

Análise de componentes principais

Prof. Caio Azevedo

Introdução

- Objetivo: construir variáveis não correlacionadas que retenham a maior parte da estrutura de variabilidade e correlação, a partir da variáveis originais, através de transformações lineares.
- Benefícios:
 - Pode-se usar metodologias de análise univariada.
 - Pode-se trabalhar com um menor número de variáveis.
 - Pode-se obter detalhes do comportamento dos dados os quais são difíceis de serem deduzidos a partir das variáveis originais.

Estrutura estatística

- $\mathbf{X} = (X_1, \dots, X_p)' \sim D_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, que $D_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ representa uma distribuição p-variada (em princípio qualquer) com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$.
- A matriz de correlações será denotada, como antes, por ρ .
- Objetivo, encontrar $\mathbf{A}_{(p \times p)}$ (não estocástica), a fim de obter $\mathbf{Y} = \mathbf{AX} = (Y_1, \dots, Y_p)'$, de sorte que $\mathcal{V}(Y_1) \geq \mathcal{V}(Y_2) \geq \dots \mathcal{V}(Y_p)$ e que $\text{Corre}(Y_i, Y_j) = 0 \forall i \neq j, i, j = 1, \dots, p$, e que $\mathcal{V}(Y_i)$ seja máximo $i = 1, 2, \dots, p$.

Cont.

$$\mathbf{Y} = \mathbf{A}\mathbf{X} = \begin{bmatrix} \mathbf{a}'_1 \\ \mathbf{a}'_2 \\ \vdots \\ \mathbf{a}'_p \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

- Portanto $\mathcal{V}(Y_i) = \mathbf{a}'_i \boldsymbol{\Sigma} \mathbf{a}_i$ e $\text{Cov}(Y_i, Y_j) = \mathbf{a}'_i \boldsymbol{\Sigma} \mathbf{a}_j$.
- Claramente $\mathcal{V}(Y_i)$ pode ser aumentada de modo ilimitado multiplicando-se cada vetor \mathbf{a}_i por uma constante positiva (o que não afetaria o fato de que $\mathbf{a}'_i \boldsymbol{\Sigma} \mathbf{a}_i = 0$). Assim, iremos restringir nossa atenção a vetores com norma unitária ($\mathbf{a}'_i \mathbf{a}_i = 1$).

Procedimento

- Primeira componente principal: combinação linear $\mathbf{a}'_1\mathbf{X}$ que maximiza $\mathcal{V}(\mathbf{a}'_1\mathbf{X})$ sujeito à $\mathbf{a}'_1\mathbf{a}_1 = 1$.
- Segunda componente principal: combinação linear $\mathbf{a}'_2\mathbf{X}$ que maximiza $\mathcal{V}(\mathbf{a}'_2\mathbf{X})$ sujeito à $\mathbf{a}'_2\mathbf{a}_2 = 1$ e $Cov(\mathbf{a}'_1\mathbf{X}, \mathbf{a}'_2\mathbf{X}) = 0$.
- i -ésima componente principal: combinação linear $\mathbf{a}'_i\mathbf{X}$ que maximiza $\mathcal{V}(\mathbf{a}'_i\mathbf{X})$ sujeito à $\mathbf{a}'_i\mathbf{a}_i = 1$ e $Cov(\mathbf{a}'_i\mathbf{X}, \mathbf{a}'_k\mathbf{X}) = 0$, para $k < i$.

Resultado

- Sejam $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \dots, (\lambda_p, \mathbf{e}_p)$, em que $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ são os autovalores associados à Σ , e \mathbf{e}_i os respectivos auto-vetores ortonormalizados, ou seja $\mathbf{e}_i' \mathbf{e}_i = 1$ e $\mathbf{e}_i' \mathbf{e}_j = 0, \forall i \neq j, i, j = 1, 2, \dots, p$.
- A i -ésima componente principal é dada por

$$\mathbf{Y}_i = \mathbf{e}_i' \mathbf{X}, i = 1, 2, \dots, p. \quad (1)$$

- Assim, temos que

$$\mathcal{V}(Y_i) = \mathbf{e}_i' \Sigma \mathbf{e}_i = \lambda_i, i = 1, 2, \dots, p \quad (2)$$

$$\text{Cov}(Y_i, Y_j) = \mathbf{e}_i' \Sigma \mathbf{e}_j = 0, \forall i \neq j, i, j = 1, 2, \dots, p \quad (3)$$

