ANÁLISE ESTATÍSTICA MULTIVARIADA

Ralph dos Santos Silva

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

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

Programa da disciplina

- Distribuições de vetores e matrizes aleatórias:
  - Normal multivariada;
  - *t*-Student multivariada; e
  - Wishart (inversa Wishart).
- Inferência sobre os parâmetros.
- Análise de variância multivariada.
- Regressão linear multivariada.
- Discriminação e classificação.
- Análise de componentes principais.
- Análise fatorial.
- Análise de conglomerados.
- Escalonamento multidimensional.

Referências

- Johnson e Wichern (2007). *Applied Multivariate Statistical Analysis*, 6th Edition. Prentice-Hall.
- Chatfield e Collins (1980). *Introduction to Multivariate Statistical Analysis*. Chapman and Hall.
- Anderson (1958). *An Introduction to Statistical Analysis*. John Wiley and Sons.

Organização de dados

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 $\boldsymbol{x}_{.i}$ 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 $p \times n$, isto é,

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

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

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

em que $\boldsymbol{x}_{j.}^{\top} = (x_{j1}, x_{j2}, \ldots, x_{jn})$ é o vetor das n observações da j-ésima variável, $j = 1, 2, \ldots, p$.

Também podemos representar a matriz $\boldsymbol{X}$ através de suas colunas.

Adotando 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

Uma seleção de 4 notas fiscais 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

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,

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

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

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

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

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

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

Agora, podemos representar de forma organizada as informações sobre variabilidade através da **matriz de 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: $\boldsymbol{S} = \boldsymbol{S}^\top$.

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

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

e a matriz de covariâncias na forma

$$\begin{aligned} \boldsymbol{S} &= \frac{1}{n-1}(\boldsymbol{X} - \overline{\boldsymbol{x}}\mathbf{1}^\top)(\boldsymbol{X} - \overline{\boldsymbol{x}}\mathbf{1}^\top)^\top \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_{.i} - \overline{\boldsymbol{x}})(\boldsymbol{x}_{.i} - \overline{\boldsymbol{x}})^\top. \end{aligned}$$

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: $\boldsymbol{R} = \boldsymbol{R}^\top$.

Algebricamente, a matriz de correlações pode ser escrita na forma

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

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

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

Ver arquivo exemplo_01.r

Representação gráfica

Podemos utilizar gráficos de dispersão para variáveis duas a duas.

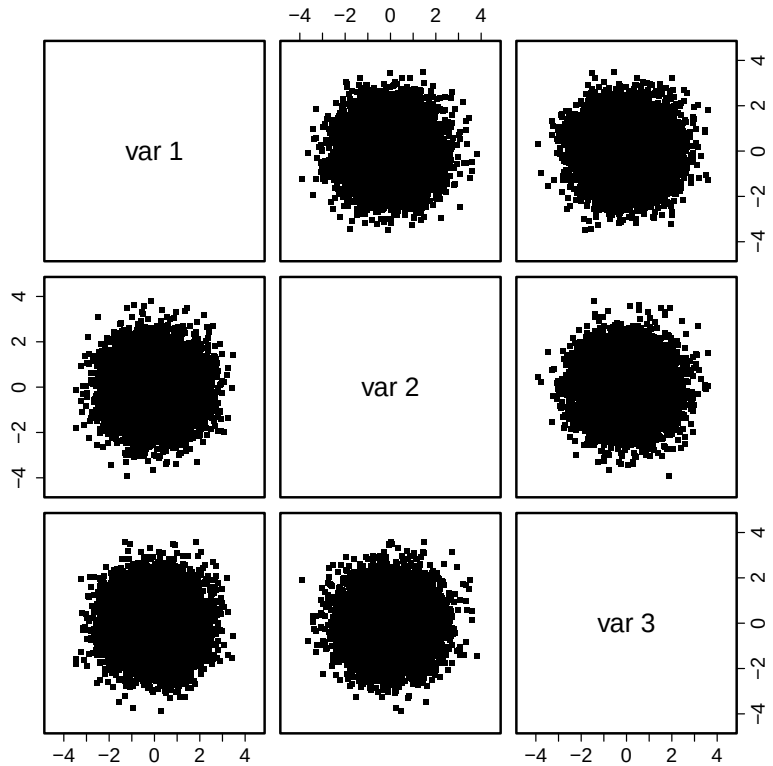

Figura 1: Exemplo de gráfico de dispersão de variáveis duas a duas.

