

# **Análisis de Componentes Principales**

Álgebra aplicada

Domingo López Rodríguez

# Índice de Contenidos

|                                                           |          |
|-----------------------------------------------------------|----------|
| <b>Introducción: El Problema de la Alta Dimensión</b>     | <b>3</b> |
| La Metáfora del Fotógrafo . . . . .                       | 4        |
| <b>Bloque I: Fundamentos Teóricos</b>                     | <b>5</b> |
| Maximización de la Varianza . . . . .                     | 5        |
| <b>Bloque II: Implementación con R</b>                    | <b>6</b> |
| Análisis Exploratorio y Correlación . . . . .             | 6        |
| Preprocesamiento: La importancia del escalado . . . . .   | 7        |
| Ejecución y Varianza Explicada . . . . .                  | 8        |
| Selección de Componentes . . . . .                        | 8        |
| Componentes Principales: Incorrelación Perfecta . . . . . | 9        |
| Interpretación Profunda: ¿Qué nos devuelve R? . . . . .   | 10       |
| El Biplot: La Visión de Conjunto . . . . .                | 11       |

## Introducción: El Problema de la Alta Dimensión

En la era del Big Data, a menudo nos enfrentamos a conjuntos de datos con decenas o cientos de variables: edad, altura, peso, ingresos, nivel educativo, presión arterial, colesterol... Visualizar o analizar simultáneamente tantas dimensiones es imposible para la mente humana. Si tenemos tan solo 25 variables, ¡existen  $\binom{25}{2} = 300$  posibles pares de correlaciones para estudiar!

El **Análisis de Componentes Principales** (PCA) es la técnica reina de reducción de dimensión. Consiste en buscar combinaciones lineales de las variables originales que representen lo mejor posible a la **variabilidad** presente en los datos. En este contexto, se asume un axioma fundamental: "variabilidad" equivale a "información".

### Biometrika y el origen del PCA



Figura 1. Karl Pearson (1857-1936)

Karl Pearson publicó el primer artículo sobre PCA en **1901** en la revista *Biometrika*. Su motivación no era la computación, sino encontrar una forma de ajustar una "línea de mejor ajuste" que fuera independiente de qué variable elegíamos como "X" o "Y". Imaginaba los datos como una nube de puntos en forma de elipsoide y buscaba los ejes principales de esa nube.

## La Metáfora del Fotógrafo

Una forma intuitiva de entender el PCA es imaginar a un fotógrafo intentando capturar una escultura tridimensional (3D) en una foto plana (2D). El fotógrafo busca el **mejor ángulo**: aquel que capture la mayor cantidad de detalles y sombras, es decir, la **máxima variabilidad** de la figura.

En términos de datos, "buscar el mejor ángulo" equivale a proyectar los puntos sobre un nuevo eje (una componente principal) de modo que los puntos proyectados estén lo más **dispersos** posible.