Prova

- Com relação à (2) e (3), seja \mathbf{E} uma matriz em que as linhas são os auto-vetores $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$ (definidos anteriormente), ou seja (ortono-normalizados)

$$\mathbf{E} = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1p} \\ e_{21} & e_{22} & \dots & e_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ e_{p1} & e_{p2} & \dots & e_{pp} \end{bmatrix} = \begin{bmatrix} \mathbf{e}'_1 \\ \mathbf{e}'_2 \\ \vdots \\ \mathbf{e}'_p \end{bmatrix}$$

- Vimos, anteriormente, que Σ for uma matriz positiva definida então $\Sigma = \mathbf{E}'\mathbf{\Lambda}\mathbf{E}$, em que $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ (autovalores), e as colunas da matriz \mathbf{E}' são formadas pelos respectivos autovetores

- Assim, temos que

$$\text{Cov}(\mathbf{Y}) = \text{Cov}(\mathbf{E}\mathbf{X}) = \mathbf{E}\text{Cov}(\mathbf{X})\mathbf{E}' = \mathbf{E}\Sigma\mathbf{E}' = \mathbf{E}\mathbf{E}'\Lambda\mathbf{E}\mathbf{E}' = \Lambda$$

pois,

$$\begin{aligned} \mathbf{E}\mathbf{E}' &= \begin{bmatrix} \mathbf{e}'_1 \\ \mathbf{e}'_2 \\ \vdots \\ \mathbf{e}'_p \end{bmatrix} \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_p \end{bmatrix} = \begin{bmatrix} \mathbf{e}'_1\mathbf{e}_1 & \mathbf{e}'_1\mathbf{e}_2 & \dots & \mathbf{e}'_1\mathbf{e}_p \\ \mathbf{e}'_2\mathbf{e}_1 & \mathbf{e}'_2\mathbf{e}_2 & \dots & \mathbf{e}'_2\mathbf{e}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{e}'_p\mathbf{e}_1 & \mathbf{e}'_p\mathbf{e}_2 & \dots & \mathbf{e}'_p\mathbf{e}_p \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \mathbf{I}_p \end{aligned}$$

Prova

- A prova de (1), vem do que fato de que $\max_{\mathbf{a} \neq \mathbf{0}} \frac{\mathbf{a}' \Sigma \mathbf{a}}{\mathbf{a}' \mathbf{a}}$ é obtido quando $\mathbf{a} = \mathbf{e}_1$ e o $\min_{\mathbf{a} \neq \mathbf{0}} \frac{\mathbf{a}' \Sigma \mathbf{a}}{\mathbf{a}' \mathbf{a}}$ é obtido quando $\mathbf{a} = \mathbf{e}_p$ (veja página 80, do livro do Johnson and Wichern (2007), 7ª edição). Como $\mathbf{e}'_1 \mathbf{e}_1 = 1$, temos que $\max_{\mathbf{a} \neq \mathbf{0}} \frac{\mathbf{a}' \Sigma \mathbf{a}}{\mathbf{a}' \mathbf{a}} = \max_{\mathbf{a} \neq \mathbf{0}} \mathbf{a}' \Sigma \mathbf{a}$ e o mesmo vale para o mínimo.
- Similarmente, $\max_{\mathbf{a} \neq \mathbf{0} \perp \mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k} \frac{\mathbf{a}' \Sigma \mathbf{a}}{\mathbf{a}' \mathbf{a}}$ é obtido quando $\mathbf{a} = \mathbf{e}_{k+1}$.

