Análise de Componentes Principais (ACP)

Baseado nos slides (AMA/LEAp 2025/26)

```
library(dplyr)
library(ggplot2)
```

1. Introdução

- A ACP é uma técnica **exploratória** para "simplificar" dados multivariados mantendo a **informação mais relevante**.
- Objetivo: reduzir a **dimensionalidade** (nº de variáveis) definindo r novas variáveis, $1 \leq r \leq p$, que expliquem o **máximo** da variabilidade das p originais.
- As novas variáveis (componentes principais) devem ser **não correlacionadas** entre si.
- **Reduzimos o nº de variáveis**, **não** o tamanho amostral.

---

1. Formulação

- Dadas p v.a. quantitativas correlacionadas $X_1, \ldots, X_p$, definem-se novas v.a. $Y_1, \ldots, Y_p$, cada uma **combinação linear** das originais, tais que:
  - $\text{Var}[Y_1] \geq \text{Var}[Y_2] \geq \cdots \geq \text{Var}[Y_p]$;
  - $\text{Cov}(Y_j, Y_k) = 0$ para $j \neq k$;
  - $\|a_j\| = 1$ ("norma 1" nos pesos) para fixar a escala das combinações lineares.
- Espera-se que as primeiras r (poucas) CPs expliquem **80–90%** da variabilidade total, permitindo descartar as restantes em análises subsequentes.

---

2. Construção das Componentes Principais — 1ª CP

- Procuramos $Y_1 = a_{11}X_1 + \cdots + a_{1p}X_p$ que **maximiza a variância** entre todas as combinações de **norma 1**:

$$\max_{\|a_1\|=1} \mathrm{Var}(Y_1) \quad \text{com} \quad a_1 = (a_{11}, \ldots, a_{1p})^\top.$$

- A restrição $\|a_1\| = 1$ é **essencial**: sem ela, a variância pode ser inflacionada reescalando os pesos.
- Solução clássica (multiplicadores de Lagrange): a_1 é o **vetor próprio (norma 1)** associado ao **maior valor próprio** da matriz de **covariâncias amostrais** das X_j.

---

2. Construção — 2ª CP e seguintes

- 2ª CP: $Y_2 = a_{21}X_1 + \cdots + a_{2p}X_p$ que maximiza a variância **sujeita a**

  1) $\|a_2\| = 1$; 2) **ortogonalidade** a Y_1 (i.e., $a_2^\top a_1 = 0$).

- 3ª CP: análoga, com $\|a_3\| = 1$ e ortogonalidade a Y_1 e Y_2.
- Em geral, obtemos p vetores próprios $a_1, \ldots, a_p$ (ordenados por valores próprios **estritamente decrescentes** $\lambda_1 > \cdots > \lambda_p \geq 0$) — assumindo covariâncias com espectro simples.

---

2. Variâncias e variância total

- Cada valor próprio **representa a variância** da respetiva CP: $\lambda_j = \mathrm{Var}[Y_j]$.
- A **variância total** das variáveis iniciais é a soma dos valores próprios: $\sum_{j=1}^p \mathrm{Var}[X_j] = \sum_{j=1}^p \lambda_j$.
- **Proporção explicada pela** j-ésima CP: $\lambda_j / \sum_{i=1}^p \lambda_i$.
- **Proporção de variância explicada acumulada** pelas primeiras r CPs: $(\lambda_1 + \cdots + \lambda_r)/\sum_{i=1}^p \lambda_i$.

---

2. Matriz de correlações

- Se as X_j estão em **unidades diferentes** (variâncias muito distintas), as CPs ficam **dominadas** pelas variáveis de maior variância.
- Solução: **estandardizar** as variáveis ou obter as CPs a partir da **matriz de correlações**.
- Numa matriz de correlações, o traço (soma dos diagonais) é p; logo, a proporção explicada pela j-ésima CP é λ_j/p.

---

3. Exercícios

(A) Caso bivariado: matriz de correlação

$$P = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}, \quad r = \rho(X_1, X_2) \neq 0.$$

1. Determine as **duas CPs** (valores próprios/vetores próprios de P).

2. Discuta, em função de r, a **importância** relativa das CPs.

(B) Caso trivariado:

$$P = \begin{bmatrix} 1 & r & 0 \\ r & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad r \neq 0.$$

1. Determine as **três CPs**.

2. Discuta a importância em função de r.

3. Compare com (A) e comente o efeito de X_3 ser **não correlacionada** com X_1 e X_2.

---

4. Observações importantes sobre a ACP (I)

I) Objetivo principal: transformar p v.a. **correlacionadas** num novo conjunto de p v.a. **não correlacionadas**, mantendo a **variabilidade total**. A redução de dimensão (reter poucas CPs) **só** é viável quando as originais estão correlacionadas; caso contrário, a ACP apenas **ordena** por variâncias.

II) As CPs são obtidas por **ordem decrescente de importância** (variância).

III) A ACP é adequada quando **todas** as variáveis estão em "pé de igualdade" (não há variáveis resposta vs. explicativas, ao contrário de MANOVA/AD).

