MÉTODOS ESTATÍSTICOS PARA METEOROLOGIA

Ralph dos Santos Silva

https://www.ralphsilva.org/meteorologia.html

Instituto de Matemática
Centro de Ciências Matemáticas e da Natureza
Universidade Federal do Rio de Janeiro

Referências

- Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*, 3th Edition. Academic Press.
- Johnson e Wichern (2007). *Applied Multivariate Statistical Analysis*, 6th Edition. Prentice-Hall.

Organização de dados

- A matéria prima a ser trabalhada aqui é um conjunto de dados multivariados, isto é, várias variáveis são observadas sobre diversos indivíduos ou objetos. Nosso objetivo será apresentar uma forma conveniente de organizar estes dados e de representá-los graficamente.
- Suponha que estejamos diante de um problema em que p variáveis foram observadas para uma amostra de n elementos. Assim, a observação para o i-ésimo elemento da amostra será um vetor p-variado denotador por $\boldsymbol{x}_{.i}$ ou $\underline{x}_{.i}$ tal que

$$\boldsymbol{x}_{.i}^{\top} = (x_{1i}, x_{2i}, \ldots, x_{pi}), \qquad i = 1, 2, \ldots, n,$$

  em que x_{ji} representa a observação da j-ésima variável do i-ésimo elemento da amostra, $j = 1, 2, \ldots, p$.
- A coleção de dados observados pode ser representada por meio de uma matriz $\boldsymbol{X}$ de dimensão $p \times n$ como segue

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{bmatrix}.$$

- Assim, as linhas da matriz $\boldsymbol{X}$ representam as p variáveis medidas e, as colunas, as n unidades amostrais.
- Podemos representar a matriz $\boldsymbol{X}$ através de suas linhas

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{1.}^{\top} \\ \boldsymbol{x}_{2.}^{\top} \\ \vdots \\ \boldsymbol{x}_{p.}^{\top} \end{bmatrix},$$

  sendo $\boldsymbol{x}_{j.}^{\top} = (x_{j1}, x_{j2}, \ldots, x_{jn})$ as n observações da j-ésima variável, $j = 1, 2, \ldots, p$.
- Também podemos representar a matriz $\boldsymbol{X}$ através de suas colunas. Adotando aqui a notação $\boldsymbol{x}_{.i}$, $i = 1, 2, \ldots, n$, para designar a i-ésima coluna de $\boldsymbol{X}$, temos

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{.1} & \boldsymbol{x}_{.2} & \ldots & \boldsymbol{x}_{.n} \end{bmatrix},$$

  em que cada $\boldsymbol{x}_{.i}$ é um vetor $p \times 1$, $i = 1, 2, \ldots, n$.

Exemplo 1:

- Uma seleção de 4 notas ficais de uma livraria universitária foi obtida de modo a investigar a natureza das vendas. Cada nota forneceu o número de livros vendidos e o valor total da venda (em dólares). Obteve-se a seguinte matriz de dados, na qual a primeira linha indica o número de livros vendidos e, a segunda, o valor da venda.

$$\boldsymbol{X} = \begin{bmatrix} 4 & 5 & 4 & 3 \\ 42 & 52 & 48 & 58 \end{bmatrix}.$$

- A representação dos dados desta forma permite o cálculo de quantidades numéricas de interesse de forma eficiente e fácil.

Estatísticas descritivas

1. A média amostral para a j-ésima variável observada pode ser representada como

$$\overline{x}_{j.} = \frac{1}{n}\sum_{i=1}^{n} x_{ji}, \qquad j = 1, 2, \ldots, p.$$

Assim, podemos definir o vetor de médias amostral $\overline{\boldsymbol{x}}$ como

$$\overline{\boldsymbol{x}} = \begin{bmatrix} \overline{x}_{1.} \\ \overline{x}_{2.} \\ \vdots \\ \overline{x}_{p.} \end{bmatrix}.$$

Algebricamente,

$$\overline{\boldsymbol{x}} = \frac{1}{n}\boldsymbol{X}\mathbf{1},$$

sendo $\mathbf{1}$ um vetor $n \times 1$ com todos os elementos iguais a 1.

2. Uma medida de dispersão para a j-ésima variável é dada pela variância amostral

$$s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ji} - \overline{x}_{j.})^2, \qquad j = 1, 2, \ldots, p.$$

3. A covariância amostral entre a j-ésima e a k-ésima variáveis é dada por

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ji} - \overline{x}_{j.})(x_{ki} - \overline{x}_{k.}), \qquad j, k = 1, 2, \ldots, p \quad \text{e} \quad j \neq k.$$

- Podemos então representar de forma organizada as informações sobre variabilidade através da matriz de (variâncias e) covariâncias amostral dada por

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \ldots & s_{pp} \end{bmatrix}.$$

Observe que a matriz de covariâncias é uma matriz simétrica: $\boldsymbol{S} = \boldsymbol{S}^\top$.
(A matriz de covariância $\boldsymbol{S}$ também é positiva definida.)

- Algebricamente, podemos escrever a matriz de dados menos o vetor de médias na forma

$$\boldsymbol{X} - \bar{\boldsymbol{x}}\mathbf{1}^\top,$$

e a matriz de covariâncias na forma

$$\boldsymbol{S} = \frac{1}{n-1}(\boldsymbol{X} - \bar{\boldsymbol{x}}\mathbf{1}^\top)(\boldsymbol{X} - \bar{\boldsymbol{x}}\mathbf{1}^\top)^\top.$$

4. Podemos também definir a matriz de correlações amostral $\boldsymbol{R}$, com elementos $r_{jk} = \dfrac{s_{jk}}{\sqrt{s_{jj}s_{kk}}}$,

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{bmatrix}.$$

Observe que a matriz de correlações é uma matriz simétrica: $\boldsymbol{R} = \boldsymbol{R}^\top$.
(A matriz de correlação $\boldsymbol{R}$ também é positiva definida.)

- Algebricamente, a matriz de correlações pode ser escrita na forma

$$\boldsymbol{S} = \frac{1}{n-1}\left[\Delta^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{x}}\mathbf{1}^\top)\right]\left[\Delta^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{x}}\mathbf{1}^\top)\right]^\top$$

com

$$\Delta^{1/2} = \text{diag}(s_{11}^{1/2}, s_{22}^{1/2}, \ldots, s_{pp}^{1/2})$$

uma matriz diagonal $p \times p$, e $\Delta^{-1/2} = \left[\Delta^{1/2}\right]^{-1}$.

- Observe que podemos relacionar algebricamente as matrizes $\boldsymbol{S}$ e $\boldsymbol{R}$ tal que

$$\boldsymbol{S} = \Delta^{1/2}\boldsymbol{R}\Delta^{1/2} \quad \text{e} \quad \boldsymbol{R} = \Delta^{-1/2}\boldsymbol{S}\Delta^{-1/2}.$$

- Em muitas aplicações as somas dos desvios quadrados da média e dos produtos cruzados de tais desvios são utilizadas. Adotaremos aqui a notação:

$$\begin{aligned} W_{jj} &= \sum_{i=1}^{n} (x_{ji} - \overline{x}_{j.})^2, \qquad j = 1, 2, \ldots, p, \quad \text{e} \\ W_{jk} &= \sum_{i=1}^{n} (x_{ji} - \overline{x}_{j.})(x_{ki} - \overline{x}_{k.}), \qquad j, k = 1, 2, \ldots, p \quad \text{e} \quad j \neq k. \end{aligned}$$

Assim, definimos a matriz $\boldsymbol{W}$ de somas dos desvios quadrados da média e produtos cruzados dos desvios da média por

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} & \ldots & w_{1p} \\ w_{21} & w_{22} & \ldots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \ldots & w_{pp} \end{bmatrix} = (\boldsymbol{X} - \overline{\boldsymbol{x}}\boldsymbol{1}^\top)(\boldsymbol{X} - \overline{\boldsymbol{x}}\boldsymbol{1}^\top)^\top.$$

Exemplo 1: (continuação)

Para os dados do Exemplo 1 e utilizando o programa R, faça os cálculos:

- do vetor de médias; e das matrizes ***S*** e ***R***. (Ver exemplo_01.r)

```
p       <- 2                          # numero de variaveis
n       <- 4                          # tamanho da amostra
X       <- matrix(0,p,n)              # definindo X
X[1, ]  <- c( 4,  5,  4,  3)          # valores de X_{1i}
X[2, ]  <- c(42, 52, 48, 58)          # valores de X_{1i}
ones    <- matrix(1,n,1)              # vetor de uns
xbar    <- (X%*%ones) / n             # vetor de medias
sdX     <- apply(X,1,"sd")            # d. padrao das variaveis
D.half  <- diag(sdX,p,p)              # matriz Delta^{1/2}
X.xbar  <- X-xbar%*%t(ones)           # matriz (X - xbarra*ones')
W       <- X.xbar%*%t(X.xbar)         # somas dos desvios ao
                                      # quadrados e cruzados
S       <- W/(n-1)                    # matriz de covariancias
ID.half <- solve(D.half)              # matriz Delta^{-1/2}
R       <- ID.half%*%S%*%ID.half      # matriz de correlacoes
```

```
> xbar # vetor de medias
     [,1]
[1,]    4
[2,]   50
> S    # matriz de covariancias
           [,1]     [,2]
[1,]  0.6666667 -2.00000
[2,] -2.0000000 45.33333
> R    # matriz de correlacoes
           [,1]       [,2]
[1,]  1.0000000 -0.3638034
[2,] -0.3638034  1.0000000
```

Usando as funções do R, podemos calcular:

```
apply(X,1,"mean") # vetor (linha) de medias
var(t(X))         # matriz de covariancias
cor(t(X))         # matriz de correlacoes
# IMPORTANTE: note a transposicao da matriz X.
```

Representação gráfica

- Podemos utilizar gráficos de dispersão para variáveis duas a duas. (Ver exemplo_02.r)

```
# Note que o R entende variaveis por colunas e
# as amostras por linha. Entao e' necessario
# transpor a matriz de dados
p       <- 3                          # numero de variaveis
n       <- 10000                      # tamanho da amostra
X       <- matrix(rnorm(p*n),p,n) # definindo X
#pdf(file="dispersao_pairs.pdf")
#par(mfrow=c(1,1),lwd=2.0,cex.lab=1.5,cex.axis=1.5,
#    lab=c(10,5,5),mar=c(0,1,0,2.5),xpd=T,cex.main=2.0)
pairs(t(X),pch=15)
#dev.off()

apply(X,1,"mean")  # vetor (linhas) de medias
var(t(X))          # matriz de covariancias
cor(t(X))          # matriz de correlacoes
```

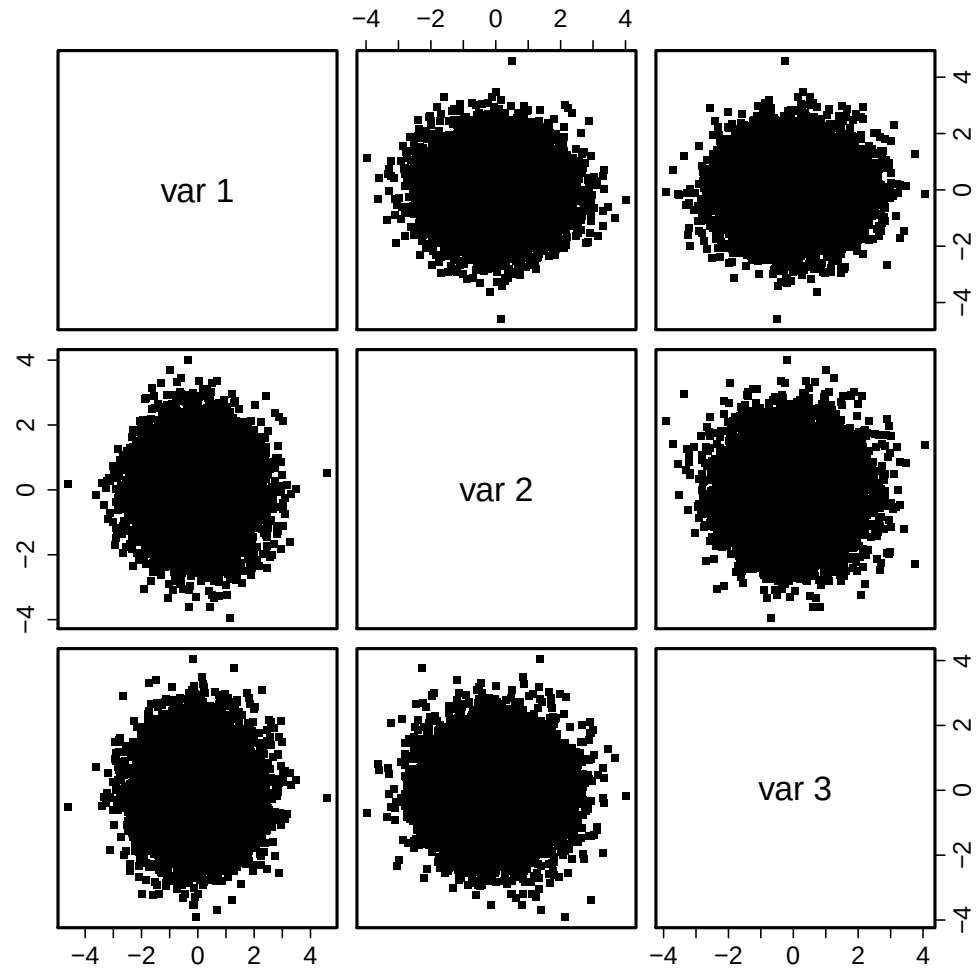

Figura 1: Exemplo de gráfico de dispersão de variáveis duas a duas.

Distâncias

- A maior parte das técnicas multivariadas baseia-se no simples conceito de distância.
- Estamos habituados a distância usual chamada distância euclideana, tal que se $P(x_1, x_2, \ldots, x_p)$ e $Q(\mu_1, \mu_2, \ldots, \mu_p)$ são dois pontos em $\mathbb{R}^p$, a distância entre P e Q é dada por

$$d_e(P, Q) = \sqrt{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 + \cdots + (x_p - \mu_p)^2}.$$

- Porém, a distância euclideana pode não ser adequada em muitos problemas, dependendo da natureza das variáveis envolvidas.
- Isto ocorre devido ao fato de que na distância euclideana cada coordenada contribui igualmente para o cálculo da mesma.
- Quando as coordenadas representam medições que são sujeitas à flutuações aleatórias de magnitudes diferentes, é frequentemente desejável ponderar coordenadas sujeitas à maior variabilidade com um peso menor do que àquelas sujeitas a uma menor variabilidade.

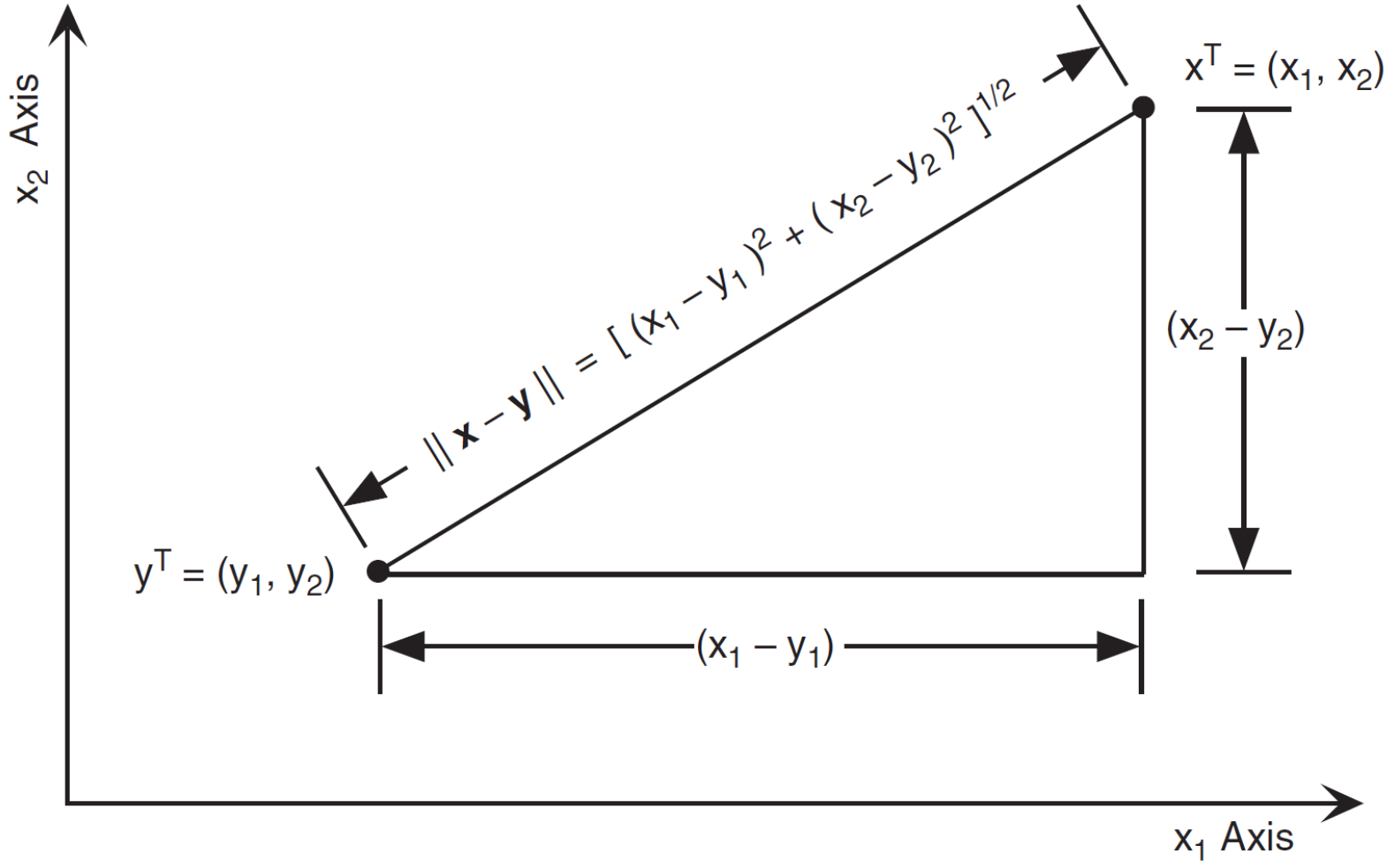

Figura 2: Exemplo da distância euclideana entre os pontos ***x*** e ***y***.

- Deseja-se uma nova medida de distância que leve em conta as diferenças em variabilidade entre as diversas variáveis incluídas na análise e a presença de correlação entre os pares de variáveis.
- Suponha primeiro um conjunto de p variáveis não correlacionadas, com variâncias distintas. Assim, de forma a equilibrar a contribuição das diversas variáveis ao cálculo da distância, podemos ponderá-las de forma inversamente proporcional aos seus desvios padrão ($\sqrt{s_{jj}}$) e calculando a distância euclideana

$$d_e(P, Q) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{D}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})},$$

  sendo $\boldsymbol{D}$ uma matriz diagonal com elementos $s_{11}, s_{22}, \ldots, s_{pp}$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ e $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^\top$.
- Uma medida de distância que leva em conta as covariâncias entre as variáveis é dada por

$$d_e(P, Q) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}, \quad \text{(distância de Mahalanobis)},$$

  sendo $\boldsymbol{S}$ a matriz de covariâncias.