Componentes principais e variáveis originais

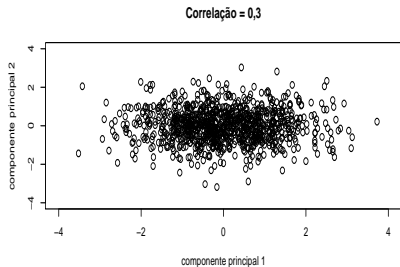
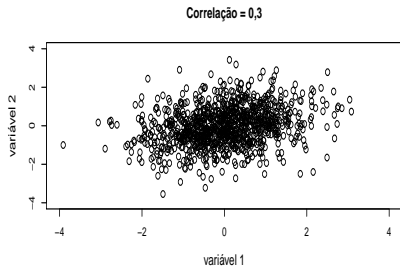
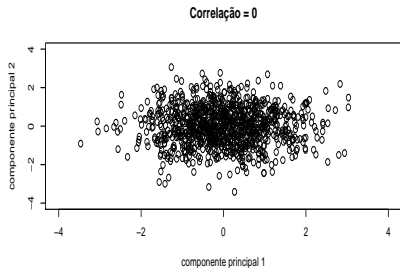
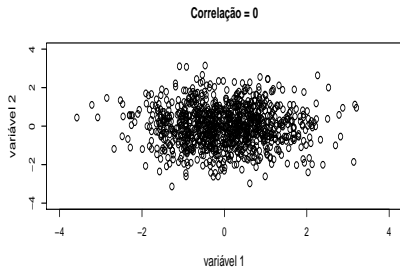
- Defina $\mathbf{a}_j = [0 \ 0 \dots \underbrace{1}_{\text{posição } j} \dots 0 \ 0]$
- Temos que $\text{Cov}(Y_i, X_j) = \text{Cov}(\mathbf{e}_i' \mathbf{X}, \mathbf{a}_j' \mathbf{X}) = \mathbf{e}_i' \text{Cov}(\mathbf{X}, \mathbf{X}) \mathbf{a}_j = \mathbf{e}_i' \boldsymbol{\Sigma} \mathbf{a}_j = \lambda_i \mathbf{e}_i' \mathbf{a}_j = \lambda_i e_{ij}$. (Provar que $\mathbf{e}_i' \boldsymbol{\Sigma} = \lambda_i \mathbf{e}_i'$).
- Assim $\text{Corre}(Y_i, X_j) = \frac{\text{Cov}(Y_i, X_j)}{\sqrt{\mathcal{V}(Y_i)} \sqrt{\mathcal{V}(X_j)}} = \frac{\lambda_i e_{ij}}{\sqrt{\lambda_i} \sigma_j} = \frac{\sqrt{\lambda_i} e_{ij}}{\sigma_j}$, em que $\sigma_j = DP(X_j)$.
- Pode-se mostrar que $\text{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^p \mathcal{V}(X_i) = \sum_{i=1}^p \lambda_i = \sum_{i=1}^p \mathcal{V}(Y_i)$

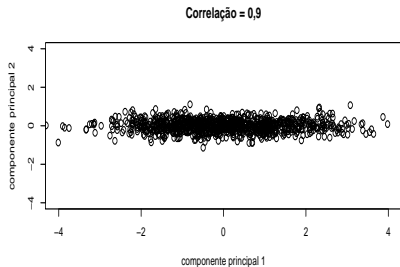
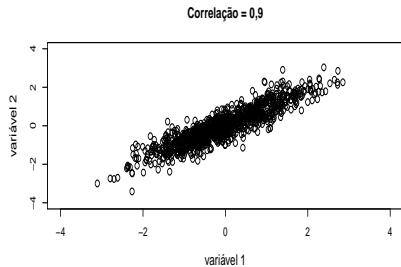
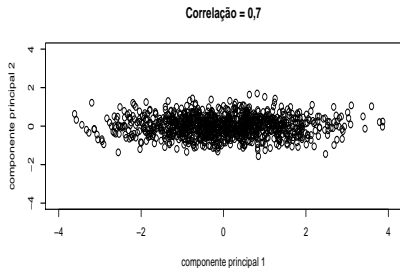
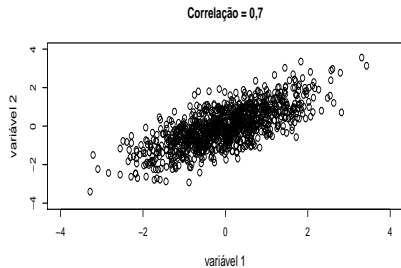
Comentários

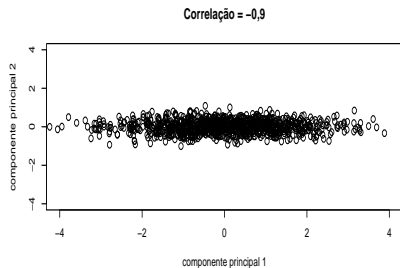
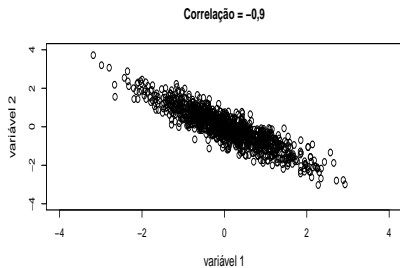
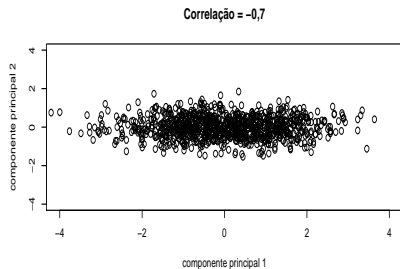
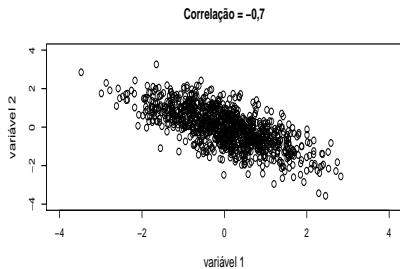
- Há problemas em se trabalhar com a matriz de covariâncias (Σ). As componentes principais tendem a ser influenciadas pela variabilidade (escala) das variáveis.
- Quanto maior a variabilidade de uma determinada variável, mais ele influenciará as componentes.
- Alternativa: trabalhar com a matriz de correlações (ρ) (o que é equivalente a trabalhar com variáveis com variância unitária).
- Nesse caso as variáveis, portanto as componentes principais, serão adimensionais.

