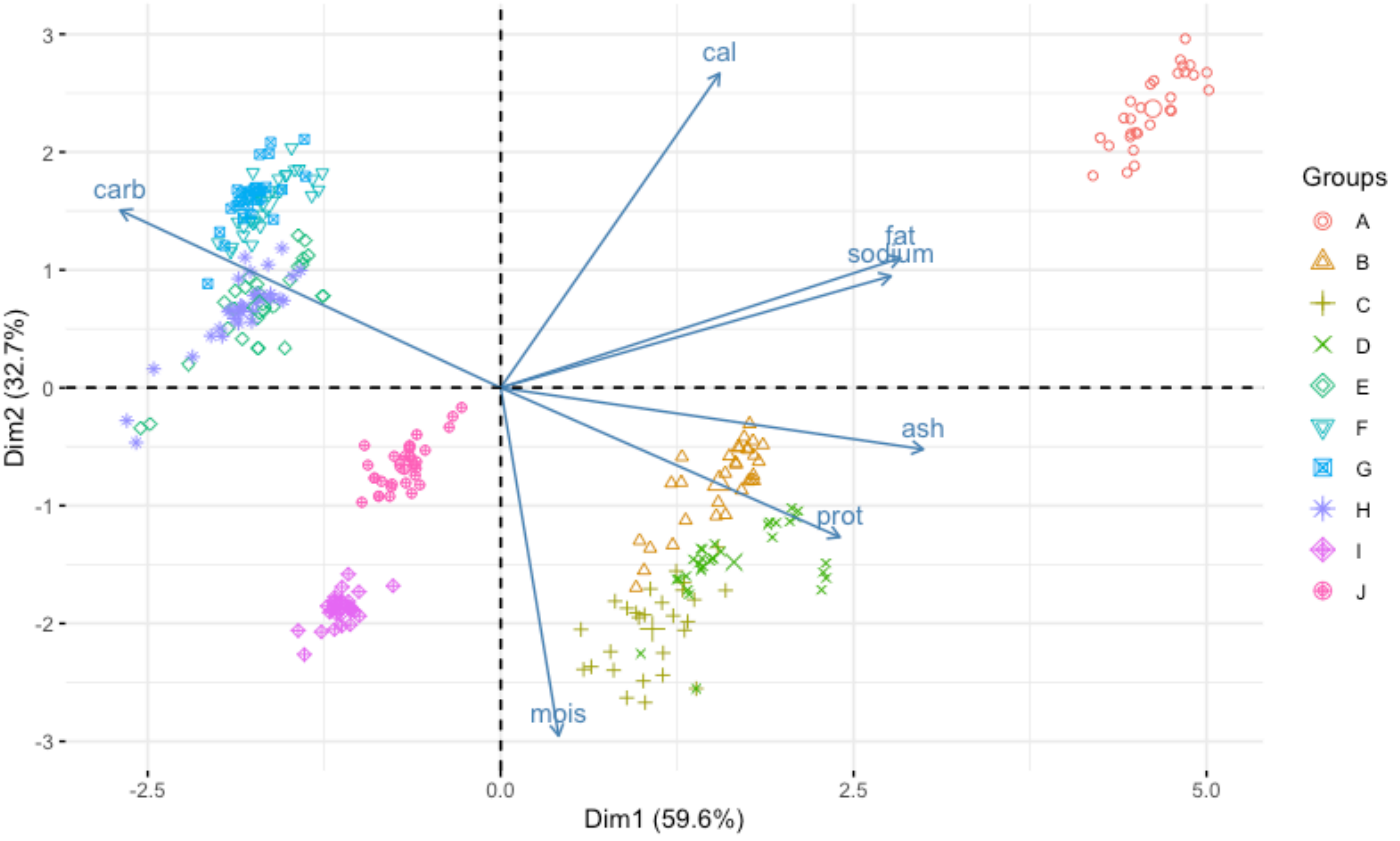
- El primer componente es

$$0.064 \cdot \text{mois} + 0.378 \cdot \text{prot} + 0.446 \cdot \text{fat} + 0.471 \cdot \text{ash} + 0.435 \cdot \text{sodium} - 0.424 \cdot \text{carb} + 0.244 \cdot \text{cal}$$

- El segundo componente es

$$-0.628 \cdot \text{mois} - 0.269 \cdot \text{prot} + 0.234 \cdot \text{fat} - 0.110 \cdot \text{ash} + 0.201 \cdot \text{sodium} + 0.320 \cdot \text{carb} + 0.567 \cdot \text{cal}$$

# Ejemplo 3: Pizza



# Otros temas

- Selección de variables (problema NP-difícil)

- PCA + otras modelos/técnicas multivariadas (e.g. SVM, análisis de discriminantes, regresión, etc.)

- Detección de outliers y observaciones influyentes (analizando los primeros y los últimos componentes)

- Rotación de componente principales (para una mejor interpretación como en análisis de factores)

- Otro tipo de datos (e.g. series de tiempo, datos no independientes, discretos, etc.)