Ver arquivo exemplo_02.r

Exemplo

O papel é fabricado em folhas contínuas com vários metros de largura. Por causa da orientação das fibras dentro do papel, ele tem uma força diferente quando medido na direção produzida pela máquina do que quando medido em ângulos retos para a direção da máquina.

Temos as seguintes variáveis:

x_1: densidade (gramas/centímetro cúbico);

x_2: força (libras) na direção da máquina; e

x_3: força (libras) na direção cruzada.

Ver arquivo exemplo_03.r

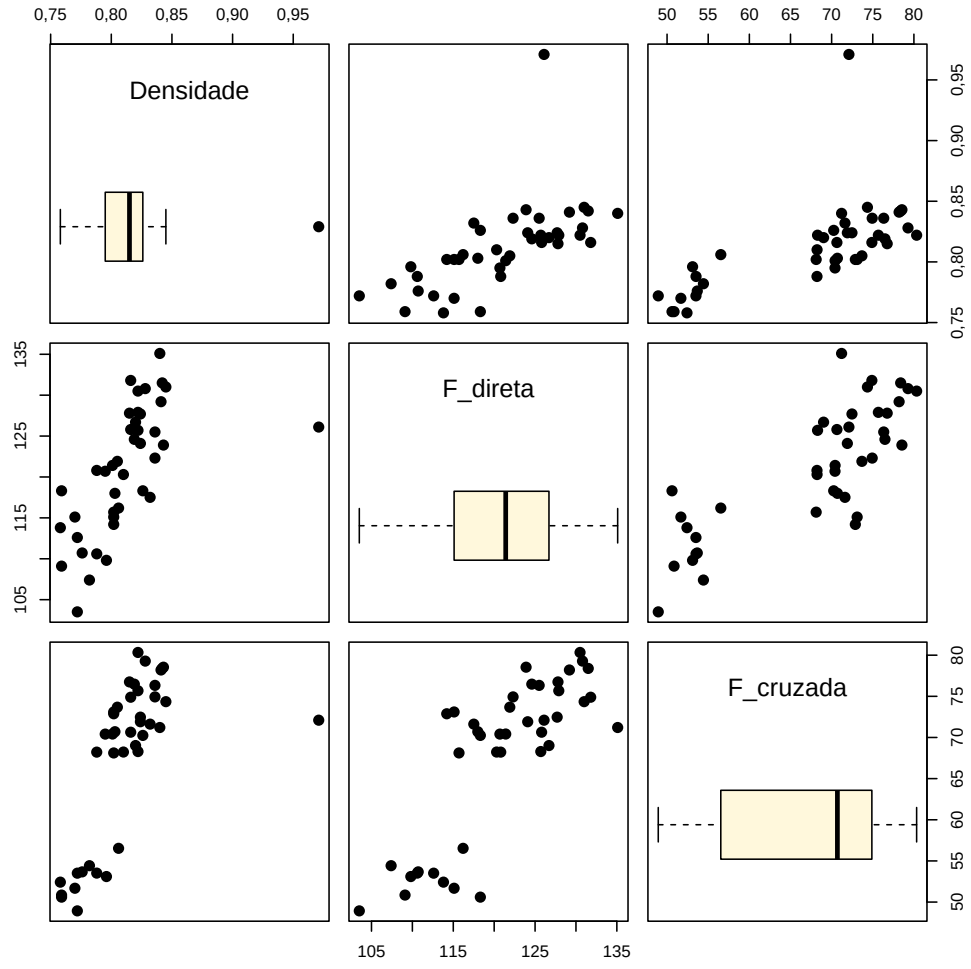

Figura 2: Gráficos de dispersão e *boxplots* das variáveis densidade, força na direção da máquina e força na direção cruzada referentes à qualidade do papel.

Distâncias

A maior parte das técnicas multivariadas se baseia 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.

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 ($\sqrt{\sigma_{jj}}$) e calculando a distância euclideana

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

em que $\boldsymbol{\Omega}$ é uma matriz diagonal com elementos $\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp}$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ 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

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

em que $\boldsymbol{\Sigma}$ é a matriz de covariâncias.

Distribuição normal

Seja X uma variável aleatória contínua com distribuição normal com média μ e variância σ^2. Então,

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

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

Temos que

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

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

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

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

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

em que a matriz de pesos é 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

Consideramos a seguir a particularização da normal multivariada para o caso bivariado.