Comentários

- O procedimento de obtenção das componentes principais continua o mesmo mas, nesse caso, trabalharemos com a matriz de correlações.
- Temos que $Cov(Y_i, X_j) = Cov(\mathbf{e}'_i \mathbf{X}, \mathbf{a}'_j \mathbf{X}) = \mathbf{e}'_i Cov(\mathbf{X}, \mathbf{X}) \mathbf{a}_j = \mathbf{e}'_i \boldsymbol{\rho} \mathbf{a}_j = \lambda_i \mathbf{e}'_i \mathbf{a}_j = \lambda_i e_{ij}$. (Provar que $\mathbf{e}'_i \boldsymbol{\rho} = \lambda_i \mathbf{e}'_i$).
- Além disso, $Corre(Y_i, X_j) = \frac{Cov(Y_i, X_j)}{\sqrt{V(Y_i)}\sqrt{V(X_j)}} = \frac{\lambda_i e_{ij}}{\sqrt{\lambda_i} 1} = \sqrt{\lambda_i} e_{ij}$.
- Pode-se também trabalhar com as variáveis padronizadas (média 0 e variância 1).
- Como determinar o número de componentes principais e como utilizá-las?







Estimação das componentes principais

- Dada uma matriz de dados $\mathbf{X}_{(n \times p)}$

Indivíduo	Variável 1	Variável 2	...	Variável p
1	X_{11}	X_{12}	...	X_{1p}
2	X_{21}	X_{22}	...	X_{2p}
\vdots	\vdots	\vdots	\ddots	\vdots
n	X_{n1}	X_{n2}	...	X_{np}

estima-se a matriz de covariâncias ($\tilde{\Sigma}$) ou a matriz de correlações amostrais ($\tilde{\rho}$), conforme visto anteriormente.

Estimação das componentes principais

- Obtem-se os auto-valores e auto-vetores ortonormalizados, digamos $(\tilde{\lambda}_1, \tilde{\mathbf{e}}_1), (\tilde{\lambda}_2, \tilde{\mathbf{e}}_2), \dots, (\tilde{\lambda}_p, \tilde{\mathbf{e}}_p)$ a partir de $\tilde{\Sigma}$ ou $\tilde{\rho}$.
- Assim $\tilde{\mathcal{V}}(Y_{ij}) = \tilde{\lambda}_j$.
- Cada componente principal será calculada como $y_{ij} = \tilde{\mathbf{e}}_j \mathbf{x}_i$, em que $\mathbf{x}_i, i = 1, \dots, n, j = 1, \dots, p$ é a i -ésima linha da matriz de dados observada.
- Alternativamente, podemos estimar as componentes através de $y_{ij} = \tilde{\mathbf{e}}_j \mathbf{z}_i$, em que \mathbf{z}_i corresponde às variáveis padronizadas (default do R).

Exemplo 4: índices de criminalidade de estados americanos

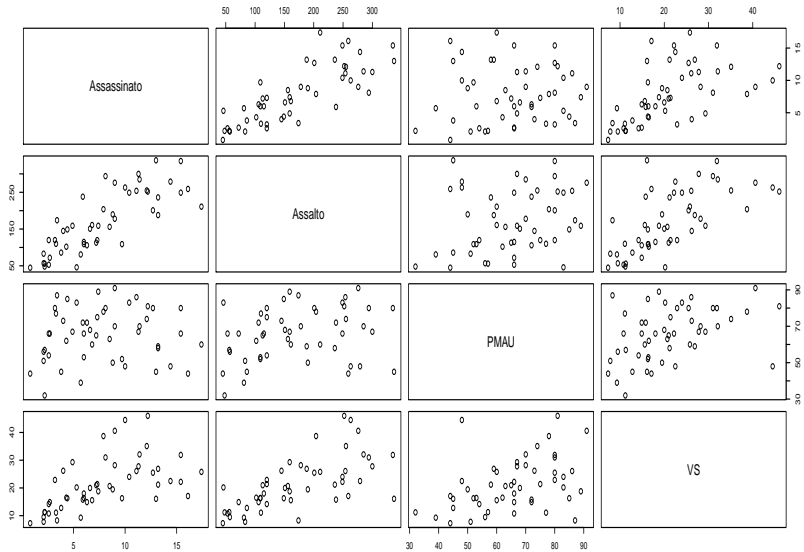
- Número de prisões efetuadas (por 100.000 habitantes) em 50 estados americanos em 1973.
- Variáveis: assalto, violência sexual (VS), assassinato e porcentagem de moradores na área urbana (PMAU).