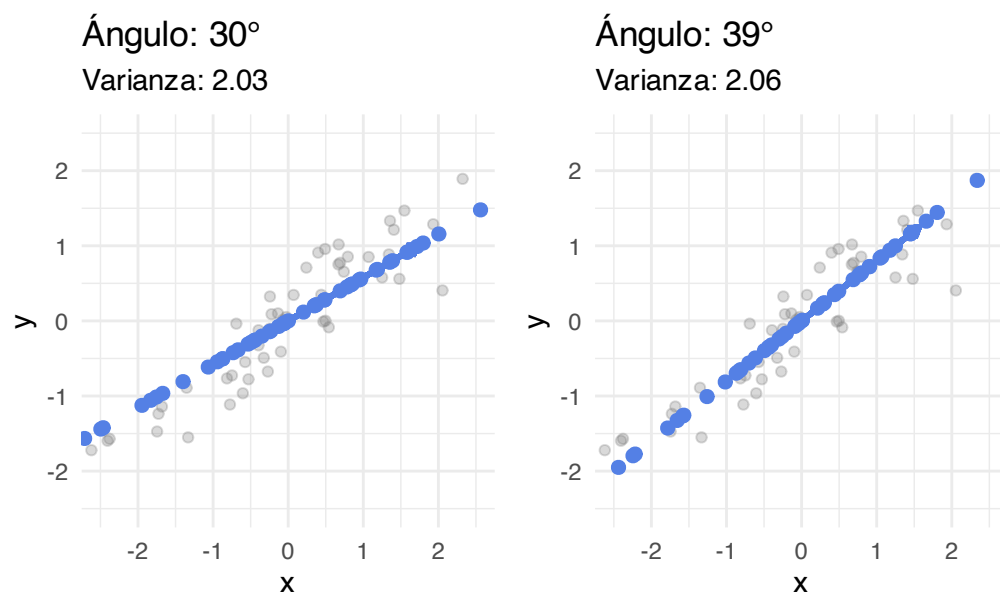


Figura 2. Concepto visual: Proyectar sobre la dirección de máxima varianza (PC1).

### El Factor 'g' de Inteligencia

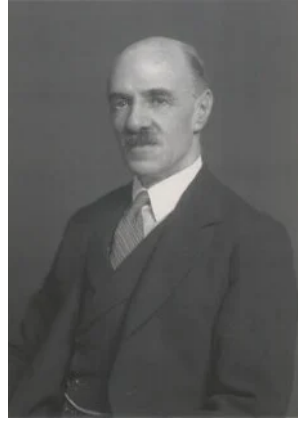


Figura 3. Charles Spearman (1863-1945)

Uno de los usos históricos más famosos del PCA (y su pariente cercano, el Análisis Factorial) fue en psicología. Charles Spearman observó que las puntuaciones en diferentes tests cognitivos estaban correlacionadas. Al aplicar estas técnicas, propuso la existencia de un **factor general de inteligencia ( $g$ )**, que sería la primera componente principal que explica la mayor parte de la habilidad cognitiva.

## Bloque I: Fundamentos Teóricos

### Maximización de la Varianza

Consideremos una matriz de datos  $X$  de tamaño  $n \times p$  ya **centrada** (media cero por columna). Buscamos una dirección, definida por un vector unitario  $\mathbf{u}$  (donde  $\|\mathbf{u}\| = 1$ ), tal que la varianza de las proyecciones de los datos sobre  $\mathbf{u}$  sea máxima.

La varianza proyectada se expresa como:

$$\text{Var}(X\mathbf{u}) = \frac{1}{n-1}(X\mathbf{u})^T(X\mathbf{u}) = \mathbf{u}^T \Sigma \mathbf{u}$$

donde  $\Sigma$  es la matriz de covarianza de los datos.

**Ejemplo 0.1** (Teorema Fundamental del PCA). Las direcciones de máxima variancia son los **autovectores** de la matriz de covarianza  $\Sigma$ . La variancia explicada en cada dirección es el **autovalor** correspondiente.

**Demostración (vía Multiplicadores de Lagrange):**

Queremos maximizar la función  $f(\mathbf{u}) = \mathbf{u}^T \Sigma \mathbf{u}$  bajo la restricción de que  $g(\mathbf{u}) = \mathbf{u}^T \mathbf{u} =$

1. Definimos el Lagrangiano como:

$$\mathcal{L}(\mathbf{u}, \lambda) = \mathbf{u}^T \Sigma \mathbf{u} - \lambda(\mathbf{u}^T \mathbf{u} - 1)$$

Si derivamos respecto al vector  $\mathbf{u}$  e igualamos el resultado a cero:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{u}} = 2\Sigma \mathbf{u} - 2\lambda \mathbf{u} = 0 \quad \implies \quad \Sigma \mathbf{u} = \lambda \mathbf{u}$$

Esto demuestra que el vector  $\mathbf{u}$  debe ser necesariamente un **autovector** de la matriz  $\Sigma$ .