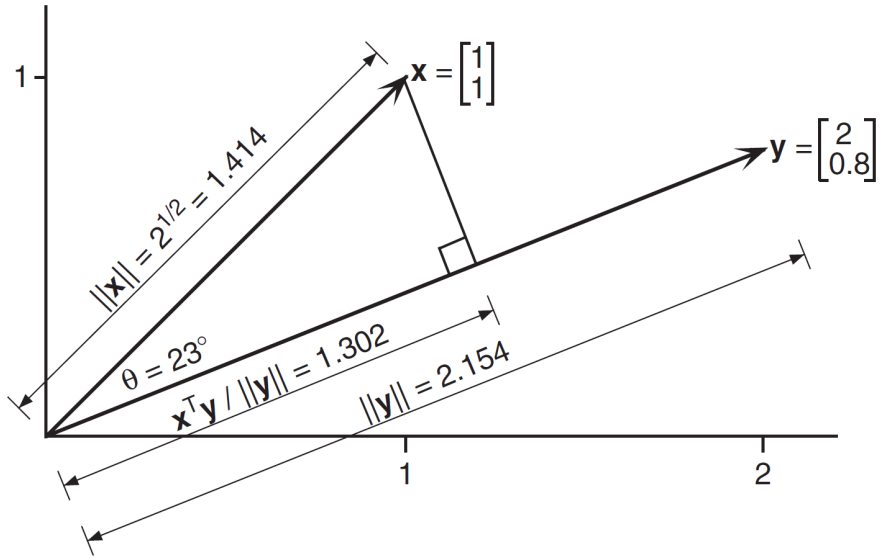

Figura 3: Exemplo do ângulo entre os vetores $\boldsymbol{x}$ e $\boldsymbol{y}$.

- Tamanho de um vetor $\boldsymbol{x}$: $\|\boldsymbol{x}\| = \left[\sum_{i=1}^{p} x_i^2\right]^{1/2}$ (norma euclideana).
- Projeção de $\boldsymbol{x}$ em $\boldsymbol{y}$: $L_{\boldsymbol{x},\boldsymbol{y}} = \dfrac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{y}\|}$.
- O ângulo θ entre dois vetores: $\theta = \cos^{-1}\left[\dfrac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}\right]$.
- Se $\boldsymbol{x}^\top \boldsymbol{y} = 0$, então $\theta = 90^o$, i.é., vetores perpendiculares (ortogonal).

Distribuição normal (univariada)

Seja X uma variável aleatória contínua com distribuição normal com média μ e variância σ^2. Então, a função de densidade de probabilidade é dada por

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},$$

para $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ e $\sigma^2 \in \mathbb{R}^+$.

Temos que

$$\begin{aligned} \Pr(|X-\mu| \leqslant \sigma) &\simeq 0,683; \\ \Pr(|X-\mu| \leqslant 2\sigma) &\simeq 0,954; \quad \text{e} \\ \Pr(|X-\mu| \leqslant 3\sigma) &\simeq 0,997. \end{aligned}$$

Sabemos também que $\mathrm{E}(X) = \mu$ e $\mathrm{Var}(X) = \sigma^2$.

Distâncias

Sabemos que a distância euclideana entre dois pontos, $P(\boldsymbol{x})$ e $Q(\boldsymbol{\mu})$, no $\mathbb{R}^p$ é dado por

$$\begin{aligned} d_e(P,Q) &= \sqrt{(x_1-\mu_1)^2+(x_2-\mu_2)^2+\cdots+(x_p-\mu_p)^2} \\ &= \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})}. \end{aligned}$$

Se $d=1$, então $d_e(x,\mu)=\sqrt{(x-\mu)^2}$.

Uma generalização da distância euclideana é dada por

$$d_e(P,Q)=\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Psi}(\boldsymbol{x}-\boldsymbol{\mu})}.$$

Neste caso, temos pesos diferentes para cada valor do vetor e também levamos em conta as possíveis interações (correlações).

No núcleo da função de densidade da normal univariada, temos

$$\left(\frac{x-\mu}{\sigma}\right)^2=\frac{(x-\mu)^2}{\sigma^2}=(x-\mu)(\sigma^{-2})(x-\mu).$$

Isto lembra o quadrado da distância euclideana modificada para $d=1$.

Assim, o núcleo de uma normal multivariada é dado pelo quadrado da distância euclideana modificada entre os pontos $\boldsymbol{x}$ e $\boldsymbol{\mu}$ no $\mathbb{R}^p$, isto é, por

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}),$$

sendo a matriz de pesos dada por $\boldsymbol{\Sigma}^{-1}$, e $\boldsymbol{\Sigma}$ corresponde a matriz de covariâncias (simétrica e positiva definida).

Note que quanto maior a variância de determinada variável (componente do vetor aleatório $\boldsymbol{X}$), menor será seu peso no cálculo da distância em questão.

Para determinar a função de densidade de probabilidade da normal multivariada precisamos somente da constante de normalização. Prova-se que a constante é dada por

$$(2\pi)^{-p/2} \left|\boldsymbol{\Sigma}\right|^{-1/2},$$

com $|\boldsymbol{A}| = \det(A)$ por definição.

Função de densidade de probabilidade da normal multivariada

Definição: A função de densidade de probabilidade da normal multivariada - com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$ - é dada por

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &= (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \end{aligned}$$

com $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{\mu} \in \mathbb{R}^p$ e $\boldsymbol{\Sigma}$ uma matriz $p \times p$ simétrica e positiva definida.

Prova-se que $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ e $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$ (matriz simétrica e positiva definida).

Notação: $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Exemplo: Normal bivariada

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \end{bmatrix}$$

e

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) \end{bmatrix},$$

sendo $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1) \Rightarrow \sigma_{12} = \sigma_{21}$. Logo,

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} &= \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix} \\ &= \frac{1}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)} \begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{11} \end{bmatrix}, \end{aligned}$$

pois $\rho_{12} = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$ (correlação entre X_1 e X_2).

Segue-se que

$$
\begin{aligned}
& (\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\
= \quad & \begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 \end{bmatrix} \frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)} \begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{11} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \\
= \quad & \frac{\sigma_{22}(x_1-\mu_1)^2+\sigma_{11}(x_2-\mu_2)^2-2\rho_{12}\sqrt{\sigma_{11}\sigma_{22}}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)} \\
= \quad & \frac{1}{1-\rho_{12}^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right].
\end{aligned}
$$

Note que se as variáveis X_1 e X_2 são não correlacionadas, tal que $\rho_{12}=0$, então a densidade conjunta pode ser reescrita como o produto de duas densidades normais. Neste caso, $f(x_1,x_2|\boldsymbol{\mu},\boldsymbol{\Sigma}) = f(x_1|\mu_1,\sigma_{11})f(x_2|\mu_2,\sigma_{22})$, e, portanto, X_1 e X_2 são variáveis aleatórias independentes.

(Ver exemplo_03.r)

```
library(mvtnorm)
nx <- 100; ny <- 100
x <- seq(-4,4,length=nx); y <- seq(-3,3,length=ny);
z <- matrix(NA,nx,ny)
mu <- c(0,0); S <- matrix(c(1, .9, .9, 1),2,2)
for(i in 1:nx)
    for(j in 1:ny)
        z[i,j] <- dmvnorm(c(x[i],y[j]), mean=mu,sigma=S)
# Grafico 1
x11(); persp(x,y,z, phi = 45, theta = 30,col=3)
# Grafico 2
x11(); contour(x,y,z,col=2,lwd=2)
#--------------------------------
w <- rmvnorm(100000,mean=mu,sigma=S)
#Grafico 3
x11(); par(mfrow=c(2,2))
plot(w,pch=".",main="",xlab="x",ylab="y")
hist(w[,2],prob=T,nclass=50,main="",xlab="y")
lines(y,dnorm(y),lwd=2,col=2)
hist(w[,1],prob=T,nclass=50,main="",xlab="x")
lines(x,dnorm(x),lwd=2,col=2)
```

Algumas propriedades da normal multivariada

Seja $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Então,

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}) &= \mathrm{E}[(X_1, \ldots, X_p)^\top] = (\mathrm{E}(X_1), \ldots, \mathrm{E}(X_p))^\top = (\mu_1, \ldots, \mu_p)^\top = \boldsymbol{\mu}. \\ \mathrm{Var}(\boldsymbol{X}) &= \mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top] = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}^\top = \boldsymbol{\Sigma}, \end{aligned}$$

em que $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$.

Note que $\mathrm{Var}(\boldsymbol{X})_{j,k} = \mathrm{Cov}(X_j, X_k)$ para $j, k = 1, \ldots, p$.

A matriz de covariâncias, $\boldsymbol{\Sigma}$, é simétrica e positiva definida.

Seja $\boldsymbol{A}$ uma matriz $m \times p$, $\boldsymbol{b}$ um vetor $m \times 1$ e $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Então,

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} \sim \mathcal{N}_m(\boldsymbol{\xi}, \boldsymbol{\Psi}),$$

em que $\boldsymbol{\xi} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$ e $\boldsymbol{\Psi} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$.

Combinação linear de variáveis aleatórias normais (correlacionadas ou não) tem distribuição normal.

Decomposição da matriz $\boldsymbol{\Sigma}$

- Podemos sempre representar a distribuição normal multivariada como uma transformação linear de variáveis aleatórias normais univariadas?
- Podemos sempre representar a distribuição normal multivariada como uma transformação linear de variáveis aleatórias normais padrão univariadas?
- Podemos sempre ver a distribuição normal multivariada como uma elipse p-dimensional?

Sim é a resposta para todas as perguntas acima.

Podemos, por exemplo, empregar a decomposição em valores singulares (SVD) a matriz de covariâncias $\boldsymbol{\Sigma}$ para obter

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top,$$

sendo $\boldsymbol{V}$ uma matriz $p \times p$ ortonormal (as colunas da matriz são ortogonais e cada uma tem norma 1) e $\boldsymbol{D}$ uma matriz $p \times p$ diagonal com elementos d_i, $i = 1, \ldots, p$, positivos.

Isto significa que qualquer vetor normalmente distribuído $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$, pode ser escrito como $\boldsymbol{X} = \boldsymbol{V}(\boldsymbol{W} - \boldsymbol{\mu}) + \boldsymbol{\mu}$, em que $\boldsymbol{W} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{D})$ e as coordenadas de $\boldsymbol{W}$, W_j, são independentes e normalmente distribuídas com média μ_j e variância d_j, $j = 1, \ldots, p$.

Suponha que $\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$ e que $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2}\boldsymbol{V}^\top = \boldsymbol{H}\boldsymbol{H}^\top$, em que $\boldsymbol{D}^{1/2}$ é a matriz formada pelas raizes quadradas dos elementos da diagonal de $\boldsymbol{D}$, e $\boldsymbol{H} = \boldsymbol{V}\boldsymbol{D}^{1/2}$. Então, para $\boldsymbol{X} = \boldsymbol{H}\boldsymbol{Z} + \boldsymbol{\mu}$, temos

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}) &= \mathrm{E}(\boldsymbol{H}\boldsymbol{Z} + \boldsymbol{\mu}) = \boldsymbol{H}\mathrm{E}(\boldsymbol{Z}) + \boldsymbol{\mu} = \boldsymbol{H}\boldsymbol{0} + \boldsymbol{\mu} = \boldsymbol{\mu} \quad \text{e} \\ \mathrm{Var}(\boldsymbol{X}) &= \mathrm{Var}(\boldsymbol{H}\boldsymbol{Z}) = \boldsymbol{H}\mathrm{Var}(\boldsymbol{Z})\boldsymbol{H}^\top = \boldsymbol{H}\boldsymbol{I}\boldsymbol{H}^\top \\ &= \boldsymbol{V}\boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top = \boldsymbol{\Sigma}. \end{aligned}$$

Portanto, $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Também podemos utilizar a **decomposição de Cholesky** para obter $\boldsymbol{\Sigma} = \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{L}\boldsymbol{L}^\top$, em que $\boldsymbol{U}$ e $\boldsymbol{L}$ são matrizes triangulares superior e inferior, respectivamente. Neste caso, temos que $\boldsymbol{L} = \boldsymbol{U}^\top$.

Gere valores da normal multivariada utilizando o SVD no programa R.

Gerando valores da normal multivariada

(Ver exemplo_04.r)

```
p     <- 3      # dimensao
n     <- 100000 # tamanho da amostra
# estrutura de covariancia (correlacao)
SIGMA <- matrix(c(1,0.9,0.8,0.9,1,0.7,0.8,0.7,1),p,p,T)
# normais independentes
x     <- matrix(rnorm(p*n),n,p) # note a transposicao
#-------------------------------------------------
# Usando o SVD
S     <- svd(SIGMA)             # decomposicao de SIGMA
T     <- S$u%*%diag(sqrt(S$d)) # construcao da matriz T
y     <- x%*%t(T)               # gerando valores
cor(y); SIGMA
#-------------------------------------------------
# Usando a decomposicao de Cholesky
T     <- chol(SIGMA)
y     <- x%*%T
cor(y); SIGMA
```

Distribuição normal multivariada: marginais e condicionais

Seja $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tal que

$$\begin{aligned}
\boldsymbol{X} &= \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \text{ e } \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \text{ com dimensões } \begin{bmatrix} q \times 1 \\ (p-q) \times 1 \end{bmatrix}, \text{ e} \\
\boldsymbol{\Sigma} &= \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \text{ com dimensões } \begin{bmatrix} q \times q & q \times (p-q) \\ (p-q) \times q & (p-q) \times (p-q) \end{bmatrix}.
\end{aligned}$$

em que $p > q \geqslant 1$. Então,

$$\begin{aligned}
&\boldsymbol{X}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \\
&\boldsymbol{X}_2 \sim \mathcal{N}_{(p-q)}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}), \quad \text{e} \\
&(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{\xi}) \sim \mathcal{N}_q(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}),
\end{aligned}$$

em que

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\xi} - \boldsymbol{\mu}_2) \\
\widetilde{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.
\end{aligned}$$

Algumas propriedades

- Se $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então $(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathcal{X}^2_p$.
- Calculando uma região (elipse) que tenha probabilidade η, i.é.,

$$\Pr((\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \leqslant \chi^2_\eta) = \eta,$$

  com χ^2_η representando o η-ésimo percentil da distribuição $\mathcal{X}^2_p$.
- Sejam $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ uma AAS de $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ e $T^2 = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu})$ (estatística T^2 de Hotelling). Então,

$$\frac{n-p}{(n-1)p} T^2 \sim \mathcal{F}_{p,n-p}.$$

  Natureza da estatística T^2:

$$\underbrace{\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{\mu})^\top}_{\mathcal{N}_p(\boldsymbol{0},\boldsymbol{\Sigma})} \left\{ \overbrace{\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^\top}^{\left(\frac{\mathcal{W}_p(n-1,\boldsymbol{\Sigma})}{n-1}\right)} \right\}^{-1} \underbrace{\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{\mu})^\top}_{\mathcal{N}_p(\boldsymbol{0},\boldsymbol{\Sigma})}.$$

Exercício

- Escreva um programa no `R` para gerar números do vetor aleatório normalmente distribuído com variâncias todas iguais a 1 (um) e correlações todas iguais a um dos valores no conjunto $\{-0,9;\ 0,0;\ 0,9;\ 0,99\}$, e varie o tamanho das amostras geradas para cada matriz de covariâncias (por exemplo, 100, 1000, etc). Considere uma normal trivariada para este exercício. Faça diversos gráficos para analisar os conjuntos de dados gerados. Por exemplo *box-plot*, histogramas, *qq-plot*, dispersão dois a dois, etc. É importante neste exercício que você trabalhe com a matriz de dados de acordo com a parte teórica da matéria, ou seja, salve os dados gerados em uma matriz $p \times n$.

Dica: verificar as funções `boxplot`, `hist`, `qqnorm` e `plot` do programa `R`.

Análise em Componentes Principais - ACP

O propósito principal da ACP é substituir as variáveis originais por um número menor de variáveis que são função das variáveis originais.

A ACP consiste na determinação de uma transformação ortogonal das variáveis originais para um novo conjunto de variáveis não correlacionadas que são obtidas em ordem decrescente de importância.

As novas variáveis, chamadas de componentes, são combinações lineares das variáveis originais.

De forma a estudar as relações entre um conjunto de p variáveis correlacionadas, pode ser útil transformar o conjunto de dados em um novo conjunto de variáveis não correlacionadas chamados Componentes Principais. Estas novas variáveis são combinações lineares das variáveis originais e são obtidas em ordem decrescente de importância tal que, por exemplo, o primeiro componente principal conta com a maior parte possível da variação total nos dados originais. A transformação ortogonal é, de fato, uma rotação espaço original das p variáveis.

A técnica de obtenção desta transformação é chamada Análise em Componentes Principais (ACP). A ACP é apropriada quando as variáveis sob investigação são todas de mesma natureza, de modo que não tenhamos, por exemplo, uma variável dependente e várias variáveis explicativas como no caso da regressão múltipla.

O objetivo usual da análise é verificar se os primeiros poucos componentes principais (CP's) compreendem a maior parte da variação total dos dados originais. Se for este o caso, então coloca-se a questão de que a dimensionalidade efetiva dos dados é menor do que p.

Em outras palavras, se algumas das variáveis originais são altamente correlacionadas, elas estão, efetivamente, "dizendo a mesma coisa" e podem portanto, existir restrições lineares sobre estas variáveis. Neste caso é esperado que os primeiros componentes sejam significativos, nos ajudem a compreender melhor os dados e sejam úteis nas análises subsequentes, nas quais poderemos lidar com um número menor de variáveis.

Na prática, em diversos casos, não é fácil interpretar os componentes e, portanto, sua principal utilidade está na redução da dimensionalidade dos dados de forma a simplificar as análises posteriores.

Componentes Principais Populacionais

Algebricamente, CP's são combinações lineares particulares das variáveis $X_1, X_2, \ldots, X_p$.

Geometricamente, estas combinações lineares representam a seleção de um novo sistema de coordenadas rotacionando-se o sistema original. Os novos eixos representam as direções com variabilidade máxima e fornecem descrição mais simples e parcimoniosa da estrutura de covariância.

Como veremos, os CP's dependem somente da estrutura de covariância $\mathbf{\Sigma}$ (ou da matriz de correlações $\boldsymbol{\mathcal{R}}$). Seu desenvolvimento não exige que os dados sejam normais multivariados. Por outro lado, componentes principais obtidos de populações normais multivariadas têm interpretações úteis em função dos elipsóides de densidade constante. Além disso, inferências podem ser feitas a partir dos componentes amostrais, quando a população é normal multivariada.

Suponha $\boldsymbol{X}$ uma v.a. p-dimensional, $\boldsymbol{X}^\top = (X_1, X_2, \ldots, X_p)$, com média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. Nosso problema aqui é encontrar um novo conjunto de variáveis, a saber $\boldsymbol{Y}^\top = (Y_1, Y_2, \ldots, Y_p)$, tal que $Y_1, Y_2, \ldots, Y_p$ são não correlacionadas e cujas variâncias decrescem da primeira para a última.

Cada Y_j é tomada como uma combinação linear dos componentes do vetor $\boldsymbol{X}$ tal que

$$Y_j = a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jp}X_p = \boldsymbol{a}_j^\top \boldsymbol{X} \tag{1}$$

em que $\boldsymbol{a}_j^\top = (a_{j1}, a_{j2}, \ldots, a_{jp})$ é um vetor de constantes.

A Equação 1 contém um fator de escala arbitrário. Portanto, impomos a condição $\boldsymbol{a}_j^\top \boldsymbol{a}_j = \sum_{k=1}^{p} a_{jk}^2 = 1$. Veremos que este procedimento particular de normalização assegura que a transformação global é ortogonal. Em outras palavras, assegura que distâncias e ângulos no novo espaço são preservados.