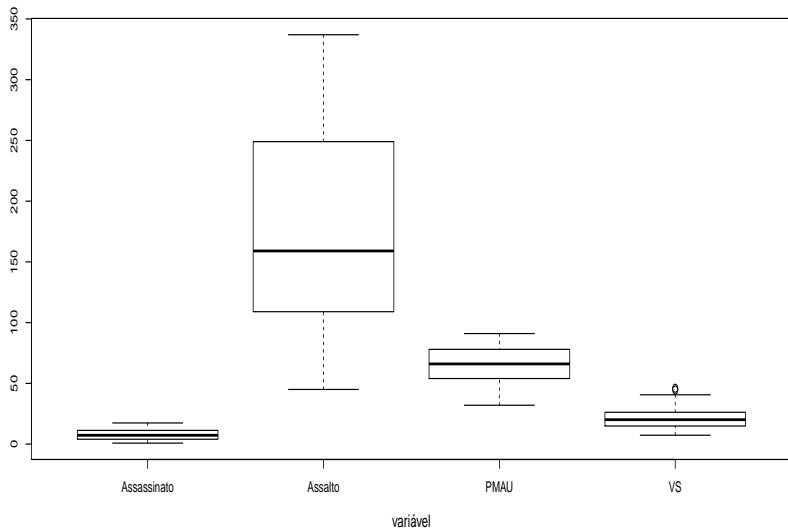
Estado	Assassinato	Assalto	PMAU	VS
Alabama	13,2	236,0	58	21,2
Alaska	10,0	263,0	48	44,5
Arizona	8,1	294,0	80	31,0
⋮	⋮	⋮	⋮	⋮

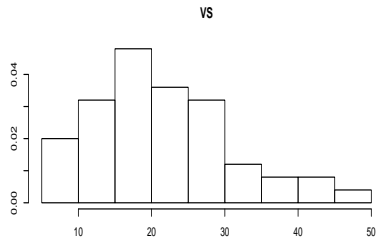
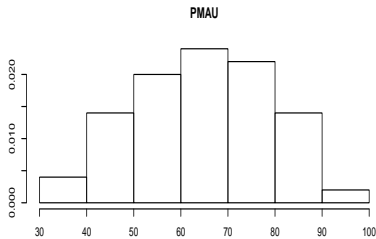
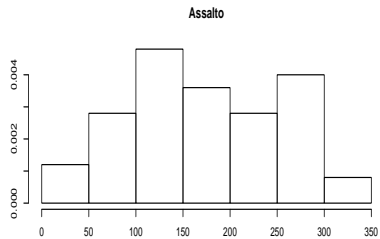
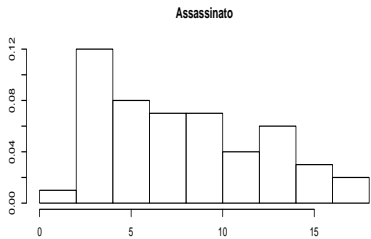
- Digitar *USArrests* no programa *R*.

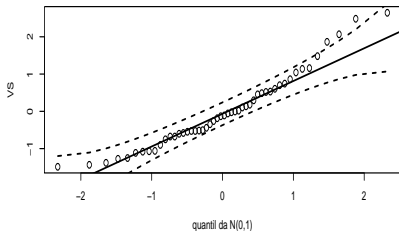
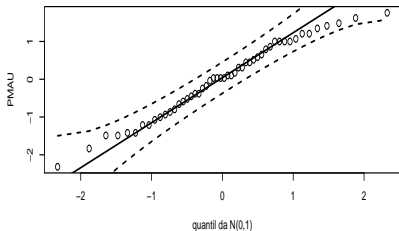
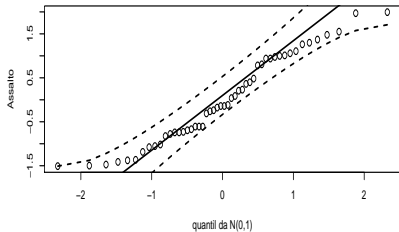
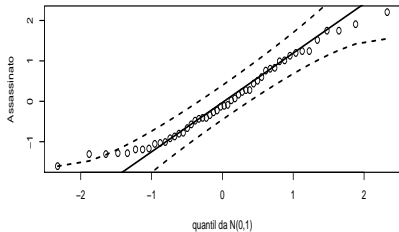
■ Medidas descritivas