IV) **Não** aplicar ACP a **variáveis qualitativas** (mesmo com codificação numérica arbitrária).

---

4. Observações importantes sobre a ACP (II)

V) Em R, ACP com `prcomp`:

```
# Se pretende CPs da matriz de correlações (recomendado com escalas distintas):
pc <- prcomp(X, center = TRUE, scale. = TRUE)
# CPs da matriz de covariâncias (todas as variáveis na mesma escala/unidade):
pc_cov <- prcomp(X, center = TRUE, scale. = FALSE)
summary(pc)     # variâncias (valores próprios), PVE, PVE acumulada
biplot(pc)      # biplot (scores e loadings)
```

VI) Quantas CPs reter?

- Meta prática: explicar **80–90%** da variabilidade com poucas CPs.

---

5. Exemplo/Exercício — Dados *iris*

Considere *iris* sem a coluna `Species`.

1) Obtenha a matriz de **covariâncias** S; calcule **valores próprios** e **vetores próprios (norma 1)** de S.

2) Verifique que:

- Valores próprios de S = **variâncias** das CPs;
- Coeficientes das CPs coincidem (a menos de sinal) com os **vetores próprios** de S;
- **Traço** de S = soma dos valores próprios = soma das variâncias das CPs.

---

3) Identifique as CPs e a **proporção de variância** explicada por cada uma; **interprete** as CPs.

4) Calcule a **matriz de correlação** dos **scores** e faça o **diagrama de dispersão** das duas primeiras CPs; comente a separação entre espécies (mesmo não usadas no ajuste).

```
#1.
#Matriz
iris_data <- iris |> select(-Species)

S <- cov(iris_data)
print("--- Matriz de Covariâncias (S) ---")
```

```
[1] "--- Matriz de Covariâncias (S) ---"
```

```
print(S)
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

```r
#Valores e vetores próprios
eig <- eigen(S)
valores_proprios_S <- eig$values
vetores_proprios_S <- eig$vectors
print("---------------")
```

```
[1] "---------------"
```

```r
print("--- Valores Próprios (de S) ---")
```

```
[1] "--- Valores Próprios (de S) ---"
```

```r
print(valores_proprios_S)
```

```
[1] 4.22824171 0.24267075 0.07820950 0.02383509
```

```r
print("---------------")
```

```
[1] "---------------"
```

```r
print("--- Vetores Próprios (de S, em colunas) ---")
```

```
[1] "--- Vetores Próprios (de S, em colunas) ---"
```

```r
print(vetores_proprios_S)
```

```
            [,1]        [,2]        [,3]       [,4]
[1,]  0.36138659 -0.65658877  0.58202985  0.3154872
[2,] -0.08452251 -0.73016143 -0.59791083 -0.3197231
[3,]  0.85667061  0.17337266 -0.07623608 -0.4798390
[4,]  0.35828920  0.07548102 -0.54583143  0.7536574
```

```r
# Ajustar a PCA
# Usamos scale. = FALSE para usar a matriz de covariâncias
pca_res <- prcomp(iris_data, scale. = FALSE)
# As variâncias das CPs são o desvio-padrão (sdev) ao quadrado
variancias_cp <- pca_res$sdev^2
print("--- Verificação 1: Valores Próprios vs Variâncias CPs ---")
```

```
[1] "--- Verificação 1: Valores Próprios vs Variâncias CPs ---"
```

```
print("Valores Próprios (de S):")
```

```
[1] "Valores Próprios (de S):"
```

```
print(valores_proprios_S)
```

```
[1] 4.22824171 0.24267075 0.07820950 0.02383509
```

```
print("Variâncias das CPs:")
```

```
[1] "Variâncias das CPs:"
```

```
print(variancias_cp)
```

```
[1] 4.22824171 0.24267075 0.07820950 0.02383509
```

```
print("----------------")
```

```
[1] "----------------"
```

```
print("----------------")
```

```
[1] "----------------"
```

```
print("----------------")
```

```
[1] "----------------"
```

```
coeficientes_cp <- pca_res$rotation
print("--- Verificação 2: Vetores Próprios vs Coeficientes CPs ---")
```

```
[1] "--- Verificação 2: Vetores Próprios vs Coeficientes CPs ---"
```

```r
print("Vetores Próprios (de S):")
```

```
[1] "Vetores Próprios (de S):"
```

```r
print(vetores_proprios_S)
```

```
            [,1]        [,2]        [,3]       [,4]
[1,]  0.36138659 -0.65658877  0.58202985  0.3154872
[2,] -0.08452251 -0.73016143 -0.59791083 -0.3197231
[3,]  0.85667061  0.17337266 -0.07623608 -0.4798390
[4,]  0.35828920  0.07548102 -0.54583143  0.7536574
```

```r
print("Coeficientes CPs (Loadings):")
```

```
[1] "Coeficientes CPs (Loadings):"
```

```r
print(coeficientes_cp)
```

```
                    PC1         PC2         PC3        PC4
Sepal.Length  0.36138659 -0.65658877  0.58202985  0.3154872
Sepal.Width  -0.08452251 -0.73016143 -0.59791083 -0.3197231
Petal.Length  0.85667061  0.17337266 -0.07623608 -0.4798390
Petal.Width   0.35828920  0.07548102 -0.54583143  0.7536574
```