O primeiro componente principal, Y_1, é encontrado escolhendo-se $\boldsymbol{a}_1$ tal que Y_1 tenha a maior variância possível. Em outras palavras, escolhemos $\boldsymbol{a}_1$ de forma a

$$\begin{aligned} \text{Maximizar} \quad & \text{Var}(\boldsymbol{a}_1^\top \boldsymbol{X}) \\ \text{Sujeito à} \quad & \boldsymbol{a}_1^\top \boldsymbol{a}_1 = 1. \end{aligned}$$

O segundo componente principal é encontrado escolhendo-se $\boldsymbol{a}_2$ tal que Y_2 tenha a (segunda) maior variância possível para todos os compostos definidos pela equação 1 que são não correlacionados com Y_1.

Similarmente, derivamos os componentes $Y_3, Y_4, \ldots, Y_p$ de forma que eles sejam não correlacionadas com os anteriores e tenham variâncias decrescentes.

Proposição 1: Seja $\boldsymbol{\Sigma}$ a matriz de covariâncias de $\boldsymbol{X}$. Sejam $(\lambda_1, \boldsymbol{v}_1), (\lambda_2, \boldsymbol{v}_2), \ldots, (\lambda_p, \boldsymbol{v}_p)$ os pares de autovalores e autovetores (ortonormalizados) de $\boldsymbol{\Sigma}$ tais que $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$.

O j-ésimo componente principal é dado por

$$Y_j = \boldsymbol{v}_j^\top \boldsymbol{X}, \quad j = 1, 2, \ldots, p.$$

Com estas escolhas, tem-se:

$$\text{Var}(Y_j) = \boldsymbol{v}_j^\top \boldsymbol{\Sigma} \boldsymbol{v}_j = \lambda_j, \quad j = 1, 2, \ldots, p$$

e

$$\text{Cov}(Y_j, Y_k) = \boldsymbol{v}_j^\top \boldsymbol{\Sigma} \boldsymbol{v}_k = 0, \quad j \neq k.$$

Se alguns λ_j forem repetidos, as escolhas dos correspondentes autovetores $\boldsymbol{v}_j$ e, portanto de Y_j, não será única.

Proposição 2: Seja $\boldsymbol{X}$ vetor aleatório de dimensão $p \times 1$ com matriz de covariâncias $\boldsymbol{\Sigma}$ tal que $(\lambda_1, \boldsymbol{v}_1), (\lambda_2, \boldsymbol{v}_2), \ldots, (\lambda_p, \boldsymbol{v}_p)$ são os pares de autovalores e autovetores (ortonormalizados) de $\boldsymbol{\Sigma}$ com $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$. Então,

$$\sum_{j=1}^{p} \text{Var}(X_j) = \sum_{j=1}^{p} \sigma_{jj} = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{j=1}^{p} \text{Var}(Y_j).$$

Dem.: Temos que $\sum_{j=1}^{p} \sigma_{jj} = \text{tr}(\boldsymbol{\Sigma})$. Mas $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{P}^\top$ com $\boldsymbol{\Delta} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$ e $\boldsymbol{P} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p]$, matriz ortogonal.

Logo,

$$\text{tr}(\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{P}^\top) = \text{tr}(\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{\Delta}) = \text{tr}(\boldsymbol{\Delta}) = \sum_{j=1}^{p} \lambda_j.$$

A Proposição 2 diz que a variância total da população é dada pela soma dos autovalores da matriz de covariâncias populacional, $\boldsymbol{\Sigma}$.

Consequentemente, a proporção da variação total devida ao j-ésimo componente é

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}, \quad \textit{para} \quad j = 1, 2, \ldots, p.$$

Se a maior parte (por exemplo 80% a 90%) da variação total, para p grande, pode ser atribuído aos dois ou três primeiros componentes, então estes componentes podem "substituir" as p variáveis originais sem muita perda de informação.

Cada entrada dos vetores que definem os componentes $\boldsymbol{v}_j^\top = [v_{j1}, v_{j2}, \ldots, v_{jp}]$ também deve ser olhada. A magnitude de v_{jk} indica a importância da k-ésima variável original no j-ésimo componente, sem olhar as demais variáveis.

Em particular, v_{jk} é proporcional ao coeficiente de correlação entre Y_j e X_k.

Proposição 3: Se $Y_j = \boldsymbol{v}_j^\top \boldsymbol{X}$, $j = 1, 2, \ldots, p$ são os componentes principais obtidos a partir da matriz de covariâncias $\boldsymbol{\Sigma}$ do vetor aleatório $\boldsymbol{X}$, então

$$\rho_{Y_j, X_k} = \frac{v_{jk}\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}, \quad j, k = 1, 2, \ldots, p$$

são os coeficientes de correlação entre Y_j e X_k.

$(\lambda_1, \boldsymbol{v}_1), (\lambda_2, \boldsymbol{v}_2), \ldots, (\lambda_p, \boldsymbol{v}_p)$ são os pares de autovalores e autovetores (ortonormalizados) de $\boldsymbol{\Sigma}$ com $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$.

(Ver exemplo_10.r)

Exemplo 1: Suponha um vetor aleatório de dimensão 3 cuja matriz de covariâncias é dada por (Ver exemplo_05.r)

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Os pares de autovalores e autovetores associados a $\boldsymbol{\Sigma}$ são:

$$\begin{array}{llll} \lambda_1 \simeq 5,83 & \text{e} & \boldsymbol{v}_1^\top \simeq [-0,383 \quad 0,924 \quad 0]; & \\ \lambda_2 \simeq 2,00 & \text{e} & \boldsymbol{v}_2^\top \simeq [\ 0 \quad\quad 0 \quad\quad 1]; & \text{e} \\ \lambda_3 \simeq 0,17 & \text{e} & \boldsymbol{v}_3^\top \simeq [\ 0,924 \quad 0,383 \quad 0]. & \end{array}$$

Portanto, os componentes principais são:

$$\begin{aligned} Y_1 &= -0,383X_1 + 0,924X_2 \\ Y_2 &= X_3 \\ Y_3 &= 0,924X_1 + 0,383X_2. \end{aligned}$$

A variável X_3 é um dos componentes principais, pois ela é não correlacionada com as outras duas variáveis.

A proporção da variação total explicada pelos dois primeiros componentes é $7,83/8 \simeq 0,98$ de modo que podemos explicar estes dados usando apenas Y_1 e Y_2, sem perder quase informação.

As correlações entre os componentes e as respectivas variáveis são dadas na tabela a seguir.

Tabela 1: Correlações entre os componentes e as respectivas variáveis.

CP	X_1	X_2	X_3
Y_1	-0,924	0,997	0
Y_2	0	0	1
Y_3	0,383	0,071	0

É informativo considerar componentes principais obtidos de populações normais multivariadas. Suponha $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Vimos que a densidade de $\boldsymbol{X}$ é constante para os pontos $\boldsymbol{x}$ em $\mathbb{R}^p$ que satifazem

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c^2.$$

As soluções desta equação correspondem ao elipsóide centrado em $\boldsymbol{\mu}$ cujos semi-eixos são dados por $\pm c\sqrt{\lambda_j}\boldsymbol{v}_j$, $j = 1, 2, \ldots, p$ em que $(\lambda_1, \boldsymbol{v}_1), (\lambda_2, \boldsymbol{v}_2), \ldots, (\lambda_p, \boldsymbol{v}_p)$ são os pares de autovalores e autovetores (ortonormalizados) de $\boldsymbol{\Sigma}$.

Um ponto caindo no j-ésimo eixo do elipsóide de densidade constante terá coordenadas proporcionais às coordenadas do vetor

$$\boldsymbol{v}_j^\top = [v_{j1}, v_{j2}, \ldots, v_{jp}],$$

no sistema de coordenadas que tem origem em $\boldsymbol{\mu}$ e eixos que são paralelos aos eixos das variáveis originais $X_1, X_2, \ldots, X_p$.

Será conveniente aqui considerar $\boldsymbol{\mu} = \mathbf{0}$, sem perda de generalidade.

Vimos que os pares de autovalores e autovetores de $\boldsymbol{\Sigma}^{-1}$ são dados por $(1/\lambda_1, \boldsymbol{v}_1), (1/\lambda_2, \boldsymbol{v}_2), \ldots, (1/\lambda_p, \boldsymbol{v}_p)$.

Com $\boldsymbol{\mu} = \mathbf{0}$, temos

$$\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} = \sum_{j=1}^{p} \frac{1}{\lambda_j} (\boldsymbol{v}_j^\top \boldsymbol{x})^2 = \sum_{j=1}^{p} \frac{1}{\lambda_j} y_j^2 = \frac{1}{\lambda_1} y_1^2 + \frac{1}{\lambda_2} y_2^2 + \cdots + \frac{1}{\lambda_p} y_p^2 = c^2,$$

e esta equação define um elipsóide (pois $\lambda_1, \lambda_2, \ldots, \lambda_p$ são positivos) em um sistema de coordenadas $y_1, y_2, ..., y_p$ definido pelos eixos cujas direções são $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$ respectivamente.

Se λ_1 é o maior autovalor, então o eixo maior está na direção do autovetor $\boldsymbol{v}_1$. Os eixos menores restantes estão nas direções definidas pelos autovetores $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$.

Resumindo: os componentes principais caem nas direções dos eixos de um elipsóide de densidade constante. Portanto, qualquer ponto sobre o j-ésimo eixo do elipsóide tem coordenadas proporcionais ao j-ésimo autovetor.

Quando $\boldsymbol{\mu} \neq \mathbf{0}$, seu componente principal centrado na média é $y_j = \boldsymbol{v}_j^\top (\boldsymbol{x} - \boldsymbol{\mu})$ que tem média zero e está na direção $\boldsymbol{v}_j$.

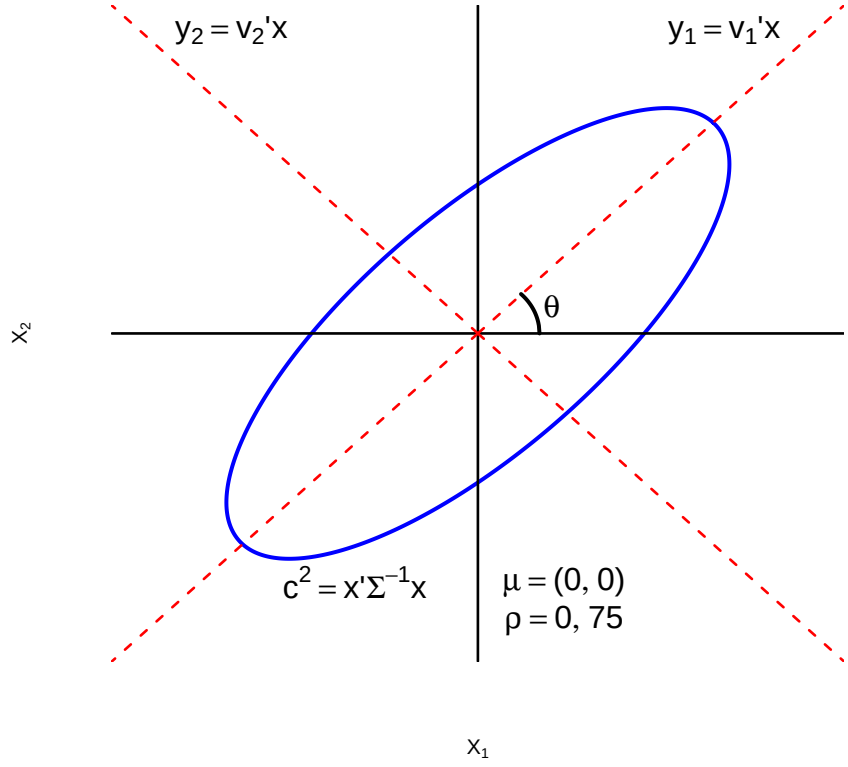

Figura 4: Elipse de densidade constante e os componentes principais para um vetor aleatório normal bivariado com $\boldsymbol{\mu} = \mathbf{0}$ e $\rho = 0,75$.

É possível ver que os CP's são obtidos pela rotação em relação à origem, dos eixos originais de um ângulo θ no sentido anti-horário até que eles coincidam com os eixos da elipse de densidade constante. Este resultado também vale para $p > 2$.

Componentes Principais Obtidos das Variáveis Padronizadas

Os componentes também podem ser obtidos via variáveis padronizadas. Na notação matricial $\boldsymbol{Z} = (\boldsymbol{\Psi}^{1/2})^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ com $\boldsymbol{\Psi}^{1/2} = \text{diag}\{\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \ldots, \sqrt{\sigma_{pp}}\}$.

Claramente $\text{E}(\boldsymbol{Z}) = \boldsymbol{0}$ e $\text{Var}(\boldsymbol{Z}) = \boldsymbol{\mathcal{R}}$.

Os componentes principais de $\boldsymbol{Z}$ podem ser obtidos da matriz de correlações $\boldsymbol{\mathcal{R}}$ de $\boldsymbol{X}$.

Os resultados anteriores todos se aplicam, com algumas simplificações, pois agora a variância total é p, pois na diagonal de $\boldsymbol{\mathcal{R}}$ os elementos são todos unitários.

Nesse caso, a proporção total da variação devida ao j-ésimo componente será dada por $\dfrac{\delta_j}{p}$, para $j = 1, 2, \ldots, p$, sendo δ_j o j-ésimo autovalor da matriz de correlações $\boldsymbol{\mathcal{R}}$.

Porém, os componentes obtidos a partir da matriz de correlações serão diferentes dos componentes obtidos pela matriz de covariâncias.

Exemplo 2: Considere a seguinte matriz de covariâncias $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$ e a correspondente matriz de correlações $\boldsymbol{\mathcal{R}} = \begin{bmatrix} 1 & 0,4 \\ 0,4 & 1 \end{bmatrix}$.

(Ver exemplo_06.r)

Os pares de autovalores e autovetores de $\boldsymbol{\Sigma}$ são dados por

$$\lambda_1 = 100,16 \quad \text{e} \quad \boldsymbol{v}_1^\top = [0,040 \quad -0,999]$$
$$\lambda_2 = 0,839 \quad \text{e} \quad \boldsymbol{v}_2^\top = [0,999 \quad 0,040].$$

Os pares de autovalores e autovetores de $\boldsymbol{\mathcal{R}}$ são dados por

$$\delta_1 = 1,4 \quad \text{e} \quad \boldsymbol{w}_1^\top = [\ 0,707 \quad 0,707]$$
$$\delta_2 = 0,6 \quad \text{e} \quad \boldsymbol{w}_2^\top = [-0,707 \quad 0,707].$$

Os respectivos componentes principais são, então,

- relativos a $\boldsymbol{\Sigma}$:
  $Y_1 = 0,040X_1 - 0,999X_2$, e
  $Y_2 = 0,999X_1 + 0,040X_2$
  com Y_1 correspondendo a 99,17% da variação total.
- relativos a $\boldsymbol{\mathcal{R}}$:
  $$Y_1^\star = 0,707\frac{(X_1-\mu_1)}{1} + 0,707\frac{(X_2-\mu_2)}{10} = 0,707Z_1 + 0,707Z_2, \quad \text{e}$$
  $$Y_2^\star = -0,707\frac{(X_1-\mu_1)}{1} + 0,707\frac{(X_2-\mu_2)}{10} = -0,707Z_1 + 0,707Z_2$$
  com $Y_1^\star$ correspondendo a 70% da variação total.

Devido à variância de X_2 ser muito maior que a de X_1, esta variável domina o primeiro componentes principal determinado pela matriz $\boldsymbol{\Sigma}$.

Quando as variáveis X_1 e X_2 são padronizadas, as variáveis resultantes contribuem igualmente para os componentes principais determinados pela matriz $\boldsymbol{\mathcal{R}}$ e as correlações de Z_1 e Z_2 com o primeiro componente são $0,837$.

Mais surpreendentemente, vemos que a importância relativa das variáveis para o primeiro componente é fortemente afetada pela padronização. Quando o primeiro componente obtido de $\boldsymbol{\mathcal{R}}$ é expresso em função de X_1 e X_2, as magnitudes relativas dos pesos que são 0,707 e 0,707/10=0,0707 estão em oposição direta aos pesos do primeiro componente relativo a $\boldsymbol{\Sigma}$ dados por 0,040 e 0,999, respectivamente.

Este exemplo demonstra que componentes principais obtidas de $\boldsymbol{\Sigma}$ são diferentes daqueles obtidos de $\boldsymbol{\mathcal{R}}$. Além disso, um conjunto de componentes principais não é uma função simples do outro. Isto sugere que a padronização não é inconsequente.

As variáveis devem provavelmente ser padronizadas se elas são medidas sobre escalas cujas variações são muito diferentes ou se as unidades de medida não são comensuráveis. Por exemplo, se X_1 representa vendas anuais de dez mil a 350 mil dólares e X_2 é a razão (renda bruta anual)/(ativos totais) que varia entre 0,01 a 0,60, então a variação total será quase que exclusivamente devida às vendas em dólar. Neste caso, esperaríamos um único componente principal importante com um peso alto em X_1. Alternativamente, se ambas as variáveis são padronizadas, suas magnitudes subsequentes serão da mesma ordem, e X_2 (ou Z_2) representarão um papel maior na construção dos componentes principais.

Componentes Principais Amostrais

Agora temos a estrutura necessária para estudar o problema de resumir a variação nas n observações sobre p variáveis em poucas combinações lineares criteriosamente escolhidas.

Suponha que os dados $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ representam n observações independentes de alguma população p-dimensional com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. Estes dados produzem um vetor de médias amostral $\overline{\boldsymbol{x}}$, a matriz de covariâncias amostral $\boldsymbol{S}$ e a matriz de correlações amostral $\boldsymbol{R}$.

O objetivo aqui será construir combinações lineares não correlacionadas a partir das características medidas que levam em conta o máximo da variação na amostra. As combinações não correlacionadas com as maiores variâncias serão chamadas componentes principais amostrais.

Proposição 4: Sejam $\boldsymbol{S}$ a matriz de covariância amostral e $(\widehat{\lambda}_j, \widehat{\boldsymbol{v}}_j)$, $j = 1, 2, \ldots, p$ seus correspondentes pares de autovalores e autovetores, com $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \cdots \geqslant \widehat{\lambda}_p \geqslant 0$.

Seja $\boldsymbol{x}$ uma observação qualquer.

Variância amostral $\widehat{\text{Var}}(\widehat{y}_j) = \widehat{\lambda}_j$, para $j = 1, 2, \ldots, p$.

Covariância amostral $\widehat{\text{Cov}}(\widehat{y}_j, \widehat{y}_k) = 0$, para $j \neq k$ e $j, k = 1, 2, \ldots, p$.

Variância amostral total $\sum_{j=1}^{p} s_{jj} = \sum_{j=1}^{p} \widehat{\lambda}_j$.

Correlação do j-ésimo componente principal amostral com a k-ésima variável

$$r_{\widehat{y}_j, x_k} = \frac{\widehat{v}_{jk}\sqrt{\widehat{\lambda}_j}}{\sqrt{s_{kk}}}, \qquad j, k = 1, 2, \ldots, p.$$

Vamos denotar

$$\widehat{\boldsymbol{\Lambda}} = \text{diag}(\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_p) \quad \text{e} \quad \widehat{\boldsymbol{\Lambda}}^{1/2} = \text{diag}\left(\sqrt{\widehat{\lambda}_1}, \sqrt{\widehat{\lambda}_2}, \ldots, \sqrt{\widehat{\lambda}_p}\right).$$

Exemplo 3: Um censo forneceu informações sobre 5 variáveis socioeconômicas em 14 setores censitários na área de Madison, Wisconsin. Os dados de 61 setores censitários estão disponíveis no arquivo T8-5.dat. As variáveis observadas foram X_1 - população (milhares), X_2 - grau profissional, X_3 - população empregada acima de 16 anos, X_4 - funcionários públicos e X_5 - valor mediano da residência.