Considere o seguinte:

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

e

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

em que $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1) \Rightarrow \sigma_{12} = \sigma_{21}$. 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 arquivo exemplo_04.r

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 representar a distribuição normal multivariada como uma transformação linear de variáveis aleatórias normais univariadas?

Podemos representar a distribuição normal multivariada como uma transformação linear de variáveis aleatórias normais **padrão** univariadas?

Podemos ver a distribuição normal multivariada como uma elipse p-dimensional?

Sim é a resposta para as três perguntas acima.

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

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

em que $\boldsymbol{V}$ é uma matriz $p \times p$ ortonormal (as colunas da matriz são ortogonais e cada uma tem norma 1) e $\boldsymbol{D}$ é uma matriz $p \times p$ diagonal com elementos d_j, $j = 1, \ldots, p$, positivos. Em geral, os elementos d_j estão ordenados de forma não crescente.

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 $\boldsymbol{L} = \boldsymbol{U}^\top$.

Ver arquivo exemplo_05.r

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

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

Revisão

Seja X uma variável aleatória contínua com função de densidade de probabilidade $f(x)$.

Momentos ordinários:

$$\mathrm{E}(X^\ell) = \int_{-\infty}^{\infty} x^\ell f(x) dx.$$

Definimos $\mu_x = \mathrm{E}(X)$ como a **média**.

Momentos centrais:

$$\mathrm{E}([X - \mu_x]^\ell) = \int_{-\infty}^{\infty} (x - \mu_x)^\ell f(x) dx.$$

Definimos $\sigma_x^2 = \mathrm{E}([X - \mu_x]^2)$ como a **variância**.

Definimos $\sigma_x = \sqrt{\sigma_x^2}$ como **desvio padrão**.

O **coeficiente de assimetria** é definido por

$$S(X) = \mathrm{E}\left(\left[\frac{X - \mu_X}{\sigma_X}\right]^3\right) = \frac{\mathrm{E}([X - \mu_X]^3)}{\sigma_X^3}$$

e o **coeficiente de curtose** por

$$K(X) = \mathrm{E}\left(\left[\frac{X - \mu_X}{\sigma_X}\right]^4\right) = \frac{\mathrm{E}([X - \mu_X]^4)}{\sigma_X^4}.$$

A quantidade $K(X) - 3$ é chamada de **excesso de curtose** porque $K(X) = 3$ para a distribuição normal[1].

Uma distribuição com excesso de curtose positivo é dita ter **caudas pesadas** (leptocúrtica).

Uma distribuição com excesso de curtose negativo é dita ter **caudas leves** (platicúrtica).

[1] Se $X \sim \mathcal{N}(\mu, \sigma^2)$, então $K(X) = 3$.

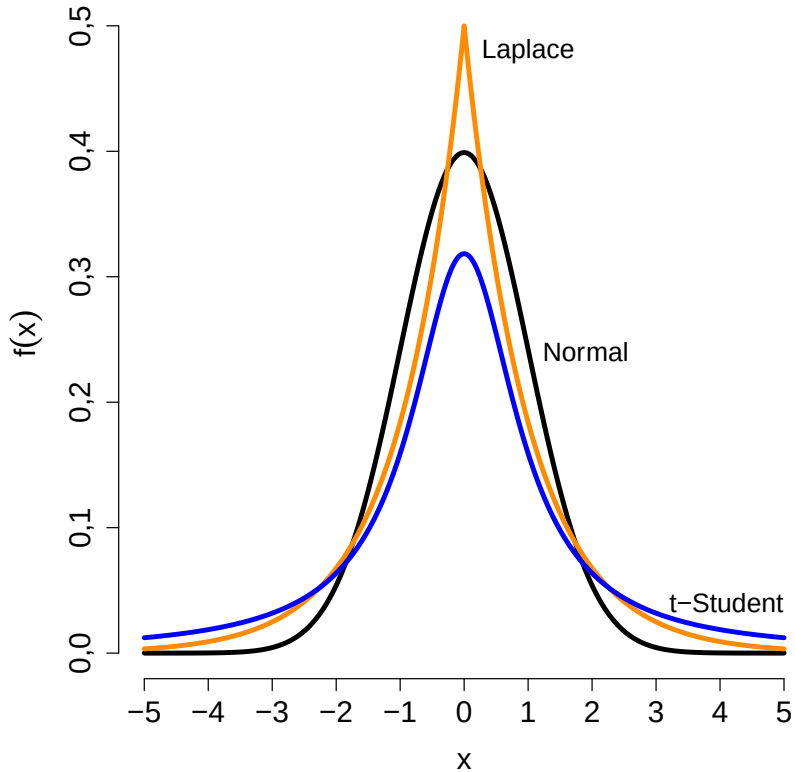

Figura 3: Comparação das funções de densidades de probabilidades da normal, Laplace e *t*-Student com 1 grau de liberdade.