Estatística	Assassinato	Assalto	PMAU	VS
Média	7,79	170,76	65,54	21,23
Var.	18,97	6.945,17	209,52	87,73
DP	4,36	83,34	14,47	9,37
CV(%)	55,93	48,80	22,09	44,11
Mínimo	0,80	45,00	32,00	7,30
Mediana	7,25	159,00	66,00	20,10
Máximo	17,40	337,00	91,00	46,00









■ Matriz de covariâncias

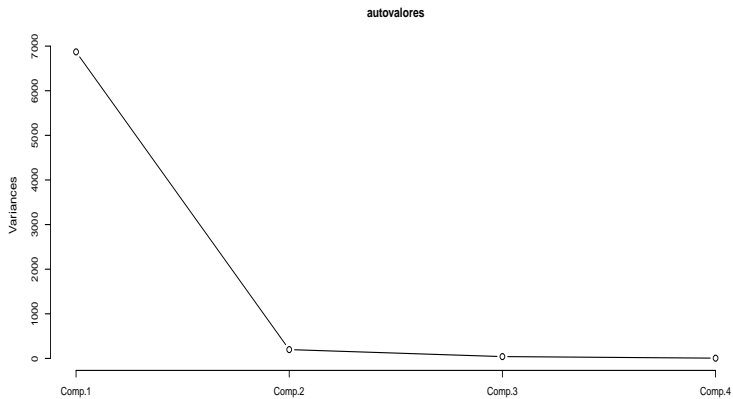
	Assassinato	Assalto	PMAU	VS
Assassinato	18,97	291,06	4,38	22,99
Assalto	291,06	6945,16	312,27	519,26
PMAU	4,38	312,27	209,51	55,76
VS	22,99	519,26	55,76	87,72

■ Autovalores

7011,11 201,99 42,11 6,16

■ Autovetores

-0,04170432	0,04482166	0,07989066	0,99492173
-0,99522128	0,05876003	-0,06756974	-0,03893830
-0,04633575	-0,97685748	-0,20054629	0,05816914
-0,07515550	-0,20071807	0,97408059	-0,07232502
0,04170432	-0,04482166	-0,07989066	-0,99492173
0,99522128	-0,05876003	0,06756974	0,03893830
0,04633575	0,97685748	0,20054629	-0,05816914
0,07515550	0,20071807	-0,97408059	0,07232502



- Variância explicada

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
PVE (%)	96,55	2,78	< 0.01	< 0.01
PVEA (%)	96,55	99,33	99,91	100,00

- Componentes principais

	Comp. 1	Comp. 2
Assassinato	0,04 (0,80)	-0,04 (-0,14)
Assalto	0,99 (0,99)	-0,06 (<-0,01)
PMAU	0,04 (0,26)	0,98 (0,96)
VS	0,07 (0,67)	0,20 (0,30)

- Utilizar a matriz de correlações.

■ Matriz de correlações

	Assassinato	Assalto	PMAU	VS
Assassinato	1,00	0,80	0,06	0,56
Assalto	0,80	1,00	0,25	0,66
PMAU	0,07	0,25	1,00	0,41
VS	0,56	0,66	0,41	1,00

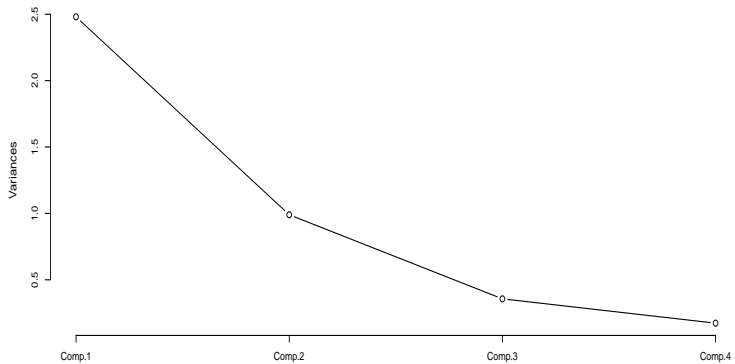
■ Autovalores

2,48 0,98 0,35 0,17

■ Autovetores

0,5358995	-0,4181809	0,3412327	-0,64922780
0,5831836	-0,1879856	0,2681484	0,74340748
0,2781909	0,8728062	0,3780158	-0,13387773
0,5434321	0,1673186	-0,8177779	-0,08902432

autovalores



■ Variância explicada

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
PVE (%)	62,00	24,74	8,91	4,32
PVEA (%)	62,00	86,75	95,67	100,00