Sustituyendo de nuevo en nuestra función de variancia:

$$\mathbf{u}^T \Sigma \mathbf{u} = \mathbf{u}^T (\lambda \mathbf{u}) = \lambda (\mathbf{u}^T \mathbf{u}) = \lambda$$

Por lo tanto, para maximizar la variabilidad explicada, debemos elegir el autovector asociado al **autovalor más grande**.  $\square$

## Bloque II: Implementación con R

Analizaremos el dataset `USArrests`, que contiene las tasas de criminalidad en 50 estados de EEUU. en 1973.

### Análisis Exploratorio y Correlación

Antes de reducir la dimensión, comprobamos si las variables están relacionadas.

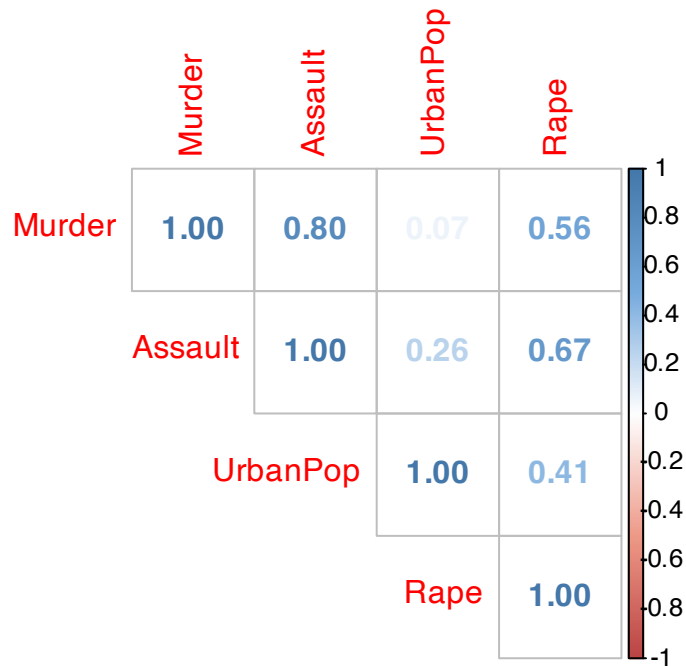
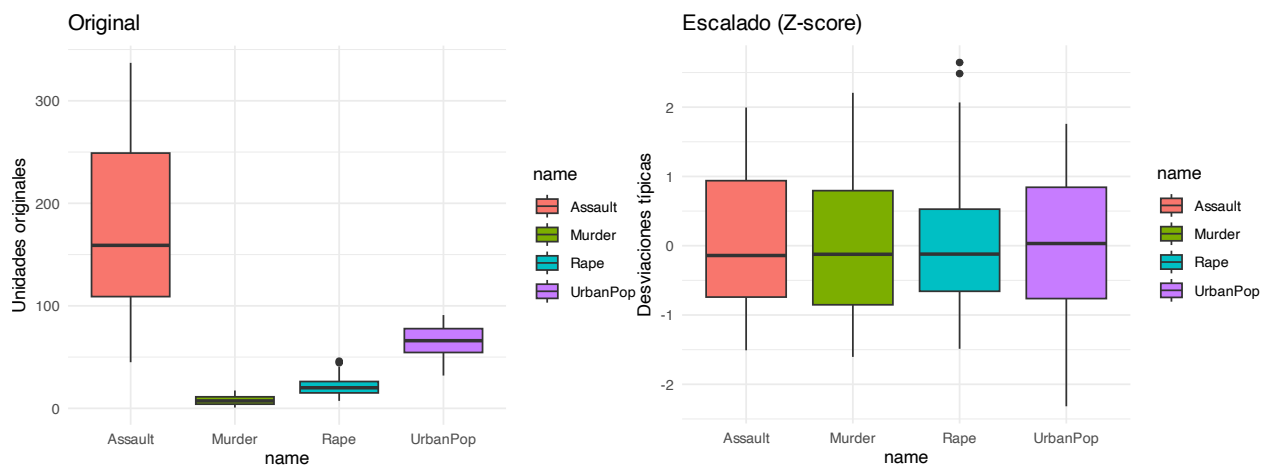


Figura 4. Matriz de correlación de las variables originales (USArrests).