- A curtose da Laplace é maior que a curtose da normal.
- A curtose da *t*-Student[2] é maior que a curtose da Laplace.

[2] Distribuição *t*-Student com 1 grau de liberdade ou distribuição de Cauchy.

Momentos amostrais (unidimensionais)

Suponha que $\{X_1, X_2, \ldots, X_n\}$ seja uma amostra aleatória de X com n observações.

- Média amostral:

$$\hat{\mu}_x = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Variância amostral:

$$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2.$$

- O coeficiente de assimetria amostral:

$$\hat{S}(X) = \frac{M^3}{\hat{\sigma}_x^3} \quad \text{com} \quad M^3 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^3.$$

- O coeficiente de curtose amostral:

$$\hat{K}(X) = \frac{M^4}{\hat{\sigma}_x^4} \quad \text{com} \quad M^4 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^4.$$

M^ℓ, $\ell = 1, 2, \ldots$, podem ser chamados de **momentos centrais amostrais**.

Testes de assimetria e curtose

Sob a hipótese de normalidade (**que a amostra aleatória é proveniente de uma distribuição normal**), temos que

$$\hat{S}(X) \approx \mathcal{N}(0, 6/n) \quad \text{e} \quad \hat{K}(X) \approx \mathcal{N}(3, 24/n)$$

para n "suficientemente grande".

Dado uma amostra aleatória $\{X_1, X_2, \ldots, X_n\}$, para testar a assimetria da distribuição dos dados, consideramos

$$\begin{cases} H_0: & S(X) = 0 \\ H_1: & S(X) \neq 0. \end{cases}$$

A estatística da razão t-Student da assimetria é

$$T = \frac{\widehat{S}(X)}{\sqrt{6/n}} \approx \mathcal{N}(0, 1).$$

Rejeitamos H_0 ao nível α de significância se $|T_{obs}| > z_{1-\alpha/2}$, em que $z_{1-\alpha/2}$ é o percentil $100(1-\alpha/2)$ da distribuição normal padrão.

Para testar o excesso de curtose da distribuição dos dados, consideramos

$$\begin{cases} H_0: & K(X) - 3 = 0 \\ H_1: & K(X) - 3 \neq 0. \end{cases}$$

A estatística da razão t-Student do excesso de curtose é

$$T = \frac{\widehat{K}(X) - 3}{\sqrt{24/n}} \approx \mathcal{N}(0, 1).$$

Rejeitamos H_0 ao nível α de significância se $|T_{obs}| > z_{1-\alpha/2}$, em que $z_{1-\alpha/2}$ é o percentil $100(1 - \alpha/2)$ da distribuição normal padrão.

Teste de normalidade de Jarque-Bera

Considere o seguinte teste de hipóteses:

$$\begin{cases} H_0: & \textbf{A amostra aleatória é proveniente de uma dist. normal;} \\ H_1: & \text{A amostra aleatória } \textbf{não} \text{ é proveniente de uma dist. normal.} \end{cases}$$

A estatística do teste de normalidade de **Jarque e Bera** é dada por

$$JB = \frac{\widehat{S}^2(X)}{6/n} + \frac{(\widehat{K}(X) - 3)^2}{24/n} \approx \chi^2_2,$$

para n "suficientemente grande".

Rejeitamos H_0 (normalidade) se $JB > \chi^\star_{1-\alpha}$, em que $\chi^\star_{1-\alpha}$ é o percentil $100(1-\alpha)$ da distribuição χ^2_2.