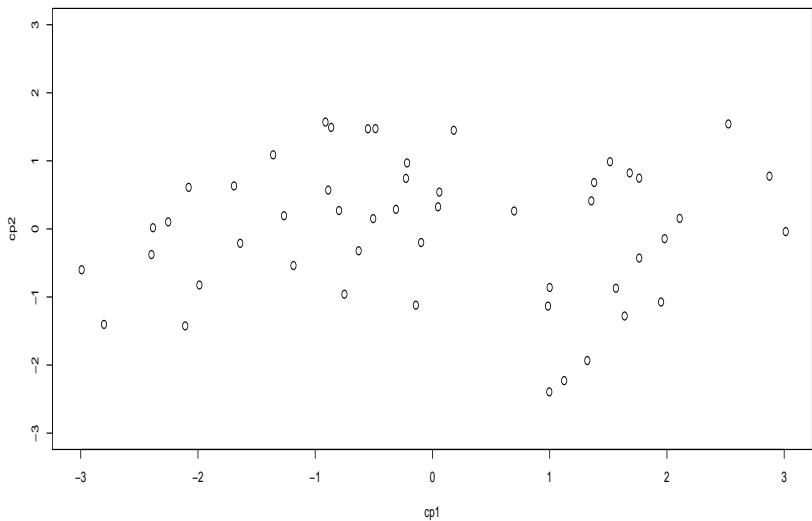
■ Componentes principais

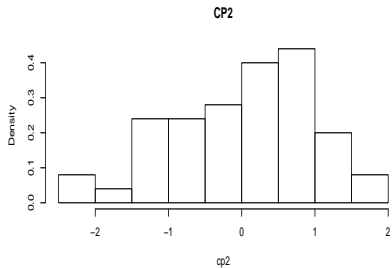
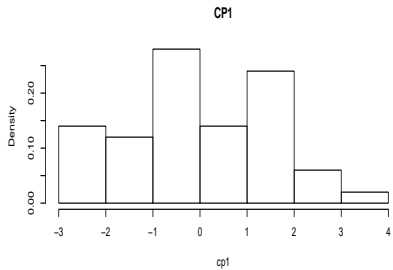
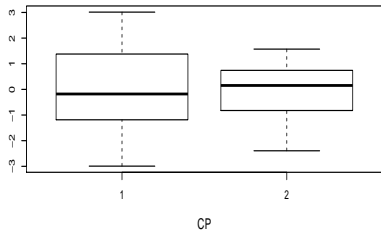
	Comp. 1	Comp. 2
Assassinato	0,53 (0,84)	-0,41(-0,41)
Assalto	0,58 (0,91)	-0,18(-0,19)
PMAU	0,27 (0,43)	0,86(0,87)
VS	0,54 (0,85)	0,16(0,17)

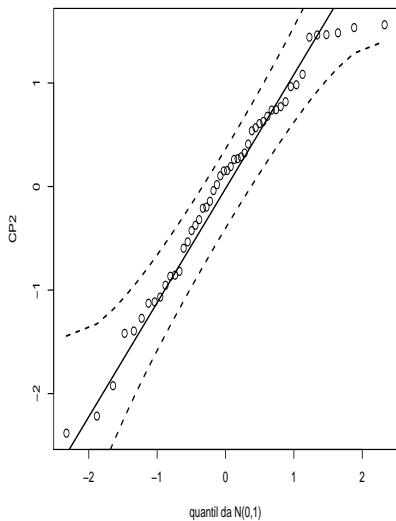
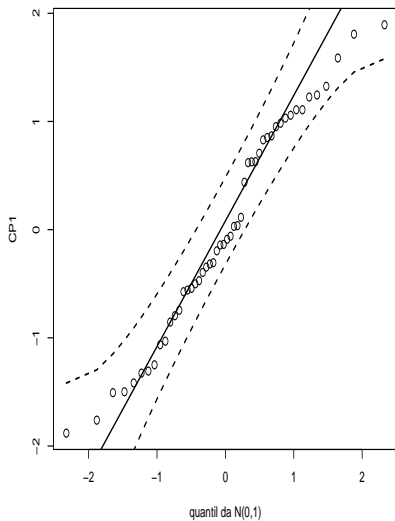
- Primeira componente: “escore ponderado” entre as variáveis, com maiores pesos para as variáveis criminais.
- Segunda componente: é um contraste entre PMAU e VS e as outras variáveis criminais, com maior peso para a variável PMAU.