PROBLEMA: A variação amostral pode ser resumida por um ou dois componentes principais? (Ver exemplo_07.r)

Vetor de médias amostral das variáveis:

Variável	pop.	prof.	emprego	func. pub.	casa
$\overline{\boldsymbol{x}}^\top$	4,47	3,96	71,42	26,91	1,64

Matriz de covariâncias amostrais das variáveis:

	pop.	prof.	emprego	func. pub.	casa
pop.	3,397	-1,102	4,306	-2,078	0,027
prof.	-1,102	9,673	-1,513	10,953	1,203
emprego	4,306	-1,513	55,626	-28,938	-0,044
func. pub.	-2,078	10,953	-28,938	89,067	0,957
casa	0,027	1,203	-0,044	0,957	0,319

Matriz de correlações amostrais das variáveis:

	pop.	prof.	emprego	func. pub.	casa
pop.	1,00	-0,19	0,31	-0,12	0,03
prof.	-0,19	1,00	-0,07	0,37	0,69
emprego	0,31	-0,07	1,00	-0,41	-0,01
func. pub.	-0,12	0,37	-0,41	1,00	0,18
casa	0,03	0,69	-0,01	0,18	1,00

No `R`, a função `prcomp` do pacote `stats` roda a ACP.

Desvios padrão:

$$\widehat{\boldsymbol{\Lambda}}^{1/2} = \begin{bmatrix} 10,3448 & 6,2986 & 2,8932 & 1,6935 & 0,3933 \end{bmatrix}$$

Rotação:

	$\widehat{\boldsymbol{v}}_1$	$\widehat{\boldsymbol{v}}_2$	$\widehat{\boldsymbol{v}}_3$	$\widehat{\boldsymbol{v}}_4$	$\widehat{\boldsymbol{v}}_5$
pop.	0,0389	-0,0711	0,1879	0,9771	-0,0577
prof.	-0,1053	-0,1298	-0,9610	0,1714	-0,1386
emprego	0,4924	-0,8644	0,0458	-0,0910	0,0050
func. pub.	-0,8631	-0,4803	0,1532	-0,0297	0,0067
casa	-0,0091	-0,0147	-0,1250	0,0817	0,9886

Importância das componentes:

	CP1	CP2	CP3	CP4	CP5
$\sqrt{\hat{\lambda}_j}$	10,345	6,299	2,893	1,693	0,393
Prop. Variância	0,677	0,251	0,053	0,018	0,001
Prop. Acumulada	0,677	0,928	0,981	0,999	1,000

Os dois primeiros componentes principais, juntos, explicam 92,8% da variância amostral total. Assim, podemos dizer que a variação amostral é bem resumida por dois componentes principais e uma redução nos dados de 61 observações de 5 variáveis para 61 observações de dois componentes é razoável. Dados os coeficientes dos vetores que definem os componentes, o primeiro parece ser essencialmente uma diferença ponderada entre as variáveis funcionário público e empregados maiores de 16 anos. O segundo CP parece ser uma soma ponderada dos dois.

Correlações entre os componentes e as variáveis

	X_1	X_2	X_3	X_4	X_5
Y_1	0,218	-0,350	-0,683	-0,946	-0,167
Y_2	-0,243	-0,263	-0,730	-0,321	-0,165

Observação: A resposta ao problema é sim, desde que de fato as variáveis X_3 e X_4 sejam mais importantes. Caso contrário, é melhor padronizar as variáveis.

O número de componentes principais

A questão de quantos componentes reter está sempre presente. Não há uma resposta definitiva para ela. Aspectos a serem considerados incluem a quantidade total da variância explicada, o tamanho relativo dos autovalores (as variâncias dos componentes amostrais), e as interpretações dos componentes. Em adição, como veremos adiante, um componente associado com um autovalor próximo de zero e, assim, considerados sem importância, podem indicar dependências lineares insuspeitas nos dados.

Uma ajuda visual útil para determinar um número apropriado de componentes é um *scree plot*. Com os autovalores ordenados em ordem decrescente, um *scree plot* é um gráfico de j versus $\widehat{\lambda}_j$. Para determinar o número apropriado de componentes, buscamos por um "cotovelo" (dobra) no *scree plot*.

O número de componentes é tomado como o ponto a partir do qual os autovalores restantes são relativamente pequenos e todos têm aproximadamente o mesmo tamanho.

Exemplo 3: (Continuação)

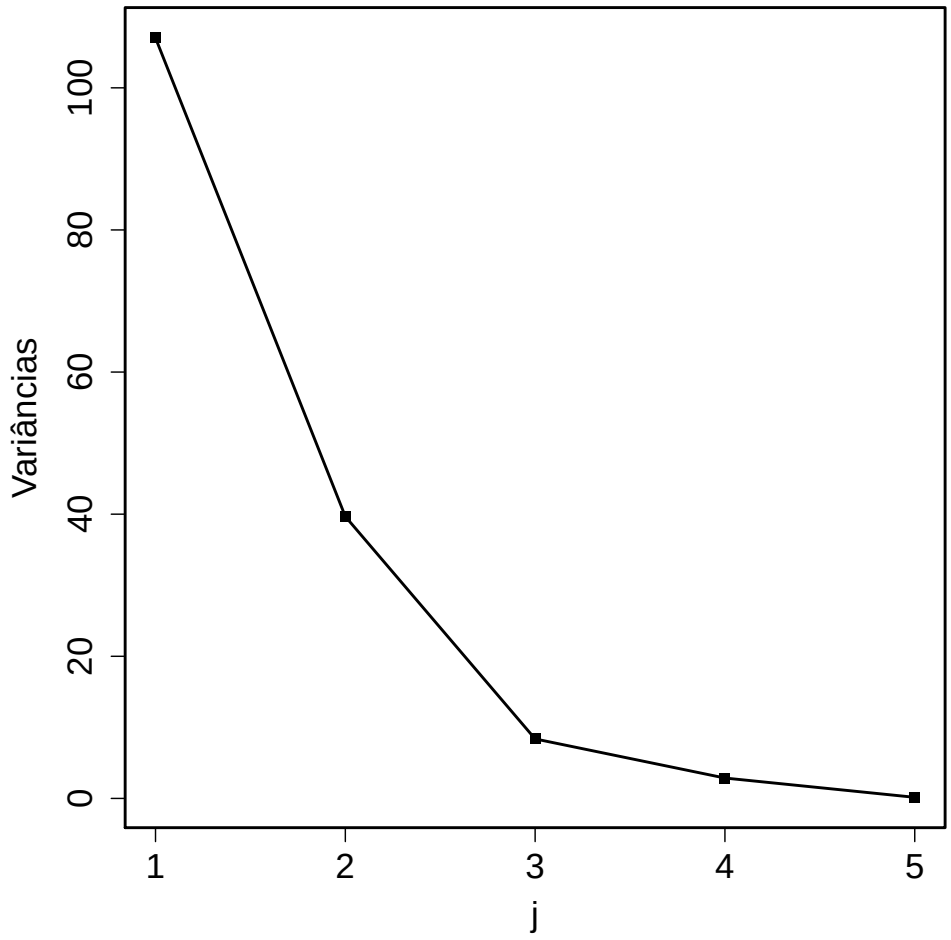

Figura 5: Autovalores ordenados de forma não crescente.

Componentes principais usando as variáveis padronizadas

Os componentes principais também podem ser obtidos via matriz de correlações amostral $\boldsymbol{R}$ e os resultados com suas devidas adaptações serão:

- Seja $(\widehat{\delta}_j, \widehat{\boldsymbol{w}}_j)$, $j = 1, 2, \ldots, p$, os pares de autovalores e autovetores de $\boldsymbol{R}$, com $\widehat{\delta}_1 \geqslant \widehat{\delta}_2 \geqslant \cdots \geqslant \widehat{\delta}_p \geqslant 0$. De maneira similar, temos $\widehat{\boldsymbol{\Delta}}$ e $\widehat{\boldsymbol{\Delta}}^{1/2}$.
- Seja $\boldsymbol{z}$ uma observação qualquer padronizada.
- O valor do j-ésimo CP para esta observação é $\widehat{y}_j^{\star} = \widehat{\boldsymbol{w}}_j^{\top} \boldsymbol{z}$.
- Variância amostral $\widehat{\text{Var}}(\widehat{y}_j^{\star}) = \widehat{\delta}_j, \quad j = 1, 2, \ldots, p.$
- Covariância amostral $\widehat{\text{Cov}}(\widehat{y}_j^{\star}, \hat{y}_k^{\star}) = 0, \quad j \neq k, \quad j, k = 1, 2, \ldots, p.$
- Variância amostral total $\text{tr}(\boldsymbol{R}) = p = \sum_{j=1}^{p} \widehat{\delta}_j$.
- Correlação do j-ésimo componente amostral com a k-ésima variável

$$r_{\widehat{y}_j^{\star}, x_k} = \widehat{w}_{jk} \sqrt{\widehat{\delta}_j}, \quad j, k = 1, 2, \ldots, p.$$

Lembrem-se de que em geral os componentes obtidos via matriz de correlações serão diferentes daqueles obtidos via matriz de covariâncias.

Exemplo 4: (Componentes principais amostrais via dados padronizados)

As taxas semanais de retorno para cinco ações (JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell, and ExxonMobil) listadas na Bolsa de Nova York foram registradas para o período de janeiro de 2004 a dezembro de 2005. As taxas semanais de retorno são definidas como (preço de fechamento da semana corrrente menos preço de fechamento da semana anterior)/(preço de fechamento da semana anterior), ajustada para *stock splits* e dividendos. Os dados estão disponíveis no arquivo T8-4.dat.

As observações sobre 103 semanas sucessivas parecem ser independentemente distribuídas, mas as taxas de retorno das ações são correlacionadas, pois como é de se esperar, ações tendem a mover-se junto em resposta à conjuntura econômica. (Ver exemplo_08.r)

Vetor de médias amostral das variáveis:

Variável	JP	Citi	Fargo	Royal	Exxon
$\overline{\boldsymbol{x}}^{\top}$	0,00106	0,00066	0,00163	0,00405	0,00404

Matriz de correlações amostrais das variáveis:

	JP	Citi	Fargo	Royal	Exxon
JP	1,000	0,632	0,510	0,115	0,154
Citi	0,632	1,000	0,574	0,322	0,213
Fargo	0,510	0,574	1,000	0,182	0,146
Royal	0,115	0,322	0,182	1,000	0,683
Exxon	0,154	0,213	0,146	0,683	1,000

Importância das componentes:

	CP1	CP2	CP3	CP4	CP5
$\sqrt{\hat{\lambda}_j}$	1,5612	1,1862	0,7075	0,63248	0,50514
Prop. Variância	0,4874	0,2814	0,1001	0,08001	0,05103
Prop. Acumulada	0,4874	0,7689	0,8690	0,94897	1,00000

Rotação:

	$\widehat{\boldsymbol{v}}_1$	$\widehat{\boldsymbol{v}}_2$	$\widehat{\boldsymbol{v}}_3$	$\widehat{\boldsymbol{v}}_4$	$\widehat{\boldsymbol{v}}_5$
JP	-0,4691	0,3680	-0,6043	0,3630	0,3841
Citi	-0,5324	0,2365	-0,1361	-0,6292	-0,4962
Fargo	-0,4652	0,3152	0,7718	0,2890	0,0712
Royal	-0,3873	-0,5850	0,0934	-0,3813	0,5947
Exxon	-0,3607	-0,6058	-0,1088	0,4934	-0,4976

- O primeiro componente é uma soma ponderada das cinco ações ou um "índice". Este componente pode ser chamado de componente geral do mercado de ações, ou simplesmente, componente de mercado.
- O segundo componente representa um contraste entre ações de bancos (JP Morgan, Citibank, Wells Fargo) e ações de petróleo (Royal Dutch Shell, ExxonMobil). Ele pode ser chamado de componente industrial.

Assim, vemos que a maior parte da variação nestes retornos de ações (77%) é devida à atividade de mercado e a atividade não correlacionada de indústria.

Os componentes restantes não são fáceis de interpretar e, coletivamente, representam variação que é provavelmente específica a cada ação.

As correlações entre os três primeiros CP's e as variáveis é dada a seguir.

Tabela 2: Correlações entre os componentes e as variáveis.

	X_1	X_2	X_3	X_4	X_5
Y_1	-0,732	-0,831	-0,726	-0.605	-0,563
Y_2	0,437	0,280	0,374	-0,694	-0,719
Y_3	-0,428	-0,096	0,546	0,066	-0,077

A seguir mostramos o *scree plot*.

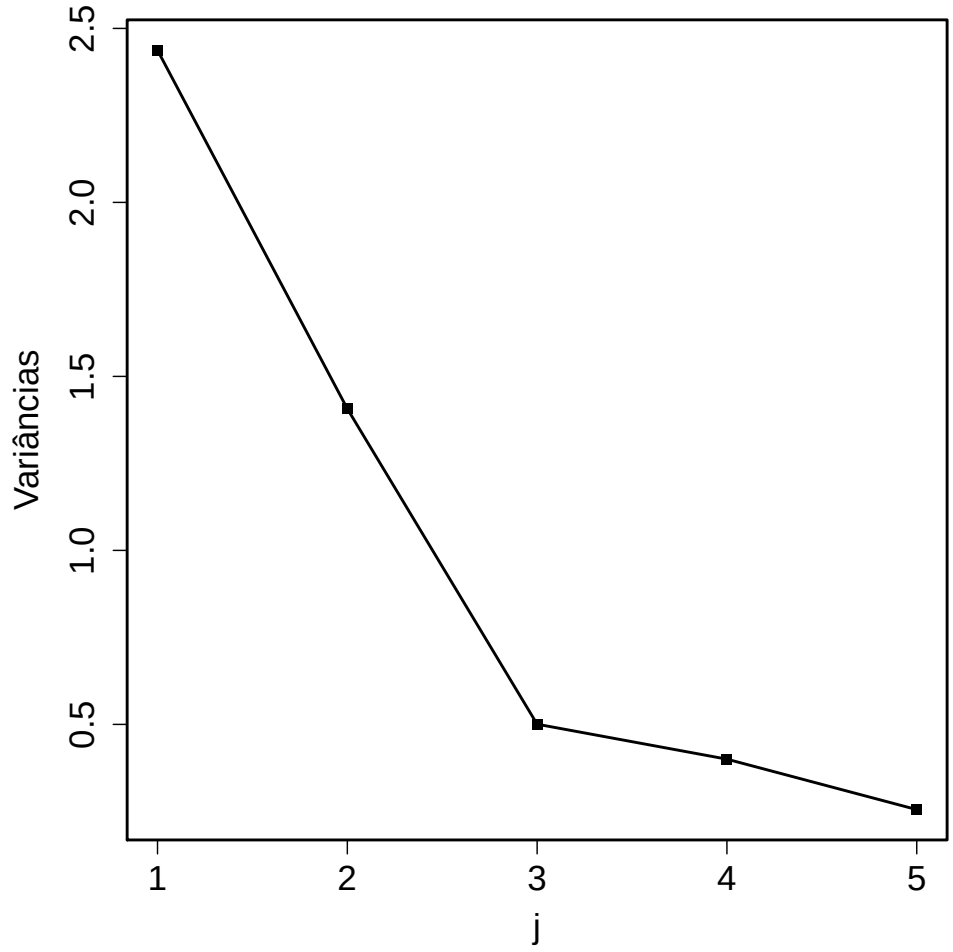

Figura 6: Autovalores ordenados de forma não crescente.

Gráficos dos componentes principais

Gráficos dos CP's podem revelar observações suspeitas bem como desvios da suposição de normalidade. Como os CP's são combinações lineares das variáveis originais, é razoável esperar que eles comportem-se de acordo com uma distribuição normal, se estivermos supondo $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Em geral é necessário verificar que os primeiros poucos CP's são aproximadamente normais, quando eles são usados como dados de entrada em análises posteriores.

Os últimos CP's podem ajudar a identificar observações suspeitas. Cada observação $\boldsymbol{x}_i$ pode ser expressa como uma combinação linear da seguinte forma:

$$\boldsymbol{x}_i = (\boldsymbol{x}_i^\top \widehat{\boldsymbol{v}}_1)\widehat{\boldsymbol{v}}_1 + (\boldsymbol{x}_i^\top \widehat{\boldsymbol{v}}_2)\widehat{\boldsymbol{v}}_2 + \cdots + (\boldsymbol{x}_i^\top \widehat{\boldsymbol{v}}_p)\widehat{\boldsymbol{v}}_p, \quad i = 1, 2, \ldots, n.$$

De fato, $\boldsymbol{Y} = \boldsymbol{P}^\top \boldsymbol{X}$, com $\boldsymbol{P}$ matriz ortogonal cujas colunas são os autovetores ortonormalizados. Logo, pré-multiplicando pela matriz $\boldsymbol{P}$ ambos os lados, obtemos $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{Y}$, ou seja, as variáveis originais também podem, obviamente, ser representadas como combinações lineares dos CP's.

Assim, as magnitudes dos últimos CP's determinarão quão bem os poucos primeiros ajustam as observações. Isto é,

$$\sum_{j=1}^{q-1} \widehat{y}_{ij}\widehat{\boldsymbol{v}}_j \text{ difere de } \boldsymbol{x}_i \text{ por } \sum_{j=q}^{p} \widehat{y}_{ij}\boldsymbol{v}_j.$$

O módulo quadrado deste último vetor é dado por

$$\widehat{y}_{iq}^2 + \widehat{y}_{i(q+1)}^2 + \cdots + \widehat{y}_{ip}^2.$$

Observações suspeitas frequentemente serão tais que pelo menos uma das coordenadas $\widehat{y}_{ij}$, $j = q, q+1, \ldots, p$ é grande.

Recomendações:

1. Para ajudar na verificação da normalidade, construa diagramas de dispersão para os pares dos poucos primeiros CP's que explicam uma proporção alta da variação total. Também construa os *qq-plots* dos valores amostrais gerados por cada componente.
2. Construa os diagramas de dispersão 2 a 2 e os qq-plots para os últimos (poucos) CP's. Eles ajudam a identificar observações suspeitas.

Inferências para grandes amostras

Vimos que a decomposição espectral da matriz de covariâncias (correlações) é a essência da ACP.

Os autovetores determinam as direções de variabilidade máxima e, os autovalores especificam as variâncias.

Quando os primeiros poucos autovalores são muito maiores que os demais, a maior parte da variação pode ser "explicada" por um número menor do que p de variáveis.

Na prática, decisões olhando a qualidade da aproximação dos CP's devem ser tomadas com base nos pares de autovalores e autovetores associados de $\boldsymbol{S}$ ou $\boldsymbol{R}$, $(\widehat{\lambda}_j, \widehat{\boldsymbol{v}}_j)$ ou $(\widehat{\delta}_j, \widehat{\boldsymbol{w}}_j)$, para $j = 1, 2, \ldots, p$.

Devido à variação inerente à amostragem, estes autovalores e autovetores serão diferentes de suas contrapartidas populacionais.

As distribuições amostrais $\widehat{\lambda}_j$ e $\widehat{\boldsymbol{v}}_j$ são difíceis de se obter e estão além dos objetivos desta disciplina.

No entanto, se a suposição de normalidade multivariada é válida para o vetor de observações $\boldsymbol{X}$, podemos estabelecer propriedades sob grandes amostras.

Suponha $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $\boldsymbol{\Sigma}$ positiva definida e tal que $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ são seus autovalores. A única exceção a este caso ocorre quando algum autovalor tem multiplicidade maior que 1.