## Preprocesamiento: La importancia del escalado

PCA es extremadamente sensible a las unidades de medida. En USArrests, la variable Assault tiene una magnitud y varianza muchísimo mayor que el resto. Sin estandarizar, PCA ignoraría la estructura de las otras dimensiones. Por ello, es imperativo escalar los datos (media 0, varianza 1).



## Ejecución y Varianza Explicada

```
# Ejecutamos PCA con escalado  
acp <- prcomp(USArrests, center = TRUE, scale = TRUE)  
summary(acp)
```

Importance of components:

|                        | PC1    | PC2    | PC3     | PC4     |
|------------------------|--------|--------|---------|---------|
| Standard deviation     | 1.5749 | 0.9949 | 0.59713 | 0.41645 |
| Proportion of Variance | 0.6201 | 0.2474 | 0.08914 | 0.04336 |
| Cumulative Proportion  | 0.6201 | 0.8675 | 0.95664 | 1.00000 |

## Selección de Componentes

Visualizamos el Scree Plot y la contribución de cada variable.

Warning in geom\_bar(stat = "identity", fill = barfill, color = barcolor, :  
Ignoring empty aesthetic: `width`.

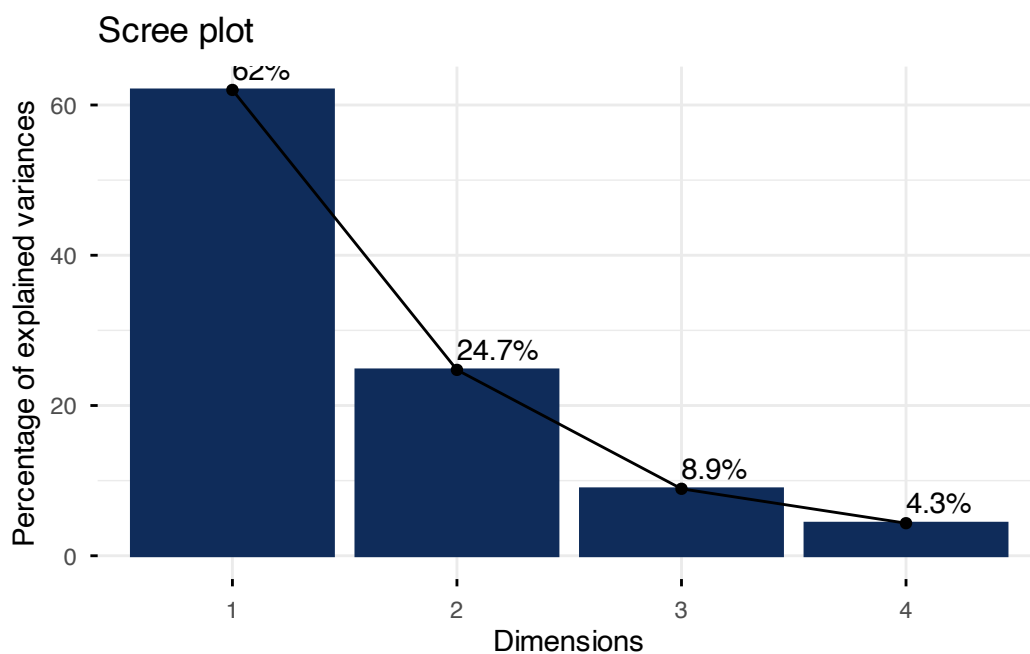


Figura 5. Scree Plot: Porcentaje de varianza explicada.

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.

i Please use ``linewidth`` instead.

i The deprecated feature was likely used in the `ggpubr` package.

Please report the issue at <https://github.com/kassambara/ggpubr/issues>.