```r
traco_S <- sum(diag(S))
soma_val_prop <- sum(valores_proprios_S)
soma_var_cp <- sum(variancias_cp)

print("--- Verificação 3: Traço(S) vs Somas ---")
```

```
[1] "--- Verificação 3: Traço(S) vs Somas ---"
```

```r
print(paste("Traço de S:", traco_S))
```

```
[1] "Traço de S: 4.57295704697987"
```

```r
print(paste("Soma dos Valores Próprios:", soma_val_prop))
```

```
[1] "Soma dos Valores Próprios: 4.57295704697986"
```

```r
print(paste("Soma das Variâncias das CPs:", soma_var_cp))
```

```
[1] "Soma das Variâncias das CPs: 4.57295704697987"
```

```r
scores <- pca_res$x
cor_scores <- cor(scores)
print("--- Matriz de Correlação dos Scores das CPs ---")
```

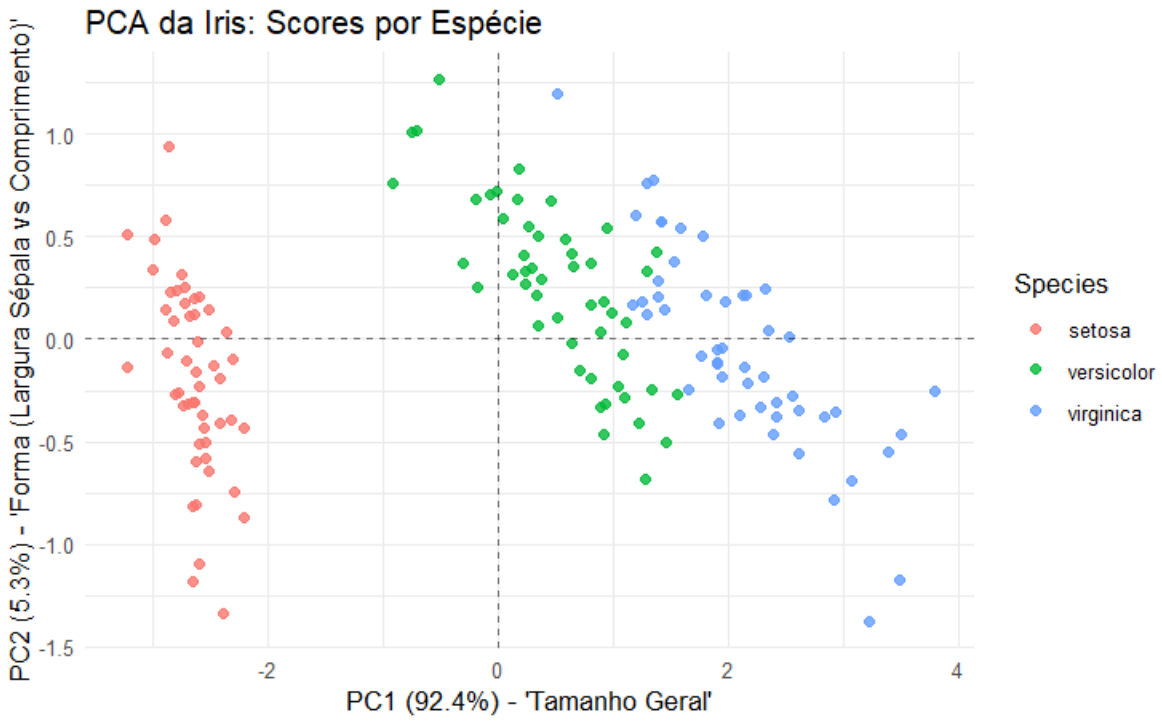
```
[1] "--- Matriz de Correlação dos Scores das CPs ---"
```

```r
print(round(cor_scores, 5)) # Arredondar para ver os zeros
```

```
    PC1 PC2 PC3 PC4
PC1   1   0   0   0
PC2   0   1   0   0
PC3   0   0   1   0
PC4   0   0   0   1
```

```r
plot_data <- bind_cols(as_tibble(scores), Species = iris$Species)

# Gráfico
ggplot(plot_data, aes(x = PC1, y = PC2, color = Species)) +
  geom_point(alpha = 0.8, size = 2) +
  labs(
    title = "PCA da Iris: Scores por Espécie",
    x = "PC1 (92.4%) - 'Tamanho Geral'",
    y = "PC2 (5.3%) - 'Forma (Largura Sépala vs Comprimento)'"
  ) +
  theme_minimal() +
  # Adiciona linhas a 0 para referência
  geom_hline(yintercept = 0, linetype = "dashed", alpha = 0.5) +
  geom_vline(xintercept = 0, linetype = "dashed", alpha = 0.5)
```


PCA da Iris: Scores por Espécie
PC2 (5.3%) - 'Forma (Largura Sépala vs Comprimento)'
PC1 (92.4%) - 'Tamanho Geral'
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2
0
2
4
Species
setosa
versicolor
virginica