Geralmente as conclusões para o caso de autovalores distintos são aplicáveis a menos que existam fortes razões para acreditar que $\boldsymbol{\Sigma}$ tenha uma estrutura especial que produza autovalores iguais. Mesmo quando a suposição de normalidade é violada, os intervalos de confiança descritos a seguir ainda fornecerão alguma indicação da incerteza sobre $\widehat{\lambda}_j$ e componentes de $\widehat{\boldsymbol{v}}_j$.

Suponha $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Seja $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$, com λ_j, $j = 1, 2, \ldots, p$, os autovalores de $\boldsymbol{\Sigma}$. Faça $\boldsymbol{\lambda}^\top = (\lambda_1, \lambda_2, \ldots, \lambda_p)$. Defina

$$\boldsymbol{V}_j = \lambda_j \sum_{i=1, i \neq j}^{p} \frac{\lambda_i}{(\lambda_i - \lambda_j)^2} \boldsymbol{v}_i \boldsymbol{v}_i^\top .$$

Então

(P1) $\sqrt{n}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \overset{a}{\sim} \mathcal{N}_p(\boldsymbol{0}, 2\boldsymbol{\Lambda}^2)$.

(P2) $\sqrt{n}(\widehat{\boldsymbol{v}}_j - \boldsymbol{v}_j) \overset{a}{\sim} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{V}_j)$.

(P3) Cada $\widehat{\lambda}_j$ é independentemente distribuído dos elementos de seu autovetor $\widehat{\boldsymbol{v}}_j$.

Logo, (P1) implica que para n grande os $\widehat{\lambda}_j$'s são independentemente distribuídos, isto é

$$\widehat{\lambda}_j \overset{ind,a}{\sim} \mathcal{N}(\lambda_j, 2\lambda_j^2).$$

Usando esta distribuição, podemos construir intervalos de confiança assintóticos 100(1-α)% para os λ_j, a saber,

$$P\left(|\widehat{\lambda}_j - \lambda_j| \leq z_{(1-\alpha/2)}\lambda_j\sqrt{\frac{2}{n}}\right) \simeq 1-\alpha$$

tal que

$$\text{IC}_{100(1-\alpha)\%}(\lambda_j) = \left(\frac{\widehat{\lambda}_j}{1+z_{(1-\alpha/2)}\sqrt{\frac{2}{n}}}, \frac{\widehat{\lambda}_j}{1-z_{(1-\alpha/2)}\sqrt{\frac{2}{n}}}\right).$$

Intervalos simultâneos assintóticos para os λ_j's via Bonferroni também podem ser construídos. Suponha que estejamos interessados em $m < p$ intervalos. Então, basta substituir o quantil acumulado da distribuição normal padrão $z_{(1-\alpha/2)}$ por $z_{(1-\alpha/(2m))}$.

(P2) implica que os $\widehat{\boldsymbol{v}}_j$ são normalmente distribuídos para n grande.

Os elementos de cada $\widehat{\boldsymbol{v}}_j$ são correlacionados e a correlação depende em grande extensão da separação dos autovalores $\lambda_1, \lambda_2, \ldots, \lambda_p$ (que é desconhecida) e do tamanho amostral n.

Erros padrão aproximados para os coeficientes $\widehat{v}_{kj}$ são dados pelas raízes quadradas dos elementos diagonais da matriz $\frac{1}{n}\widehat{\boldsymbol{E}}_j$ em que os elementos de $\widehat{\boldsymbol{E}}_j$ são obtidos a partir de $\boldsymbol{E}_j$ substituindo suas entradas por suas estimativas $\widehat{\lambda}_j$ e $\widehat{\boldsymbol{v}}_j$.

Exemplo 4: (Continuação) Voltando ao exemplo das ações pedem-se intervalos de confiança assintóticos simultâneos de 95% para os dois primeiros autovalores. Usar Bonferroni.

Neste caso $\alpha = 0,05$, $m = 2$, implicando $1 - \dfrac{\alpha}{2m} = 0,9875$ e $z_{(0,9875)} \simeq 2,24$.

Temos também $n = 103$, $1 + z_{\left(1-\dfrac{\alpha}{2m}\right)}\sqrt{\frac{2}{n}} \simeq 1,31$ e $1 + z_{(1-\frac{\alpha}{2m})}\sqrt{\frac{2}{n}} \simeq 0,69$.

Além disso, $\widehat{\lambda}_1 \simeq 2,44$ e $\widehat{\lambda}_2 \simeq 1,41$.

Assim, os IC's simultâneos são dados por

$$\begin{aligned} IC_{95\%}(\lambda_1) &= \left(\frac{2,44}{1,31}; \frac{2,44}{0,69}\right) \simeq (1,86;\ 3,54) \\ IC_{95\%}(\lambda_2) &= \left(\frac{1,41}{1,31}; \frac{1,41}{0,69}\right) \simeq (1,08;\ 2,04) \end{aligned}$$

Análise de Correlação Canônica - ACC

A ACC visa identificar e quantificar as associações entre dois conjuntos de variáveis.

A ACC foca na correlação entre uma combinação linear das variáveis de um conjunto e uma combinação linear das variáveis em outro conjunto.

Primeiro determina-se o par de combinações lineares tendo a maior correlação.

Depois, determina-se o par de combinações lineares tendo a maior correlação dentre todos os pares não correlacionados com o par inicial, e assim por diante.

Os pares de combinações lineares são chamados de variáveis canônicas e suas correlações de correlações canônicas.

Suponha dois grupos de variáveis:

$$\boldsymbol{X}^{(1)}\ (p \times 1) \quad \text{e} \quad \boldsymbol{X}^{(2)}\ (q \times 1) \quad \text{com} \quad p \leqslant q,$$

tal que

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}^{(1)}) &= \boldsymbol{\mu}^{(1)}, & \mathrm{E}(\boldsymbol{X}^{(2)}) &= \boldsymbol{\mu}^{(2)}, \\ \mathrm{Var}(\boldsymbol{X}^{(1)}) &= \boldsymbol{\Sigma}_{11}, & \mathrm{Var}(\boldsymbol{X}^{(2)}) &= \boldsymbol{\Sigma}_{22}, \quad \text{e;} \\ \mathrm{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) &= & \boldsymbol{\Sigma}_{12} &= \boldsymbol{\Sigma}_{21}^{\top}. \end{aligned}$$

Então, podemos organizar as variáveis da seguinte forma:

$$\begin{aligned} \boldsymbol{X} &= \begin{bmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{bmatrix}, \\ \mathrm{E}(\boldsymbol{X}) &= \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}, \quad \text{e} \\ \mathrm{Cov}(\boldsymbol{X}) &= \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \end{aligned}$$

A ACC visa resumir as associações entre os conjuntos de variáveis $\boldsymbol{X}^{(1)}$ e $\boldsymbol{X}^{(1)}$ em termos de poucas covariâncias (correlações) escolhidas com cuidado.

Sejam $U = \boldsymbol{a}^\top \boldsymbol{X}^{(1)}$ e $V = \boldsymbol{b}^\top \boldsymbol{X}^{(2)}$ tal que

$$\begin{aligned} \text{Var}(U) &= \boldsymbol{a}^\top \text{Cov}(\boldsymbol{X}^{(1)})\boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{\Sigma}_{11}\boldsymbol{a}, \\ \text{Var}(V) &= \boldsymbol{b}^\top \text{Cov}(\boldsymbol{X}^{(2)})\boldsymbol{b} = \boldsymbol{b}^\top \boldsymbol{\Sigma}_{22}\boldsymbol{b}, \quad \text{e} \\ \text{Cov}(U, V) &= \boldsymbol{a}^\top \text{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})\boldsymbol{b} = \boldsymbol{a}^\top \boldsymbol{\Sigma}_{12}\boldsymbol{b}. \end{aligned}$$

O interesse está nos valores de $\boldsymbol{a}$ e $\boldsymbol{b}$ que maximizem

$$\text{Cor}(U, V) = \frac{\boldsymbol{a}^\top \boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^\top \boldsymbol{\Sigma}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}^\top \boldsymbol{\Sigma}_{22}\boldsymbol{b}}}.$$

Definições

- O primeiro par de variáveis canônicas é o par de combinações lineares U_1 e V_1 tendo variâncias unitárias e que maximiza Cor(U, V).
- O segundo par de variáveis canônicas é o par de combinações lineares U_2 e V_2 tendo variâncias unitárias e que maximiza Cor(U, V) dentre todas as escolhas não correlacionadas com U_1 e V_1.
- O raciocínio prosegue.

Suponha que $\boldsymbol{\Sigma}$ tenha posto completo. Então,

$$\text{máx}_{\boldsymbol{a},\boldsymbol{b}}\text{Cor}(U, V) = \rho_1^\star$$

é dado por

$$U_1 = \boldsymbol{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{X}^{(1)} \quad \text{e} V_1 = \boldsymbol{f}_1^\top \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{X}^{(2)}$$

em que $\boldsymbol{a}_1^\top = \boldsymbol{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}$ e $\boldsymbol{b}_1^\top = \boldsymbol{f}_1^\top \boldsymbol{\Sigma}_{22}^{-1/2}$.

O k-ésimo par de variáveis canônicas, $k = 1, 2, \ldots, p$,

$$U_k = \boldsymbol{e}_k^\top \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{X}^{(1)} \quad \text{e} V_k = \boldsymbol{f}_k^\top \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{X}^{(2)}$$

maximiza

$$\text{Cor}(U_k, V_k) = \rho_k^\star.$$

Seja $\boldsymbol{A}$ uma matriz simétrica e positiva definida. Então, pela decomposição do valor singular (*SVD*) temos que $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top$. Logo, $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^\top$ chama-se raiz quadrada de $\boldsymbol{A}$ e consequentemente $\boldsymbol{A}^{-1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}^\top$.

Vale ressaltar que $\rho_1^{\star 2} \geqslant \rho_2^{\star 2} \geqslant \cdots \geqslant \rho_p^{\star 2}$ são os autovalores de $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ e $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ os autovetores associados.

Tem-se que $\rho_1^{\star 2} \geqslant \rho_2^{\star 2} \geqslant \cdots \geqslant \rho_p^{\star 2}$ são também os p maiores autovalores de $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$ e $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_p$ os autovetores associados. Cada $\boldsymbol{f}_i$ é proporcional a $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{e}_i$.

Tem-se também que

$$\begin{aligned}
\text{Var}(U_k) &= \text{Var}(V_k) = 1 \\
\text{Cov}(U_k, U_\ell) &= \text{Cor}(U_k, U_\ell) = 0, \quad k \neq \ell \\
\text{Cov}(V_k, V_\ell) &= \text{Cor}(V_k, V_\ell) = 0, \quad k \neq \ell \\
\text{Cov}(U_k, V_\ell) &= \text{Cor}(U_k, V_\ell) = 0, \quad k \neq \ell,
\end{aligned}$$

para $k, \ell = 1, \ldots, p$.

Se as variáveis $\boldsymbol{X}^{(1)}$ e $\boldsymbol{X}^{(2)}$ forem padronizadas em $\boldsymbol{Z}^{(1)} = \boldsymbol{V}^{-1/2}(\boldsymbol{X}^{(1)} - \boldsymbol{\mu}^{(1)})$ e $\boldsymbol{Z}^{(2)} = \boldsymbol{V}^{-1/2}(\boldsymbol{X}^{(2)} - \boldsymbol{\mu}^{(2)})$, então a correlação canônica não se altera.

Além disso, se $\boldsymbol{a}_k$ é o vetor de coeficientes para a k-ésima variável canônica, então $\boldsymbol{a}_k^\top \boldsymbol{V}^{1/2}$ é o vetor de coeficientes da k-ésima variável canônica construída das variáveis padronizadas $\boldsymbol{Z}^{(1)}$.

Exemplo

Suponha que $\boldsymbol{Z}^{(1)} = (Z_1^{(1)}, Z_2^{(1)})^\top$ e $\boldsymbol{Z}^{(2)} = (Z_1^{(2)}, Z_2^{(2)})^\top$ sejam variáveis padronizadas tal que $\boldsymbol{Z} = (\boldsymbol{Z}^{(1)\top}, \boldsymbol{Z}^{(2)\top})^\top$,

$$\text{Cor}(\boldsymbol{Z}) = \left[\begin{array}{c:c} \mathcal{R}_{11} & \mathcal{R}_{12} \\ \hdashline \mathcal{R}_{21} & \mathcal{R}_{22} \end{array}\right] = \left[\begin{array}{cc:cc} 1,0 & 0,4 & 0,5 & 0,6 \\ 0,4 & 1,0 & 0,3 & 0,4 \\ \hdashline 0,5 & 0,3 & 1,0 & 0,2 \\ 0,6 & 0,4 & 0,2 & 1,0 \end{array}\right].$$

Então,

$$\mathcal{R}_{11}^{-1/2} = \begin{bmatrix} 1,0681 & -0,2229 \\ -0,2229 & 1,0681 \end{bmatrix}, \quad \mathcal{R}_{22}^{-1} = \begin{bmatrix} 1,0417 & -0,2083 \\ -0,2083 & 1,0417 \end{bmatrix} \quad \text{e}$$

$$\mathcal{R}_{11}^{-1/2}\mathcal{R}_{12}\mathcal{R}_{22}^{-1}\mathcal{R}_{21}\mathcal{R}_{11}^{-1/2} = \begin{bmatrix} 0,4370 & 0,2178 \\ 0,2178 & 0,1097 \end{bmatrix}.$$

Para $\mathcal{R}_{11}^{-1/2}\mathcal{R}_{12}\mathcal{R}_{22}^{-1}\mathcal{R}_{21}\mathcal{R}_{11}^{-1/2}$, tem-se que os autovalores são dados por $\rho_1^{\star 2} = 0,54572$ e $\rho_1^{\star 2} = 0,00091$, e o primeiro autovetor por $\boldsymbol{e}_1 = (0,8947;\ 0,4468)^\top$.

Assim, $\boldsymbol{a}_1 = \mathcal{R}_{11}^{-1/2}\boldsymbol{e}_1 = (0,8560;\ 0,2777)^\top$.

Tem-se que $\boldsymbol{f}_1 \propto \mathcal{R}_{22}^{-1/2}\mathcal{R}_{21}\mathcal{R}_{11}^{-1/2}\boldsymbol{e}_1$ e $\boldsymbol{b}_1 = \mathcal{R}_{22}^{-1/2}\boldsymbol{f}_1$. Logo,

$$\boldsymbol{b}_1 \propto \mathcal{R}_{22}^{-1}\mathcal{R}_{21}\boldsymbol{a}_1 = (0,4025;\ 0,5442)^\top.$$

Agora, devemos padronizar $\boldsymbol{b}_1$ para que

$$\text{Var}(V_1) = \text{Var}(\boldsymbol{b}_1^\top \boldsymbol{Z}^{(2)}) = \boldsymbol{b}_1^\top \mathcal{R}_{22}\boldsymbol{b}_1 = 1.$$

Assim,

$$\begin{bmatrix}0,4025 & 0,5442\end{bmatrix}\begin{bmatrix}1,0 & 0,2\\ 0,2 & 1,0\end{bmatrix}\begin{bmatrix}0,4025\\ 0,5442\end{bmatrix} = 0,5457.$$

Usando $\sqrt{0,5457} = 0,7387$, tem-se que

$$\boldsymbol{b}_1 = \frac{1}{0,7387}\begin{bmatrix}0,4025\\ 0,5442\end{bmatrix} = \begin{bmatrix}0,5448\\ 0,7366\end{bmatrix}.$$

O primeiro par de variáveis canônicas é dador por

$$\begin{aligned} U_1 &= \boldsymbol{a}_1^\top \boldsymbol{Z}^{(1)} = 0,8560 Z_1^{(1)} + 0,2777 Z_2^{(1)} \\ V_1 &= \boldsymbol{b}_1^\top \boldsymbol{Z}^{(2)} = 0,5448 Z_1^{(2)} + 0,7366 Z_2^{(2)} \end{aligned}$$

e a correlação canônica

$$\rho_1^\star = \sqrt{\rho_1^{\star 2}} = \sqrt{0,54572} = 0,7387.$$

Esta é a maior correlação possível entre combinações linear dos conjuntos $\boldsymbol{Z}^{(1)}$ e $\boldsymbol{Z}^{(2)}$.

A segunda correlação canônica, $\rho_2^\star = \sqrt{0,00092} = 0,03$, é muito pequena (não tem informação relevante).

Variáveis Canônicas Populacionais

- É melhor trabalhar com variáveis padronizadas para evitar problemas de interpretação das variáveis canônica.
- Variáveis padronizadas são livres de escala.
- As variáveis canônicas podem ser “identificadas” em termos de variáveis por assunto.
- O cálculo da correlação entre as variáveis canônicas e as originais auxiliam o trabalho de interpretação (mas é necessário cuidade devido à unidimensionalidade).

Sejam $\boldsymbol{A}_{(p\times p)} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p)^\top$ e $\boldsymbol{B}_{(q\times q)} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_q)^\top$ tal que

$$\boldsymbol{U} = \boldsymbol{A}\boldsymbol{X}^{(1)} \quad \text{e} \quad \boldsymbol{V} = \boldsymbol{B}\boldsymbol{X}^{(2)}$$

com foco nas p primeiras variáveis canônicas de $\boldsymbol{V}$. Logo,

$$\begin{aligned}
\text{Cov}(\boldsymbol{U}, \boldsymbol{X}^{(1)}) &= \text{Cov}(\boldsymbol{A}\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(1)}) = \boldsymbol{A}\boldsymbol{\Sigma}_{11} \\
\text{Cor}(\boldsymbol{U}, \boldsymbol{X}^{(1)}) &= \text{Cov}(\boldsymbol{U}, \boldsymbol{V}_{11}^{-1/2}\boldsymbol{X}^{(1)}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}^{(1)}, \boldsymbol{V}_{11}^{-1/2}\boldsymbol{X}^{(1)}) = \boldsymbol{A}\boldsymbol{\Sigma}_{11}\boldsymbol{V}_{11}^{-1/2},
\end{aligned}$$

$$\begin{aligned}
\text{Cor}(\boldsymbol{U}, \boldsymbol{X}^{(2)}) &= \boldsymbol{A}\boldsymbol{\Sigma}_{12}\boldsymbol{V}_{22}^{-1/2}, \\
\text{Cor}(\boldsymbol{V}, \boldsymbol{X}^{(1)}) &= \boldsymbol{B}\boldsymbol{\Sigma}_{21}\boldsymbol{V}_{11}^{-1/2}, \quad \text{e} \\
\text{Cor}(\boldsymbol{V}, \boldsymbol{X}^{(2)}) &= \boldsymbol{B}\boldsymbol{\Sigma}_{22}\boldsymbol{V}_{22}^{-1/2}.
\end{aligned}$$

No caso de variáveis padronizadas $\boldsymbol{Z}^{(1)}$ e $\boldsymbol{Z}^{(2)}$, tem-se que

$$\begin{aligned}
\text{Cor}(\boldsymbol{U}, \boldsymbol{Z}^{(1)}) &= \boldsymbol{A}_{\boldsymbol{Z}}\mathcal{R}_{11}, & \text{Cor}(\boldsymbol{U}, \boldsymbol{Z}^{(2)}) &= \boldsymbol{A}_{\boldsymbol{Z}}\mathcal{R}_{12}, \\
\text{Cor}(\boldsymbol{V}, \boldsymbol{Z}^{(1)}) &= \boldsymbol{B}_{\boldsymbol{Z}}\mathcal{R}_{21} \quad \text{e} & \text{Cor}(\boldsymbol{V}, \boldsymbol{Z}^{(2)}) &= \boldsymbol{B}_{\boldsymbol{Z}}\mathcal{R}_{22}.
\end{aligned}$$

Vale ressaltar que a padronização não afeta a correlação. Portanto, $\text{Cor}(\boldsymbol{U}, \boldsymbol{X}^{(1)}) = \text{Cor}(\boldsymbol{U}, \boldsymbol{Z}^{(1)})$, e assim por diante.