Warning: ``aes_string()`` was deprecated in `ggplot2 3.0.0`.

i Please use tidy evaluation idioms with ``aes()``.

i See also ``vignette("ggplot2-in-packages")`` for more information.

i The deprecated feature was likely used in the `factoextra` package.

Please report the issue at <https://github.com/kassambara/factoextra/issues>.

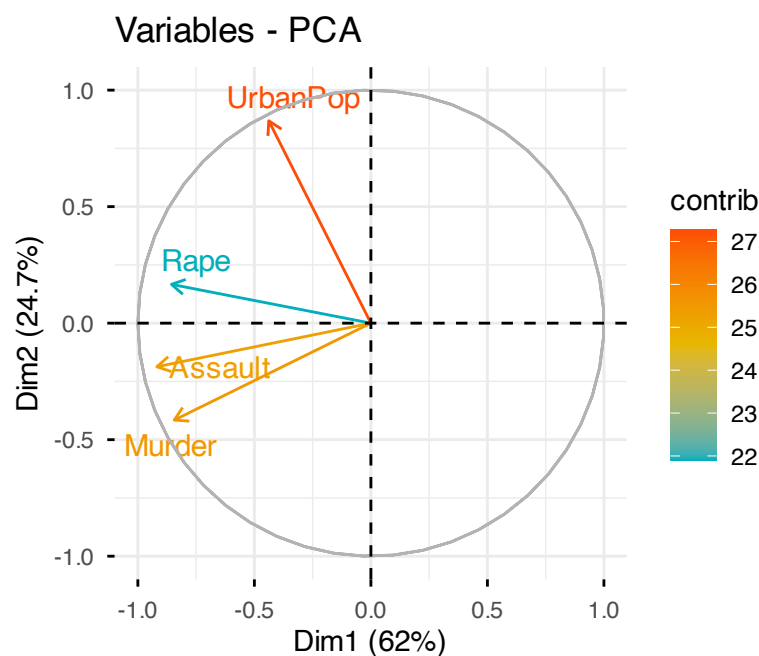


Figura 6. Contribuciones de las variables a las dos primeras componentes.

Las dos primeras componentes explican el **86.7%** de la información total. La PC1 agrupa los delitos violentos, mientras que la PC2 captura principalmente el nivel de urbanización.

## Componentes Principales: Incorrelación Perfecta

Uno de los mayores beneficios del PCA es que las nuevas variables son **ortogonales** (incorreladas) entre sí.

Warning in `ind1:ind2`: numerical expression has 2 elements: only the first used

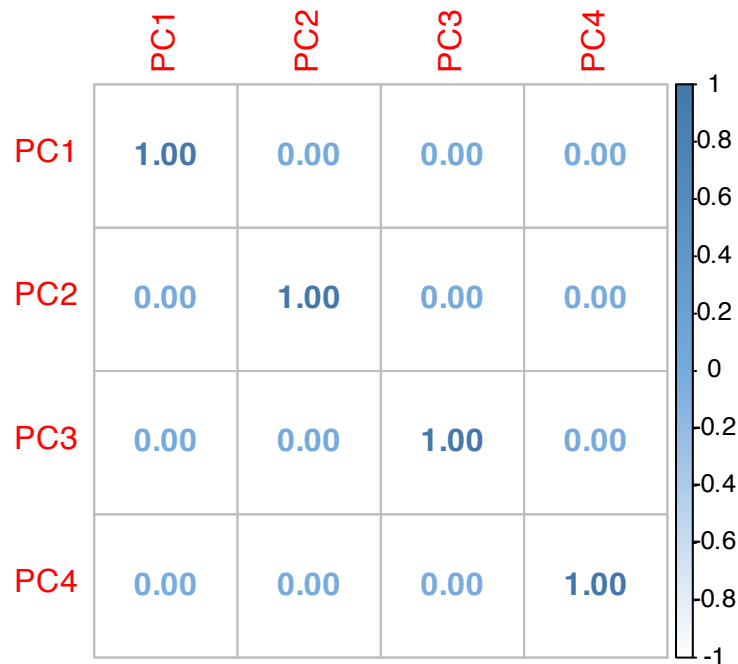


Figura 7. Matriz de correlación entre las Componentes Principales.