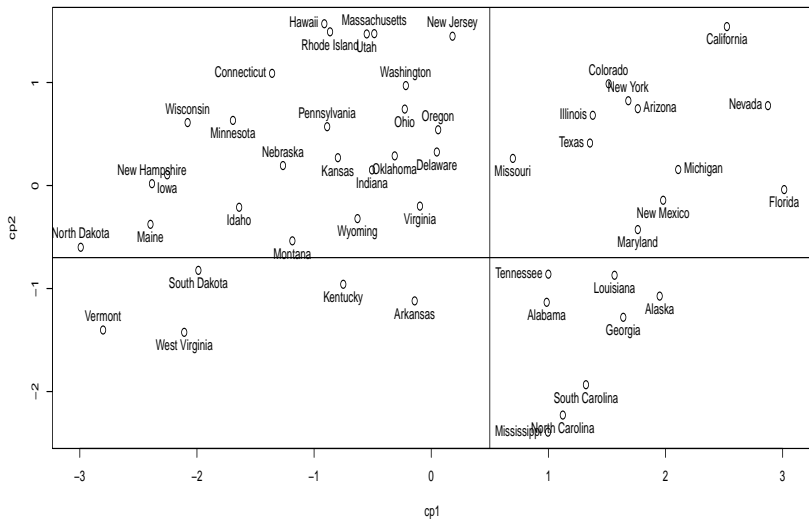
- Função “princomp”. Resultado salvo num objeto chamado result.

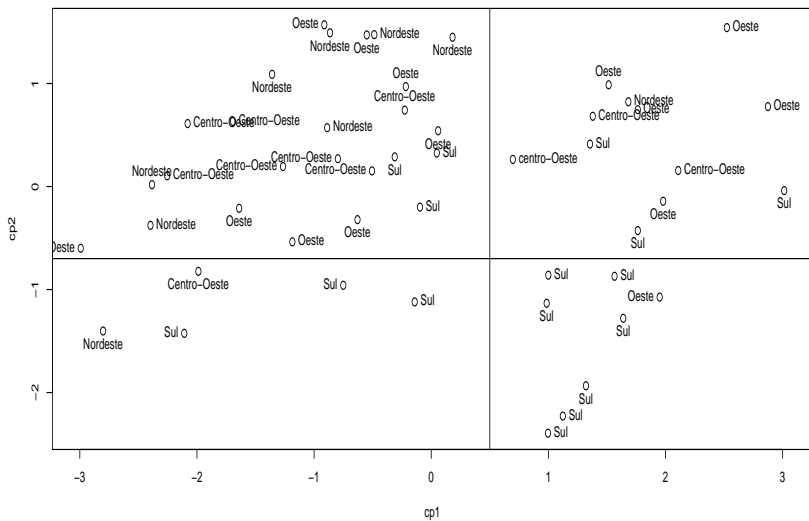
Comando	Significado
result\$loadings	coeficientes que geram as CP's (geralmente os autovetores negativos)
result\$scores	estimativa das componentes principais
screepLOT(result)	screepLOT

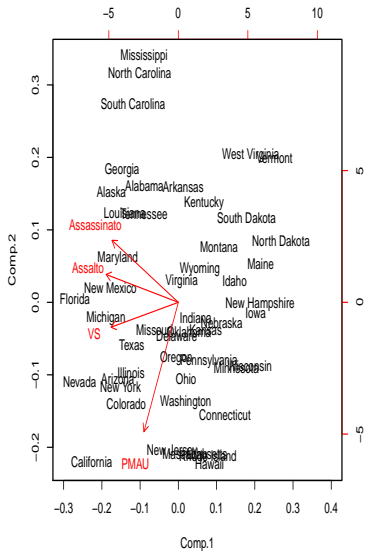












Biplot

- Um gráfico com quatro eixos.
- Os pontos representam as observações enquanto que os vetores (setas) representam as variáveis.
- Os eixos da parte de baixo e da esquerda se referem aos valores das duas componentes principais divididos pelos respectivos desvios-padrão (das componentes) vezes a raiz quadrada do tamanho da amostra ($y_{ij}^* = \frac{y_{ij}}{\sqrt{\lambda_i n}}$).
- Os eixos das partes de cima e da direita representam os valores dos coeficientes das componentes multiplicados pelos respectivos desvios-padrão (das componentes) vezes a raiz quadrada do tamanho da amostra ($e_{ij}^* = e_{ij} \sqrt{\lambda_i n}$).