Exemplo (Continuação)

Tem-se que

$$\mathcal{R}_{11} = \begin{bmatrix} 1,0 & 0,4 \\ 0,4 & 1,0 \end{bmatrix}, \quad \mathcal{R}_{22} = \begin{bmatrix} 1,0 & 0,2 \\ 0,2 & 1,0 \end{bmatrix} \quad \text{e} \quad \mathcal{R}_{12} = \begin{bmatrix} 0,5 & 0,6 \\ 0,3 & 0,4 \end{bmatrix}.$$

Com $p = 1$, $\boldsymbol{A_Z} = \begin{bmatrix} 0,8560 & 0,2777 \end{bmatrix}$ e $\boldsymbol{B_Z} = \begin{bmatrix} 0,5448 & 0,7366 \end{bmatrix}$ tal que

$$\text{Cor}(U_1, \boldsymbol{Z}^{(1)}) = \begin{bmatrix} 0,8560 & 0,2777 \end{bmatrix} \begin{bmatrix} 1,0 & 0,4 \\ 0,4 & 1,0 \end{bmatrix} = \begin{bmatrix} 0,9671 & 0,6201 \end{bmatrix} \quad \text{e}$$

$$\text{Cor}(V_1, \boldsymbol{Z}^{(2)}) = \begin{bmatrix} 0,5448 & 0,7366 \end{bmatrix} \begin{bmatrix} 1,0 & 0,2 \\ 0,2 & 1,0 \end{bmatrix} = \begin{bmatrix} 0,6921 & 0,8456 \end{bmatrix}.$$

- A primeira variável em $\boldsymbol{Z}^{(1)}$ é mais relacionada a variável canônica U_1.
- A segunda variável em $\boldsymbol{Z}^{(2)}$ é mais relacionada a variável canônica V_1.
- Em ambos os casos, as correlações reinforçam a informação dada em $\boldsymbol{A_Z}$ e $\boldsymbol{B_Z}$.
- Note que as correlações aumentam a importância de $Z_2^{(1)}$ em $\boldsymbol{Z}^{(1)}$, e de $Z_1^{(2)}$ em $\boldsymbol{Z}^{(2)}$ porque ignoram a contribuição das variáveis restantes em cada grupo.

Variáveis Canônicas Amostrais e Correlações Canônicas Amostrais

Sejam $\boldsymbol{X}^{(1)}$ e $\boldsymbol{X}^{(2)}$ matrizes de amostras aleatórias de dois grupos de variáveis tal que

$$\overline{\boldsymbol{x}} = \begin{bmatrix} \overline{\boldsymbol{x}}^{(1)} \\ \hdashline \overline{\boldsymbol{x}}^{(2)} \end{bmatrix} \quad \text{e} \quad \boldsymbol{S} = \left[\begin{array}{c:c} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \hdashline \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{array}\right].$$

As combinações lineares

$$\widehat{U} = \widehat{\boldsymbol{a}}^{\top} \boldsymbol{x}^{(1)} \quad \text{e} \quad \widehat{V} = \widehat{\boldsymbol{b}}^{\top} \boldsymbol{x}^{(2)}$$

têm correlação amostral dada por

$$\text{Cor}(\widehat{U}, \widehat{V}) = \frac{\widehat{\boldsymbol{a}}^{\top} \boldsymbol{S}_{12} \widehat{\boldsymbol{b}}^{\top}}{\sqrt{\widehat{\boldsymbol{a}}^{\top} \boldsymbol{S}_{11} \widehat{\boldsymbol{a}}^{\top}} \sqrt{\widehat{\boldsymbol{b}}^{\top} \boldsymbol{S}_{22} \widehat{\boldsymbol{b}}^{\top}}}.$$

O primeiro par de variáveis canônicas amostrais é o par de combinações lineares $\widehat{U}$ e $\widehat{V}$ que maximiza a correlação acima.

Sejam $\widehat{\rho}_1^{\star 2} \geqslant \widehat{\rho}_2^{\star 2} \geqslant \cdots \geqslant \widehat{\rho}_p^{\star 2}$ e $\widehat{\boldsymbol{e}}_1, \widehat{\boldsymbol{e}}_2, \ldots, \widehat{\boldsymbol{e}}_p$, os autovalores e autovetores, respectivamente, de $\boldsymbol{S}_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1/2}$ com $p \leqslant q$.

Sejam $\widehat{\boldsymbol{f}}_1, \widehat{\boldsymbol{f}}_2, \ldots, \widehat{\boldsymbol{f}}_p$, autovetores de $\boldsymbol{S}_{22}^{-1/2}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1/2}$ podendo ser calculado como

$$\widehat{\boldsymbol{f}}_k = (1/\widehat{\rho}_k^{\star 2})\boldsymbol{S}_{22}^{-1/2}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\widehat{\boldsymbol{e}}_k, \quad k = 1, 2, \ldots, p.$$

Então, o k-ésimo par de variáveis canônicas é dado por

$$\widehat{U}_k = \widehat{e}_k^{\top}\boldsymbol{S}_{11}^{-1/2}\boldsymbol{x}^{(1)} \quad \text{e} \quad \widehat{V}_k = \widehat{f}_k^{\top}\boldsymbol{S}_{22}^{-1/2}\boldsymbol{x}^{(2)}$$

com a k-ésima correlação canônica amostral dada por

$$\text{Cor}(\widehat{U}_k, \widehat{V}_k) = \widehat{\rho}_k^{\star}.$$

As variáveis canônicas amostrais têm variâncias amostrais unitárias, e as correlações amostrais entre elas é zero:

$$s_{\widehat{U}_k,\widehat{U}_k} = s_{\widehat{V}_k,\widehat{V}_k} = 1 \quad \text{e} \quad s_{\widehat{U}_k,\widehat{U}_\ell} = s_{\widehat{V}_k,\widehat{V}_\ell} = s_{\widehat{U}_k,\widehat{V}_\ell} = 0.$$

Sejam $\widehat{\boldsymbol{A}}_{(p\times p)} = (\widehat{\boldsymbol{a}}_1, \widehat{\boldsymbol{a}}_2, \ldots, \widehat{\boldsymbol{a}}_p)^\top$ e $\widehat{\boldsymbol{B}}_{(q\times q)} = (\widehat{\boldsymbol{b}}_1, \widehat{\boldsymbol{b}}_2, \ldots, \widehat{\boldsymbol{a}}_q)^\top$ tal que

$$\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{A}}\boldsymbol{x}^{(1)} \quad \text{e} \quad \widehat{\boldsymbol{V}} = \widehat{\boldsymbol{B}}\boldsymbol{x}^{(2)}.$$

$$\begin{aligned}
\text{Cor}(\widehat{\boldsymbol{U}}, \boldsymbol{x}^{(1)}) &= \widehat{\boldsymbol{A}}\boldsymbol{S}_{11}\boldsymbol{D}_{11}^{-1/2}, \\
\text{Cor}(\widehat{\boldsymbol{U}}, \boldsymbol{x}^{(2)}) &= \widehat{\boldsymbol{A}}\boldsymbol{S}_{12}\boldsymbol{D}_{22}^{-1/2}, \\
\text{Cor}(\widehat{\boldsymbol{V}}, \boldsymbol{x}^{(1)}) &= \widehat{\boldsymbol{B}}\boldsymbol{S}_{21}\boldsymbol{D}_{11}^{-1/2}, \quad \text{e} \\
\text{Cor}(\widehat{\boldsymbol{V}}, \boldsymbol{x}^{(2)}) &= \widehat{\boldsymbol{B}}\boldsymbol{S}_{22}\boldsymbol{D}_{22}^{-1/2},
\end{aligned}$$

em que $\boldsymbol{D}_{11} = \text{diag}(\boldsymbol{S}_{11})$ e $\boldsymbol{D}_{22} = \text{diag}(\boldsymbol{S}_{22})$.

No caso de variáveis padronizadas $\boldsymbol{z}^{(1)}$ e $\boldsymbol{z}^{(2)}$, tem-se que

$$\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{A}}_{\boldsymbol{z}}\boldsymbol{z}^{(1)} \quad \text{e} \quad \widehat{\boldsymbol{V}} = \widehat{\boldsymbol{B}}_{\boldsymbol{z}}\boldsymbol{z}^{(2)}$$

com $\widehat{\boldsymbol{A}}_{\boldsymbol{z}} = \widehat{\boldsymbol{A}}\boldsymbol{D}_{11}^{1/2}$ e $\widehat{\boldsymbol{B}}_{\boldsymbol{z}} = \widehat{\boldsymbol{B}}\boldsymbol{D}_{22}^{1/2}$.

(Ver exemplo_10.r)

Técnicas multivariadas que dizem respeito à “separação” de conjuntos distintos de objetos (ou observações) e à alocação de novos objetos (observações) a grupos previamente definidos.

Podemos enumerar os principais objetivos aqui como:

Discriminação: Descrever grafica e algebricamente os aspectos que diferenciam os grupos de objetos (observações). Determinar “discriminantes” entre grupos.

Classificação: Alocar objetos em classes previamente definidas. A ênfase aqui está na construção de uma regra que pode ser usada para designar de forma ótima um novo objeto às classes existentes.

Exemplo 1: (Diagnóstico médico) Suponha que dispõe-se de uma amostra de n fichas de pacientes para os quais foram registrados p sintomas que podem ser representados por um vetor $\boldsymbol{x}$ e cujo diagnóstico foi uma entre k doenças possíveis. Um novo paciente apresenta vetor de sintomas $\boldsymbol{x}_0$. Como utilizar a informação amostral para diagnosticar a doença do novo paciente?

Uma função que separa objetos pode servir algumas vezes como "alocadora" e, reciprocamente, uma regra que aloca objetos pode sugerir um procedimento de discriminação.

Na prática, os dois principais objetivos se sobrepõem e a distinção entre separação e alocação fica obscurecida.

Separação e Classificação para o Caso de Duas Populações

Sejam π_1 e π_2 as duas populações.

Os objetos são separados ou classificados com base em p medidas $\boldsymbol{X}^\top = (X_1, X_2, \ldots, X_p)$. Os valores observados $\boldsymbol{x}$ diferem de alguma forma de uma população para outra.

Se $\boldsymbol{x}$ vem da população π_1 dizemos que a distribuição caracterizada pela densidade de probabilidade conjunta de $\boldsymbol{X}$ é dada por $f_1(\boldsymbol{x})$, caso contrário, a densidade é dada por $f_2(\boldsymbol{x})$.

As regras de classificação costumam ser desenvolvidas a partir de amostras de aprendizagem: amostras para as quais a classificação de todos os elementos é conhecida.

Essencialmente, o conjunto de todos os resultados amostrais é dividido em duas regiões complementares, R_1 e R_2 tal que se uma nova observação cair em R_1 ela será classificada em π_1 e, caso contrário, será classificada em π_2.

Regras de classificação não fornecem métodos livres de erro. Muitas vezes não é clara a distinção entre as medidas observadas de cada população: os grupos podem se sobrepor de alguma forma. Logo, será possível classificar um objeto de π_2 em π_1 e vice-versa.

Um bom procedimento de classificação deve resultar numa taxa de erro de classificação pequena.

Probabilidades a Priori

Pode ocorrer que uma população tenha verossimilhança maior do que a outra porque uma população é muito maior. A regra de classificação deve levar em conta estas "probabilidades" a priori de cada população.

Notação: Sejam p_j, $j = 1, 2$ tal que $p_j > 0$, $j = 1, 2$ e $p_1 + p_2 = 1$ tais probabilidades.

Custos de Classificação Incorreta

Também pode ocorrer que classificar um objeto de π_1 em π_2 represente um erro muito mais sério do que o recíproco. A regra de classificação deve levar em conta os custos de classificação incorreta.

Seja Ω o espaço amostral - isto é, a coleção de todos os valores possíveis do vetor $\boldsymbol{x}$. Sejam $R_1 \subset \Omega$ o conjunto de valores para o qual classificamos o objeto com sendo de π_1 e, $R_2 = \Omega \setminus R_1$ o conjunto de valores para o qual classificamos o objeto como sendo de π_2.

Se $p = 2$, podemos representar graficamente esta situação.

A probabilidade condicional $\Pr(2|1)$ de classificar um objeto de π_1 em π_2 é

$$\Pr(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} = \Pr(2|1).$$

Similarmente, a probabilidade condicional $\Pr(1|2)$ de classificar um objeto de π_2 em π_1 é

$$\Pr(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} = \Pr(1|2).$$

Desse modo, a probabilidade global de classificação incorreta pode ser obtida como a soma do produto das probabilidades condicionais por suas probabilidades a priori:

P(classificar incorretamente em π_1) $= \Pr(\boldsymbol{X} \in R_1|\pi_2)\Pr(\pi_2) = \Pr(1|2)p_2$.

P(classificar incorretamente em π_2) $= \Pr(\boldsymbol{X} \in R_2|\pi_1)\Pr(\pi_1) = \Pr(2|1)p_1$.

Os esquemas de classificação costumam ser avaliados em função de suas probabilidades de classificação incorreta, mas observe que estas probabilidades ignoram os custos de classificação incorreta.

Suponha a seguinte tabela de custos de classificação:

População Real	Classificado em π_1	Classificado em π_2
π_1	0	$C(2\|1)$
π_2	$C(1\|2)$	0

Para qualquer regra, o custo esperado de classificação incorreta (CECI) é dado por

$$\text{CECI} = C(2|1)\Pr(2|1)p_1 + C(1|2)\Pr(1|2)p_2.$$

Uma regra de classificação razoável deve ter um CECI tão pequeno quanto possível.

Proposição 1: As regiões R_1 e R_2 que minimizam o CECI são definidas pelos valores de $\boldsymbol{x}$ para os quais valem:

$$R_1 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqslant \frac{C(1|2)}{C(2|1)} \frac{p_2}{p_1}.$$

(Razão de densidades) $\geqslant$ (Razão de custos) $\times$ (Razão de probabilidades a priori).

$$R_2 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{C(1|2)}{C(2|1)} \frac{p_2}{p_1}.$$

A implementação dessa regra requer o conhecimento da razão de densidades para uma nova observação $\boldsymbol{x}_0$, da razão de custos e da razão de probabilidades a priori. É, em geral, mais simples atribuir valores para as razões do lado direito da desigualdade acima do que atribuir um valor para cada probabilidade a priori e custo de classificação incorreta.

Casos especiais da regra estabelecida pela Proposição 1:

(1a) probabilidades a priori iguais,

$$R_1 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqslant \frac{C(1|2)}{C(2|1)};$$

(1b) custos de classificação incorreta iguais,

$$R_1 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqslant \frac{p_2}{p_1};$$

(1c) probabilidades a priori iguais e custos iguais,

$$R_1 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqslant 1.$$

No caso especial (1c), observe que a regra reduz-se à comparação de densidades tal que se $f_1(\boldsymbol{x}_0) \geqslant f_2(\boldsymbol{x}_0)$, $\boldsymbol{x}_0$ é classificado em π_1. Caso, contrário, $\boldsymbol{x}_0$ é classificado em π_2.

Outros critérios podem ser usados para obter uma regra de classificação ótima. Por exemplo, podemos ignorar os custos de classificação incorreta e escolher R_1 e R_2 que minimizam a probabilidade total de classificação incorreta (PTCI).

$$\begin{aligned} \text{PTCI} &= \text{Pr(classificar incorretamente uma observação)} \\ &= p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}. \end{aligned}$$

Matematicamente, este problema é equivalente a minimizar o custo esperado de classificação incorreta quando os custos de classificação incorreta são iguais (1b).

Poderíamos também alocar uma nova observação $\boldsymbol{x}_0$ a população com maior probabilidade a posteriori $\Pr(\pi_i|\boldsymbol{x}_0)$, $i = 1, 2$.

Pelo teorema de Bayes

$$\begin{aligned}
\Pr(\pi_1|\boldsymbol{x}_0) &= \frac{\Pr(\pi_1 \text{ ocorrer e observarmos } \boldsymbol{x}_0)}{\Pr(\text{observarmos } \boldsymbol{x}_0)} \\
&= \frac{\Pr(\boldsymbol{x}_0|\pi_1)p_1}{\Pr(\boldsymbol{x}_0|\pi_1)p_1 + \Pr(\boldsymbol{x}_0|\pi_2)p_2} \\
&= \frac{p_1 f_1(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)}, \quad \text{e} \\
\Pr(\pi_2|\boldsymbol{x}_0) &= \frac{p_2 f_2(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)}.
\end{aligned}$$

Se

$$\Pr(\pi_1|\boldsymbol{x}_0) \geqslant \Pr(\pi_2|\boldsymbol{x}_0)$$

classificamos $\boldsymbol{x}_0$ em π_1. Caso contrário, classificamos $\boldsymbol{x}_0$ em π_2. Observe que essa regra é equivalente a regra (1b), que considera os custos de classificação incorreta iguais.

Classificação em Uma de Duas Populações Normais Multivariadas

Primeiro, suponha que as populações tenham matrizes de covariâncias iguais, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$.

$$f_j(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right\}, \quad j = 1, 2,$$

com $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}$ desconhecidos.

Proposição 2: A regra do CECI mínimo é dada por

$$\begin{cases} R_1: & \left\{\boldsymbol{x}\middle|(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \geqslant \ln\left(\frac{C(1|2)p_2}{C(2|1)p_1}\right)\right\}; \\ R_2: & \text{caso contrário.} \end{cases}$$

Observação: Nas aplicações $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}$ são desconhecidos, por essa razão, a regra dada pela proposição 2 deve ser modificada.

O estimador de $\boldsymbol{\Sigma}$ é dado por

$$\boldsymbol{S}_c = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{(n_1 + n_2 - 2)}.$$

A regra resultante da substituição pelos vetores de média amostral e a matriz $\boldsymbol{S}_c$ é:

$$\begin{cases} R_1: & \left\{ \boldsymbol{x} \middle| (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1} \boldsymbol{x} - \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \geqslant \ln\left(\frac{C(1|2)p_2}{C(2|1)p_1}\right) \right\}; \\ R_2: & \text{caso contrário.} \end{cases}$$

Se as probabilidades a priori são iguais e os custos de classificação incorreta também são iguais, então a regra acima se simplifica para

$$R_1 : \left\{ \boldsymbol{x} \middle| (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1} \boldsymbol{x} \geqslant \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \right\}.$$

Faça

$$
\begin{aligned}
\widehat{y} &= \widehat{\boldsymbol{a}}^{\top}\boldsymbol{x}, \quad \text{com} \\
\widehat{\boldsymbol{a}}^{\top} &= (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^{\top}\boldsymbol{S}_c^{-1}, \\
\widehat{m} &= \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^{\top}\boldsymbol{S}_c^{-1}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) = \frac{1}{2}(\overline{y}_1 + \overline{y}_2) \quad \text{com} \\
\overline{y}_j &= \widehat{\boldsymbol{a}}^{\top}\overline{\boldsymbol{x}}_j, \quad j = 1, 2.
\end{aligned}
$$

Resumindo, a regra estimada do CECI mínimo é equivalente a criar duas populações normais univariadas para os valores y, tomando-se uma combinação linear apropriada das observações de π_1 e π_2 e, então, designar $\boldsymbol{x}_0$ a π_1 ou a π_2 dependendo se $\widehat{y}_0 = \widehat{\boldsymbol{a}}^{\top}\boldsymbol{x}_0$ cai à direta ou à esquerda do ponto médio $\widehat{m}$ entre as duas médias amostrais $\overline{y}_1$ e $\overline{y}_2$.