## Interpretación Profunda: ¿Qué nos devuelve R?

El objeto `acp` resultante de `prcomp` contiene:

- **sdev**: Desviaciones típicas (raíces de los autovalores). La varianza es  $\text{sdev}^2$ .
- **rotation**: Matriz de Loadings (autovectores). Pesos de las variables originales.
- **x**: Los Scores. Coordenadas de los individuos en el nuevo espacio.

**Ejemplo 0.2** (Desglose de un caso: California). Cerca de un tercio de la variabilidad se resume en un solo score.

```
# Valores escalados de California (fila 5 de USArrests)
calif_scaled <- scale(USArrests)[5, ]

# Pesos de la PC1
weights_pc1 <- acp$rotation[, 1]

# Producto escalar (manual)
score_manual <- sum(calif_scaled * weights_pc1)

score_manual
```

[1] -2.498613

California tiene un score en PC1 de -2.5. Un valor muy negativo indica que ese estado tiene niveles de criminalidad muy por encima de la media.

## El Biplot: La Visión de Conjunto

Visualización combinada de variables e individuos.

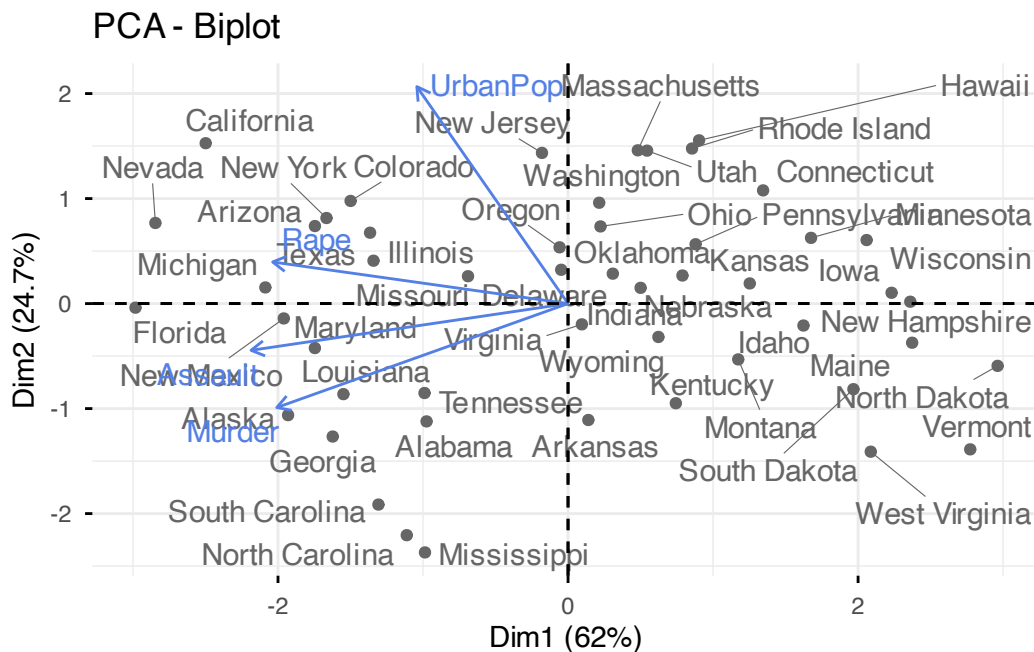


Figura 8. Biplot de Arrestos en EEUU.

### Guía de interpretación de los estados:

- **California e Illinois:** Situados en el cuadrante superior izquierdo, indicando altos niveles de delitos y alta urbanización.
- **Carolina del Norte:** Situado en el cuadrante inferior izquierdo, reflejando criminalidad alta pero baja urbanización.
- **Vectores:** La cercanía entre las flechas de *Murder* y *Assault* destaca su alta correlación.