Como os parâmetros são substituídos por suas estimativas não se pode mais assegurar que a regra resultante minimize o custo esperado de classificação incorreta em uma particular aplicação.

Porém, parece razoável esperar que ela deva comportar-se bem para tamanhos amostrais grandes.

Resumindo: se os dados parecem ser normais multivariados, a estatística de classificação do lado esquerdo

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1} \boldsymbol{x} - \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)$$

pode ser calculada para cada nova observação $\boldsymbol{x}_0$. Essas observações são classificadas comparando-se os valores da estatística com o valor de

$$\ln\left(\frac{C(1|2)p_2}{C(2|1)p_1}\right).$$

Exemplo 2: Um biólogo obteve medidas sobre $n = 25$ lagartos conhecidos cientificamente como *Cophosaurus texanus*. O peso (*mass*) é dados em gramas, enquanto que o comprimento da abertura do focinho (*svl*) e a extensão dos membros posteriores (*hls*) são dados em milímetros. Os dados estão disponíveis no arquivo T1-3.DAT. Além das três medidas, o biólogo identificou o gênero de cada lagarto *m*-macho, *f*-fêmea. Construir uma regra de classificação de gênero a partir das três medidas usando os dados disponíveis. (Ver exemplo_11.r)

Utilizaremos a função `lda` do `R` disponível no pacote `mass`.

Tabela 3: Probabilidade a prior dos grupos.

Grupo 1	Grupo 2
0,48	0,52

Como nada foi dito na função do `R`, o mesmo adota prioris iguais às proporções amostrais.

Tabela 4: Grupo de médias.

Grupo	*mass*	*svl*	*hls*
Fêmea	7,012	63,042	118,000
Macho	10,233	73,346	139,769

Tabela 5: Coeficientes de discriminação linear.

Variável	LD_1
mass	-0,5723
svl	-0,0908
hls	0,2949

Tabela 6: Erros de classificação.

Grupo	Classificação	
	Fêmea	Macho
Fêmea	11	1
Macho	0	13

Observação: Especificando prioris iguais, a tabela acima não apresentará erros de classificação reaplicada a amostra de aprendizagem.

Escala

Para qualquer constante $c \neq 0$, o vetor $c\widehat{\boldsymbol{a}} = c\boldsymbol{S}_c^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$ também servirá como coeficientes discriminantes. O vetor $\widehat{\boldsymbol{a}}$ é frequentemente "normalizado" para facilitar a interpretação de seus elementos. Duas das normalizações mais comuns são apresentadas a seguir.

1. Faça

$$\widehat{\boldsymbol{a}}^* = \frac{\widehat{\boldsymbol{a}}}{\sqrt{\widehat{\boldsymbol{a}}^\top \widehat{\boldsymbol{a}}}}$$

   tal que $\widehat{\boldsymbol{a}}^*$ tenha comprimento unitário.

2. Faça

$$\widehat{\boldsymbol{a}}^* = \frac{\widehat{\boldsymbol{a}}}{\widehat{a}_1}$$

   tal que o primeiro elemento de $\widehat{\boldsymbol{a}}^*$ seja 1.

Em ambos os casos, $\widehat{\boldsymbol{a}}^*$ é da forma $c\widehat{\boldsymbol{a}}$.

Abordagem de Fisher para Classificação em Uma de Duas Populações

Fisher de fato chegou a estatística de classificação

$$R_1: \quad \left\{ \boldsymbol{x} \middle| (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1} \boldsymbol{x} \geqslant \frac{1}{2} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1} (\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \right\} \quad \text{ou}$$

$$R_1: \quad \left\{ y | \widehat{y} \geqslant (\overline{y}_1 + \overline{y}_2)/2 \right\}, \quad \left(\widehat{y} = \boldsymbol{a}^\top \boldsymbol{x}, \quad \boldsymbol{a} = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1}, \quad \overline{y}_j = \widehat{\boldsymbol{a}}^\top \overline{\boldsymbol{x}}_j \right),$$

usando um argumento completamente diferente. A ideia de Fisher foi transformar as observações multivariadas $\boldsymbol{x}$ em observações univariadas y tal que os y's provenientes de π_1 e de π_2 sejam tão separados quanto possível.

Fisher sugeriu tomar combinações lineares de $\boldsymbol{x}$ para criar os y's, porque elas são funções simples de $\boldsymbol{x}$ e podem ser manipuladas facilmente.

A abordagem de Fisher não assume que as populações sejam normais. No entanto, implicitamente, assume que as matrizes de covariâncias das populações sejam iguais, porque uma estimativa combinada da matriz de covariâncias é usada.

Uma combinação linear fixada de $\boldsymbol{x}$ toma os valores $y_{11}, y_{12}, \ldots, y_{1n_1}$ para as observações de π_1 e os valores $y_{21}, y_{22}, \ldots, y_{2n_2}$ para as observações de π_2.

A separação destes dois conjuntos de valores univariados é avaliada em função da diferença entre as médias amostrais $\overline{y}_1$ e $\overline{y}_2$ expressa em unidades de desvio padrão.

$$\text{Separação} = \frac{|\overline{y}_1 - \overline{y}_2|}{s_y} \quad \text{com}$$

$$s_y^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(y_{1i} - \overline{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \overline{y}_2)^2\right].$$

O objetivo é selecionar a combinação linear de $\boldsymbol{x}$ que alcança a separação máxima das médias amostrais $\overline{y}_1$ e $\overline{y}_2$.

Proposição 3: A combinação linear

$$y = \widehat{\boldsymbol{a}}^\top \boldsymbol{x} = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1} \boldsymbol{x} \quad \text{maximiza a razão}$$

$$\frac{\text{distância quadrada entre médias}}{\text{variância amostral de y}} = \frac{(\overline{y}_1 - \overline{y}_2)^2}{s_y^2} = \frac{(\widehat{\boldsymbol{a}}^\top \boldsymbol{d})^2}{\widehat{\boldsymbol{a}}^\top \boldsymbol{S}_c \widehat{\boldsymbol{a}}}.$$

Regra de Alocação: Função Discriminante (Linear) de Fisher

Sejam

$$\begin{aligned} \widehat{y}_0 &= (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1} \boldsymbol{x}_0 \\ \widehat{m} &= \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1} (\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) = \frac{1}{2}(\overline{y}_1 + \overline{y}_2). \end{aligned}$$

Aloque $\boldsymbol{x}_0$ a π_1 se $\widehat{y}_0 \geqslant \widehat{m}$.

Caso contrário, aloque $\boldsymbol{x}_0$ a π_2.

Classificação É Uma Boa Ideia?

Para duas populações, a separação máxima relativa que pode ser obtida considerando-se combinações lineares das observações multivariadas é igual a distância

$$D^2 = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2).$$

Isto é conveniente, porque D^2 pode ser usada, em certas situações, para testar se as médias das populações π_1 e π_2 diferem sigificativamente. Consequentemente, um teste de diferença entre vetores de média pode ser visto como um teste para a “significância” da separação que pode ser alcançada.

Suponha que as populações π_1 e π_2 sejam normais multivariadas com uma matriz de covariâncias comum $\mathbf{\Sigma}$. Então, vimos que um teste de $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ usa a estatística

$$\left(\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}\right)\left(\frac{n_1 n_2}{n_1 + n_2}\right) D^2,$$

que sob H_0 tem distribuição $\mathcal{F}_{p,n_1+n_2-p-1}$.

Se H_0 é rejeitada, podemos concluir que a separação entre as duas populações é significante.

Observação: Separação significante não necessariamente implicará em boa classificação. A eficácia de um procedimento de classificação pode ser avaliada independentemente de qualquer teste de separação. Em contraste, se a separação não é significante, a busca por uma regra de classificação útil será, provavelmente, infrutífera.

Classificação de Populações Normais - Caso $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

As regras de classificação são mais complicadas quando as matrizes de covariâncias das populações são desiguais. Considere novamente a razão das densidades normais multivariadas, agora considerando as covariâncias desiguais. Neste caso, os fatores fora do termo exponencial não simplificam e não é possível colocar o termo dentro da exponencial em evidência.

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right)^{1/2} \exp\left\{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + (\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right]\right\}.$$

Nesse caso, as regiões de classificação, segundo o critério do custo esperado de classificação incorreta mínimo, serão dadas por (na escala logaritmo natural):

$$\begin{cases} R_1: & R_1: -\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} - k \geqslant \ln\left(\frac{C(1|2)}{C(2|1)}\frac{p_2}{p_1}\right); \\ R_2: & \text{caso contrário,} \end{cases}$$

com

$$k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2).$$

As regiões de classificação são quadráticas em $\boldsymbol{x}$. Quando $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, o termo quadrático $\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}$ se anula, e as regiões resultantes são aquelas obtidas anteriormente no caso de variâncias iguais.

Proposição 4: Sob normalidade multivariada com covariâncias desiguais, aloque $\boldsymbol{x}_0$ a π_1 se

$$-\frac{1}{2}\boldsymbol{x}_0^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}_0 + (\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}_0 - k \geqslant \ln\left(\frac{C(1|2)}{C(2|1)}\frac{p_2}{p_1}\right).$$

Caso contrário, aloque $\boldsymbol{x}_0$ a π_2.

Na prática, a regra de classificação acima é implementada substituindo-se os parâmetros populacionais por estimativas $\overline{\boldsymbol{x}}_1$, $\overline{\boldsymbol{x}}_2$ e $\boldsymbol{S}_1$ e $\boldsymbol{S}_2$.

Regra de Classificação Quadrática

Populações normais, covariâncias desiguais:

Aloque $\boldsymbol{x}_0$ a π_1 se

$$-\frac{1}{2}\boldsymbol{x}_0^\top(\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1})\boldsymbol{x}_0 + (\overline{\boldsymbol{x}}_1^\top\boldsymbol{S}_1^{-1} - \overline{\boldsymbol{x}}_2^\top\boldsymbol{S}_2^{-1})\boldsymbol{x}_0 - k \geqslant \ln\left(\frac{C(1|2)}{C(2|1)}\frac{p_2}{p_1}\right).$$

Caso contrário, aloque $\boldsymbol{x}_0$ a π_2.

Classificação com funções quadráticas é bem complicada quando se tem mais de duas medidas e pode levar a resultados estranhos. Isto é particularmente verdade quando os dados não são (essencialmente) normais multivariados.

As regiões de classificação podem ser uma união de regiões disjuntas do espaço amostral.

Em muitas aplicações, a cauda inferior da distribuição de π_1 será menor do que a prescrita por uma distribuição normal e a regra quadrática poderá levar a altas taxas de erro de classificação. Uma desvantagem séria da regra quadrática é que ela é bem sensível a desvios da normalidade.

Se os dados não são normais multivariados, duas opções estão disponíveis. A primeira, envolve transformar os dados não normais, e depois testar a igualdade das matrizes de covariâncias para verificar se é a regra linear ou a quadrática que devem ser usadas.

Os testes usuais para homogeneidade das covariâncias são fortemente afetados sob não normalidade. A conversão de dados não normais para dados normais deve sempre ser feita antes de realizar tais testes.

Como segunda opção, podemos usar uma regra linear (ou quadrática) sem nos preocuparmos com a forma das distribuições populacionais e esperar que elas irão funcionar razoavelmente bem.

Estudos mostraram, porém, que existem casos não normais para os quais uma função de classificação linear tem uma performance ruim, mesmo se as matrizes de covariâncias das duas populações são iguais.

Moral da história: sempre verificar a performance de qualquer procedimento de classificação. Em último caso, isto deve ser feito com o conjunto de dados usado para construir a regra. O ideal é que se tenha uma quantidade de dados suficientemente grande que podem ser repartidos em amostras de treinamento e de validação. A amostra de treinamento/aprendizagem é usada para construir a regra, e a amostra de validação é usada para avaliar a performance da regra construída.

Avaliação das Funções de Classificação

A avaliação envolve calcular taxas de erro ou probabilidades de classificação incorreta.

Como as densidades são em geral desconhecidas, concentraremo-nos sobre as taxas de erro associadas à função de classificação amostral.

Taxa de Erro Ótima (TEO) - regra de classificação segundo o critério da probabilidade total de classificação incorreta (PTCI) mínima.

$$\text{TEO} = p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}$$

Exemplo 3: Suponha duas populações normais multivariadas com matrizes de covariâncias iguais, $p_1 = p_2 = 1/2$ e também $C(2|1) = C(1|2)$ tal que $\ln\left(\frac{C(1|2)}{C(2|1)}\frac{p_2}{p_1}\right) = 0$.

Neste caso,

$$R_1 : \quad \left\{ \boldsymbol{x} \middle| (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \geqslant \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right\} \quad \text{ou}$$

$$R_1 : \quad \left\{ \boldsymbol{x} \middle| \boldsymbol{a}^\top \boldsymbol{x} \geqslant \frac{1}{2}\boldsymbol{a}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right\}.$$

Fazendo $Y = \boldsymbol{a}^\top \boldsymbol{X}$ teremos $\sigma_Y^2 = \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a} = \delta^2$.

$$\text{PTCI} = \frac{1}{2}\int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{2}\int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} = \Phi\left(-\frac{\delta}{2}\right).$$

Se $\delta^2 = 2,56$, teremos $\text{PTCI} = \Phi(-0,8) = 0,2119$.

A regra de classificação aqui irá alocar cerca de 21% dos itens incorretamente.

Este exemplo ilustra como a TEO pode ser calculada quando as funções de densidade são conhecidas. Como em geral os parâmetros populacionais são desconhecidos, eles deverão ser estimados e a avaliação da taxa de erro não será tão direta.

A performance da função de classificação amostral pode, em princípio, ser avaliada calculando-se a taxa de erro real (TER).

$$\text{TER} = p_1 \int_{\widehat{R}_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{\widehat{R}_1} f_2(\boldsymbol{x}) d\boldsymbol{x}.$$

$\widehat{R}_1$ e $\widehat{R}_2$ são as regiões de classificação determinadas pelas amostras de tamanhos n_1 e n_2, respectivamente.

A TER indica como a função de clasificação amostral se comportará em amostras futuras. Como a TEO, geralmente ela não poderá ser calculada, pois depende das densidades f_1 e f_2. Porém, uma estimativa de uma quantidade relacionada a TER pode ser calculada e será apresentada aqui.

A taxa de erro real aparente (TERA) pode ser calculada a partir da matriz de "confusão" (tabela de dupla entrada indicando as frequências de classificações corretas e incorretas em cada população (grupo)).

População	Classificação em	
	π_1	π_2
π_1	n_{1c}	$n_{1M} = n_1 - n_{1c}$
π_2	$n_{2M} = n_2 - n_{2c}$	n_{2c}

$$\text{TERA} = \frac{n_{1M} + n_{2M}}{n_1 + n_2}.$$

Observe que a TERA nada mais é do que a proporção amostral de classificações incorretas considerando-se a amostra de treinamento.

A TERA é uma medida intuitiva e simples, mas tem um viés: tende a subestimar a TER, a menos que n_1 e n_2 sejam suficientemente grandes.

Estimativas de taxas de erro melhores do que a TERA e que não exigem a suposição das distribuições populacionais podem ser construídas.

Um procedimento é dividir a amostra total em uma amostra de treinamento e outra de validação. A amostra de treinamento é usada para construir a função de classificação e, a de validação, para avaliar a função obtida. A taxa de erro é determinada pela porporção amostral de classificações incorretas na amostra de validação.

Apesar deste método superar o problema do viés, ele padece de dois defeitos:

1. requer amostras muito grandes;
2. a função avaliada não é a função de interesse. Em última análise, quase todos os dados devem ser usados para construir a regra. Caso contrário, informação importante pode estar sendo desperdiçada.

Uma segunda abordagem, que parece funcionar bem, é chamada procedimento de validação "reter um fora" (*holdout*) de Lachenbruch.

1. Comece em π_1. Omita uma de suas observações e desenvolva a função de classificação com as restantes $n_1 - 1 + n_2$.
2. Classifique a observação omitida com a função obtida.
3. Repita os passos (1) e (2) até que todas as observações de π_1 sejam classificadas. Defina $n_{1M}^{(H)}$ como o número de classificações incorretas neste grupo.
4. Repita os passos (1), (2) e (3) para as observações de π_2 e defina $n_{2M}^{(H)}$ como o número de classificações incorretas neste grupo.

$$\widehat{\Pr(2|1)} = \frac{n_{1M}^{(H)}}{n_1}, \quad \widehat{\Pr(1|2)} = \frac{n_{2M}^{(H)}}{n_2} \quad \text{e} \quad \widehat{PTCI} = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}.$$

Para amostras moderadas $\widehat{\text{PTCI}}$ é uma estimativa não viesada do valor esperado da TERA (taxa de erro aparente).

Deve ser intuitivamente claro que classificação boa (taxas de erro pequenas) dependerá da separação dos grupos. O mais separados são os grupos, mais provavelmente uma regra de classificação útil será desenvolvida.

Como veremos, regras de alocação apropriadas para o caso envolvendo probabilidades a priori iguais e custos de classificação incorreta iguais correspondem às funções designadas para populações separadas o máximo possível. É nesta situação que começamos a perder a distinção entre classificação e separação.

Classificação em Uma de g Populações ($g > 2$)

Pelo menos em teoria, a extensão para a classificação em um de g grupos, $g > 2$ é imediata. Porém, não muito é conhecido sobre as propriedades das funções de classificação amostrais correspondentes, e em particular, sobre suas taxas de erro investigadas.

A "robustez" da estatística linear de classificação em dois grupos para, por exemplo, covariâncias desiguais ou distribuições não normais pode ser estudada a partir de experimentos simulados. Para mais de duas populações, esta abordagem não leva a conclusões gerais, porque as propriedades dependem sobre onde as populações estão localizadas, e existem muitas configurações para serem convenientemente estudadas.

Como antes, a abordagem aqui será desenvolver regras ótimas teóricas e, então indicar as modificações exigidas para as aplicações reais.

Regra do Custo Esperado de Classificação Incorreta Mínimo

Notação:

- $f_k(\boldsymbol{x})$ - função de densidade de probabilidade conjunta para o k-ésimo grupo, $k = 1, 2, \ldots, g$.
- $p_1, p_2, \ldots, p_g$ - probabiliades a priori de cada grupo tais que $p_k > 0$, $\forall k$ e $\sum_{k=1}^{g} p_k = 1$.
- $C(k|j)$ - custo de classificação incorreta de uma observação de π_j em π_k, $\forall j, k = 1, 2, \ldots, g$ e $j \neq k$. Se $j = k$, então $c(k|k) = 0$.
- R_k - região de classificação em π_k tal que $\cup_{k=1}^{g} R_k = \Omega$, $R_j \cap R_k = \emptyset$ para $j \neq k$.

A probabilidade de classificar uma observação de π_j em π_k é

$$\Pr(k|j) = \int_{R_k} f_j(\boldsymbol{x}) d\boldsymbol{x} \quad \text{para} \quad k \in \{1, 2, ..., g\}, \quad k \neq j \quad \text{e}$$

$$\Pr(j|j) = 1 - \sum_{k=1, k\neq j}^{g} \Pr(k|j).$$

O custo esperado de classificação incorreta de uma observação proveniente de π_1 será dado por

$$\begin{aligned} \text{CECI}(1) &= \Pr(2|1)C(2|1) + \Pr(3|1)C(3|1) + \cdots + \Pr(g|1)C(g|1) \\ &= \sum_{k=2}^{g} P(k|1)c(k|1). \end{aligned}$$

Este custo esperado condicional ocorre com probabilidade p_1, a probabilidade a priori de π_1.

De maneira similar, podemos obter os custos esperados de classificação incorreta condicionais $\text{CECI}(2), \text{CECI}(3), \ldots, \text{CECI}(g)$.

Multiplicando os custos condicionais pelas respectivas probabilidades a priori temos o custo esperado de classificação incorreta dado por

$$\text{CECI} = \sum_{j=1}^{g} p_j \left(\sum_{k=1, k\neq j}^{g} \Pr(k|j)C(k|j) \right).$$

Proposição 5: As regiões de classificação que minimizam o custo esperado de classificação incorreta são definidas por

- Aloque $\boldsymbol{x}$ a π_j, $j = 1, 2, \ldots, g$ na qual

$$\sum_{j=1, j\neq k}^{g} p_j f_j(\boldsymbol{x}) C(k|j) \quad \text{é um mínimo.}$$

- Se os custos de classificação incorreta são todos iguais a unidade, observe que a regra alocará $\boldsymbol{x}$ à população π_k, $k = 1, 2, \ldots, g$ para a qual,

$$\sum_{j=1, j\neq k}^{g} p_j f_j(\boldsymbol{x}) \quad \text{é um mínimo.}$$

Observe que esta soma será um mínimo se o termo deixado de fora, $p_k f_k(\boldsymbol{x})$, for um máximo.

Regra do CECI Mínimo para Custos de Classificação Incorreta Iguais

- Aloque $\boldsymbol{x}_0$ à população π_k se

$$p_k f_k(\boldsymbol{x}_0) > p_j f_j(\boldsymbol{x}_0), \quad \forall j \neq k,$$

ou equivalentemente,

- Aloque $\boldsymbol{x}_0$ à população π_k se

$$\ln\left(p_k f_k(\boldsymbol{x}_0)\right) > \ln\left(p_j f_j(\boldsymbol{x}_0)\right), \quad \forall j \neq k.$$

Esta regra é equivalente à regra que maximiza a probabilidade a posteriori $\Pr(\pi_k|\boldsymbol{x}_0)$.

Deve-se ter em mente que as regras do CECI mínimo têm três componentes: probabilidades a priori, custos de classificação incorreta e funções de densidade. Estes componentes devem ser especificados (ou estimados) antes da regra poder ser implementada.

Exemplo 4: (Classificação de nova observação em uma de três populações conhecidas)

Suponha os seguintes custos de classificaÇão incorreta, probabilidades a priori e densidades avaliadas em $\boldsymbol{x}_0$ uma nova observação.

População	Classificação em		
	π_1	π_2	π_3
π_1	$C(1\|1) = 0$	$C(2\|1) = 10$	$C(3\|1) = 50$
π_2	$C(1\|2) = 500$	$C(2\|2) = 0$	$C(3\|2) = 200$
π_3	$C(1\|3) = 100$	$C(2\|3) = 50$	$C(3\|3) = 0$
Prioris	$p_1 = 0,05$	$p_2 = 0,60$	$p_3 = 0,35$
$f_j(\boldsymbol{x}_0)$	$f_1(\boldsymbol{x}_0) = 0,01$	$f_2(\boldsymbol{x}_0) = 0,85$	$f_3(\boldsymbol{x}_0) = 2$

Classificar $\boldsymbol{x}_0$ em uma das três populações.

Usando a regra do CECI mínimo, alocaremos $\boldsymbol{x}_0$ a π_k, $k = 1, 2, 3$ para a qual

$$\sum_{j=1, j\neq k}^{3} p_j f_j(\boldsymbol{x}) C(k|j) \quad \text{é um mínimo.}$$

k	$\sum_{j=1, j\neq k}^{3} p_j f_j(\boldsymbol{x}) C(k\|j)$
1	325
2	**35,055**
3	102,025

Como o menor valor ocorre para $k = 2$, alocamos $\boldsymbol{x}_0$ a π_2.

Se os custos de classificação incorreta fossem todos iguais, designaríamos $\boldsymbol{x}_0$ a π_k, $k = 1, 2, 3$ na qual $p_k f_k(\boldsymbol{x}_0) > p_j f_j(\boldsymbol{x}_0), \quad \forall j \neq k$.

k	$p_k f_k(\boldsymbol{x}_0)$
1	0,0005
2	0,5100
3	**0,7000**

Nesse caso, $\boldsymbol{x}_0$ seria alocada a π_3.

Classificação com Populações Normais

$$f_k(\boldsymbol{x}) = (2\pi)^{-p/2}|\Sigma_k|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\}, \quad k = 1, 2, \ldots, g.$$

Se considerarmos todos os custos iguais a unidade, a regra resultante será:

- Aloque $\boldsymbol{x}_0$ a π_k se

$$\begin{aligned}\ln(p_k f_k(\boldsymbol{x}_0)) &= \ln p_k - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) \\ &= \max_{1\leqslant j\leqslant g}\ln(p_j f_j(\boldsymbol{x}_0))\end{aligned}$$

A constante $p\ln(2\pi)/2$ pode ser desprezada, pois ela é igual para todas as populações. Portanto, podemos definir um escore discriminante quadrático para a k-ésima população dado por

$$d_k^Q(\boldsymbol{x}) = \ln p_k - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k), \quad k = 1, 2, \ldots, g.$$

O escore quadrático $d_k^Q(\underline{x})$ é composto pelas contribuições da variância generalizada $|\Sigma_k|$, da probabilidade a priori p_k, e da distância quadrada de $\boldsymbol{x}$ à média da população π_k.

Regra da Probabilidade Total de Classificação Incorreta Mínima - Populações Normais, Covariâncias Desiguais

- Aloque $\boldsymbol{x}_0$ a π_k se o escore quadrático

$$d_k^Q(\boldsymbol{x}_0) = \max_{1 \leqslant j \leqslant g} \{d_j^Q(\boldsymbol{x}_0)\}.$$

Na prática $\boldsymbol{\mu}_k$ e $\boldsymbol{\Sigma}_k$ são desconhecidas para todo $k = 1, 2, \ldots, g$, mas um conjunto de treinamento cujas classificações corretas das observações são conhecidas está em geral disponível para a construção de estimativas. As quantidades amostrais relevantes para a população π_k são $\overline{\boldsymbol{x}}_k$ e $\boldsymbol{S}_k$ com n_k o número de observações da k-ésima população.

$$\widehat{d}_k^Q(\boldsymbol{x}) = \ln p_k - \frac{1}{2} \ln |\boldsymbol{S}_k| - \frac{1}{2}(\boldsymbol{x} - \overline{\boldsymbol{x}}_k)^\top \boldsymbol{S}_k^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}_k), \quad k = 1, 2, \ldots, g.$$

Assim, a regra estimada é alocar $\boldsymbol{x}_0$ a π_k se o escore quadrático estimado

$$\widehat{d}_k^Q(\boldsymbol{x}_0) = \max_{1 \leqslant j \leqslant g} \{\widehat{d}_j^Q(\boldsymbol{x}_0)\}.$$

Uma simplificação aqui é possível para o caso em que $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$.

Neste caso os escores discriminantes passam a ser lineares em $\boldsymbol{x}$ e simplificam para

$$d_k(\boldsymbol{x}) = \ln p_k + \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k, \quad k = 1, 2, \ldots, g.$$

Uma estimativa de $d_k(\boldsymbol{x})$ é baseada em

$$\boldsymbol{S}_c = \frac{1}{n_1 + n_2 + \cdots + n_g - g} \sum_{k=1}^{g} (n_k - 1) \boldsymbol{S}_k$$

e é dada por

$$\widehat{d}_k(\boldsymbol{x}) = \ln p_k + \overline{\boldsymbol{x}}_k^\top \boldsymbol{S}_c^{-1} \boldsymbol{x} - \frac{1}{2} \overline{\boldsymbol{x}}_k^\top \boldsymbol{S}_c^{-1} \overline{\boldsymbol{x}}_k, \quad k = 1, 2, \ldots, g.$$

Consequentemente, temos a regra estimada dada por "aloque $\boldsymbol{x}_0$ a π_k se

$$\widehat{d}_k(\boldsymbol{x}_0) = \max_{1 \leqslant j \leqslant g} \{\widehat{d}_j(\boldsymbol{x}_0)\}".$$

Método de Discriminação de Fisher para Várias Populações

O propósito principal na análise discriminante de Fisher (ADF) é separar populações (grupos). No entanto, como veremos, o produto final pode levar a uma regra de classificação. Na ADF não é necessário supor a normalidade das g populações, embora supõe-se que as covariâncias sejam iguais.

Denote $\overline{\boldsymbol{\mu}}$ o vetor de médias combinado das g populações, isto é,

$$\overline{\boldsymbol{\mu}} = \frac{1}{g}\sum_{k=1}^{g}\boldsymbol{\mu}_k.$$

Denote a matriz $\boldsymbol{B}_{\boldsymbol{\mu}}$ a matriz $p \times p$ de somas de quadrados e produtos cruzados

$$\boldsymbol{B}_{\boldsymbol{\mu}} = \sum_{k=1}^{g}(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})^\top.$$

Faça $Y = \boldsymbol{a}^\top \boldsymbol{X}$ tal que $\mathrm{E}(Y|\pi_k) = \boldsymbol{a}^\top \boldsymbol{\mu}_k$ e $\mathrm{Var}(Y|\pi_k) = \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}$, $k = 1, 2, \ldots, g$.

Consequentemente $\mu_{kY} = \boldsymbol{a}^\top \boldsymbol{\mu}_k$ depende da população na qual $\boldsymbol{X}$ foi observada.

Média Global: $\overline{\mu}_Y = \boldsymbol{a}^\top \frac{1}{g}\sum_{k=1}^{g}\boldsymbol{\mu}_k = \boldsymbol{a}^\top \overline{\boldsymbol{\mu}}$.

Medida de separação dos grupos:

$$\begin{aligned}\frac{\sum_{k=1}^{g}(\mu_{kY}-\overline{\mu}_Y)^2}{\sigma_Y^2} &= \frac{\sum_{k=1}^{g}(\boldsymbol{a}^\top\boldsymbol{\mu}_k-\boldsymbol{a}^\top\overline{\boldsymbol{\mu}})^2}{\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a}}\\ &= \frac{^\top\boldsymbol{B}_\mu\boldsymbol{a}}{\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a}}.\end{aligned}$$

Esta razão mede a variabilidade entre os grupos de valores de Y relativa à variabilidade comum dentro dentro dos grupos.

Podemos selecionar $\boldsymbol{a}$ que maximiza esta razão.

Em geral $\boldsymbol{\Sigma}$ e $\boldsymbol{\mu}_k$ são desconhecidos, mas dispõe-se de uma amostra de treinamento a partir da qual podemos estimar estas quantidades.

Sejam $\overline{\boldsymbol{x}}_k$, $\boldsymbol{S}_k$ as estimativas em cada grupo e

$$\begin{aligned}\overline{\boldsymbol{x}} &= \frac{1}{g}\sum_{k=1}^{g}\overline{\boldsymbol{x}}_k,\\ \widehat{\boldsymbol{B}}_{\boldsymbol{\mu}} &= \sum_{k=1}^{g}(\overline{\boldsymbol{x}}_k-\overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_k-\overline{\boldsymbol{x}})^\top \quad \text{e}\\ \widehat{\boldsymbol{W}} &= \sum_{k=1}^{g}\sum_{j=1}^{n_k}(\boldsymbol{x}_{jk}-\overline{\boldsymbol{x}}_k)(\boldsymbol{x}_{jk}-\overline{\boldsymbol{x}}_k)^\top.\end{aligned}$$

Discriminantes Lineares Amostrais de Fisher (DALF)

Sejam $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_s$, $s \leq \text{mín}\{g-1, p\}$ autovalores não-nulos de $\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{B}}_{\mu}$ e $\widehat{\boldsymbol{v}}_1, \widehat{\boldsymbol{v}}_2, \ldots, \widehat{\boldsymbol{v}}_s$ os autovetores correspondentes tal que $(\widehat{\boldsymbol{v}}_k^{\top}\boldsymbol{S}_c^{-1}\widehat{\boldsymbol{v}}_k = 1)$.

Então, o vetor $\widehat{\boldsymbol{a}}$ que maximiza a razão $\dfrac{\boldsymbol{a}^{\top}\widehat{\boldsymbol{B}}_{\mu}\boldsymbol{a}}{\boldsymbol{a}^{\top}\widehat{\boldsymbol{W}}\boldsymbol{a}}$ é dado por $\widehat{\boldsymbol{a}}_1 = \widehat{\boldsymbol{v}}_1$.

A combinação linear $\widehat{\boldsymbol{a}}_1^{\top}\boldsymbol{x}$ é chamada primeiro discriminante amostral.

A escolha $\widehat{\boldsymbol{a}}_2 = \widehat{\boldsymbol{v}}_2$ produz o segundo discriminante amostral e continuamos até obter o k-ésimo discriminante amostral $\widehat{\boldsymbol{a}}_k = \widehat{\boldsymbol{v}}_k$, $k \leq s$.

Comentários Finais

Regressão Logística

As funções de classificação discutidas até aqui são baseadas em variáveis quantitativas. A regressão logística é uma abordagem apropriada para classificação quando algumas ou todas as variáveis são qualitativas. Na sua configuração mais simples, a variável resposta Y está restrita a dois valores. Por exemplo, Y pode representar gênero: macho/fêmea, ou empregado/desempregado, aprovado/reprovado, etc.

Quando a resposta assume apenas dois valores possíveis é comum codificá-la como 0 ou 1 e, o interesse passa a ser estimar a probabilidade da variável assumir o valor 1 dado o vetor de covariáveis $\boldsymbol{x}$, que representa a proporção na população codificada com o valor 1.

Esta modelagem pode então ser usada para fins de classificação em um de dois grupos, e a ideia pode ser estendida para vários grupos, substituindo a dsitribuição binomial pela multinomial.

Inclusão de Variáveis Qualitativas

Neste parte assumimos que as variáveis de discriminação $X_1, X_2, \ldots, X_p$ são contínuas. Com frequência, uma variável qualitativa ou categórica pode ser útil como variável discriminante (classificadora). Esta situação é frequentemente contornada criando-se uma variável X cujo valor numérico é 1 se o objeto possui a tal característica e zero, caso contrário. A variável é, então, tratada como uma variável de medida nos procedimentos de classificação e discriminação usuais.

Exceto para classificação logística, há pouca teoria disponível para lidar com o caso em que algumas variáveis são contínuas e outras são qualitativas. Experimentos de simulação indicaram que a funçãoo discriminante linear de Fisher pode comportar-se tanto pobremente como satisfatoriamente, dependendo das correlações entre as variáveis contínuas e qualitativas. Krzanowski: “Uma correlação baixa em uma população, mas uma correlação alta na outra, ou uma mudança no sinal das correlações entre as duas populações poderiam indicar condições desfavoráveis à função discriminante linear de Fisher”. Esta é uma área problemática e que precisa de mais estudo.

Árvores de Classificação

Uma abordagem de classificação completamente diferente dos métodos discutidos aqui foi desenvolvida. (Breiman, L., 1. Friedman, R Olshen, and C. Stone. Classification and Regression Trees. Belmont, CA: Wadsworth, Inc., 1984.) Ela é computacionalmente intensiva. A abordagem, chamada árvore de classificação e regressão (CART), é proximamente relacionada com as técnicas de conglomeração divisivas. Inicialmente, todos os objetos são considerados em um único grupo. O grupo é então dividido em dois subgrupos, usando, por exemplo, altos valores de uma variável para um grupo e baixos valores dessa mesma variável para o outro grupo. Os dois subgrupos são então cada um dividido novamente, agora usando valores de uma segunda variável. O processo de divisão continua até que um ponto de parada adequado seja atingido. Os valores das variáveis divisoras podem ser categorias ordenados ou não. É este aspecto que torna o CART tão geral.

Redes Neurais

Uma rede neural é um procedimento computacional intensivo para transformar entradas em saídas programadas usando redes altamente conectadas de unidades de processamento relativamente simples (neurônios ou nós). Suas três características essenciais são as unidades básicas de computação (neurônios ou nós), a arquitetura da rede descrevendo as conexões entre as unidades de computação, e o algoritmo de treinamento usado para encontrar valores dos parâmetros da rede (pesos) para realizar uma tarefa particular.

As unidades de computação são conectadas umas às outras no sentido de que a saída de uma unidade pode servir como entrada para outra unidade. Cada unidade de computação transforma uma entrada em uma saída usando alguma função pré-especificada que é tipicamente monótona, mas de alguma forma arbitrária. Esta função depende de constantes (parâmetros) cujos valores devem ser determinados com um conjunto de treinamento de entradas e saídas.

Arquitetura da rede é a organização das unidades computacionais e os tipos de conexão permitidos. Em aplicações estatísticas, as unidades computacionais são arrumadas em uma série de camadas com conexões entre nós em camadas diferentes, mas não entre nós da mesma camada. A camada que recebe as entradas iniciais é chamada camada de entrada. A camada final é chamada camada de saída. Todas as camadas entre as camadas de entrada e saída são chamadas camadas ocultas.

Redes Neurais podem ser usadas para discriminação e classificação. Quando elas são usadas com este fim, as variáveis de entrada são as medidas $X_1, X_2, \ldots, X_p$, e a variável de saída é a variável categórica que indica de qual grupo veio a observação de entrada. A experiência inidca que redes neurais apropriadamente construídas comportam-se tão bem quanto à regressão logística e as funções discriminantes discutidas aqui. Os autores sugerem a seguinte referência para uma boa discussão do uso de redes neurais em aplicações da estatística: Stem, H. S. Neural Networks in Applied Statistics. *Technometrics*, 38, (1996), 205-214.

Seleção de Variáveis

Em algumas aplicações da análise discriminante, os dados estão disponíveis para um grande número de variáveis. Mucciardi e Gose (A Comparison of Seven Techniques for Choosing Subsets of Pattern Recognition Properties. *IEEE Trans. Computers*, C20 (1971), 1023-1031.) estudaram uma análise discriminante baseada em 157 variáveis. Neste caso, seria obviamente desejável selecionar um subconjunto menor de variáveis que contivesse quase toda a informação original para efeitos da classificação. Este é o propósito da análise discriminante passo-a-passo *stepwise*, e vários programas de computador dispõem destas funções de seleção de variável.

Se uma análise discriminante *stepwise* (ou qualquer outro método de seleção) é empregado, os resultados devem ser interpretados com cautela. (Veja Murray, G. D. A Cautionary Note on Selection of Variables in Discriminant Analysis. *Applied Statistics*, 26, no. 3 (1977),246-250.) Não há garantia de que o subconjunto selecionado seja o "melhor", sem olhar o critério usado para fazer a seleção. Por exemplo, subconjuntos selecionados com base na minimização da taxa de erro aparente ou maximização do "poder de discriminação" podem comportar-se pobremente em amostras futuras. Problemas associados com procedimentos de seleção de variáveis são ampliados se existem correlações altas entre as variáveis ou entre ombinações lineares das variáveis.

A escolha de um subconjunto de variáveis que parece ser ótima para um dado conjunto de dados é especialmente preocupante se a classificação é o objetivo. No mínimo, a função de classificação obtida deve ser avaliada com uma amostra de validação. Como Murray (1977) sugeriu, uma ideia melhor pode ser dividir a amostra em um número de lotes e determinar o “melhor” subconjunto para cada lote. O número de vezes que uma dada variável aparece nos melhores subconjuntos fornece uma medida do valor dessa variável para classificações futuras.