Ver arquivos exemplo_06.r

Distribuição gama

Seja X uma variável aleatória com distribuição gama com parâmetros de forma α e de escala β. Então, a função de densidade de probabilidade é dada por

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\} I_{[0,\infty)}(x), \quad \alpha > 0 \quad \text{e} \quad \beta > 0.$$

Temos que

$$\begin{aligned} \mathrm{E}(X) &= \frac{\alpha}{\beta} \\ \mathrm{Var}(X) &= \frac{\alpha}{\beta^2} \end{aligned}$$

Denotaremos a distribuição gama com parâmetros α e β por $\mathcal{G}(\alpha,\beta)$.

Distribuição inversa gama

Seja Y uma variável aleatória com distribuição gama com parâmetros de forma α e de escala β e $X = 1/Y$. Então, a função de densidade de probabilidade de X é dada por

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left\{-\frac{\beta}{x}\right\} I_{[0,\infty]}(x), \quad \alpha > 0 \quad \text{e} \quad \beta > 0.$$

Temos que

$$\begin{aligned} \mathrm{E}(X) &= \frac{\beta}{\alpha-1}, \quad \text{para} \quad \alpha > 1 \\ \mathrm{Var}(X) &= \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \quad \text{para} \quad \alpha > 2 \end{aligned}$$

Denotaremos a distribuição gama invertida com parâmetros α e β por $\mathcal{IG}(\alpha,\beta)$.

Distribuição qui-quadrada

Seja X uma variável aleatória com distribuição qui-quadrada com (parâmetro) n graus de liberdade. Então, a função de densidade de probabilidade é dada por

$$f(x|n) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{n/2-1} \exp\left\{-\frac{x}{2}\right\} \boldsymbol{I}_{[0,\infty]}(x), \quad \alpha > 0 \quad \text{e} \quad \beta > 0.$$

Temos que

$$\begin{aligned} \mathrm{E}(X) &= n \\ \mathrm{Var}(X) &= 2n \end{aligned}$$

Denotaremos a distribuição qui-quadrada com n graus de liberdade por χ^2_n.

Distribuição *t*-Student

Seja X uma variável aleatória com distribuição *t*-Student com parâmetros localização μ, de escala σ^2 e ν graus de liberdade. Então, a função de densidade de probabilidade é dada por

$$\begin{aligned} f(x|\mu,\sigma^2,\nu) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{1}{2}\right)\nu^{1/2}\sigma}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}, \\ &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi^{1/2}\nu^{1/2}\sigma}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}, \quad x\in\mathbb{R}, \end{aligned}$$

em que $\mu\in\mathbb{R}$, $\nu>0$ e $\sigma^2>0$. Temos que

$$\begin{aligned} \mathrm{E}(X) &= \mu, && \text{se} \quad \nu>1 \\ \mathrm{Var}(X) &= \frac{\nu}{\nu-2}\sigma^2, && \text{se} \quad \nu>2. \end{aligned}$$

Denotaremos a distribuição *t*-Student por $t_\nu(\mu,\sigma^2)$.

Distribuição F-Snedecor

Seja X uma variável aleatória com distribuição F-Snedecor com (parâmetros) graus de liberdades n e m. Então, a função de densidade de probabilidade é dada por

$$f(x|n,m) = \left(\frac{n}{m}\right)^{n/2} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{n/2-1} \left(1+\frac{n}{m}x\right)^{-(n+m)/2} \boldsymbol{I}_{[0,\infty)}(x),$$

em que $n > 0$ e $m > 0$.

Denotaremos a distribuição F-Snedecor com n e m graus de liberdade por $\mathcal{F}_{n,m}$.

Resultado

Se Z e W são variáveis aleatórias independentes tal que $Z \sim \mathcal{N}(0,1)$, $W \sim \chi^2_n$ e

$$T = \frac{Z}{\sqrt{W/n}},$$

então $T \sim t_n(0,1)$.

Prova

$$\begin{aligned} F_T(t) &= \Pr(T \leqslant t) = \Pr(Z \leqslant t\sqrt{W/n}) \\ &= \int_0^\infty \Pr(Z \leqslant t\sqrt{w/n} \Big| W = w) f_W(w) dw \\ &= \int_0^\infty \Phi(t\sqrt{w/n}) f_W(w) dw. \end{aligned}$$

$$
\begin{aligned}
f_T(t) &= \frac{\partial F_T(t)}{\partial t} = \frac{\partial}{\partial t}\int_0^\infty \Phi(t\sqrt{w/n}) f_W(w)dw \\
&= \int_0^\infty \frac{\partial}{\partial t}\Phi(t\sqrt{w/n}) f_W(w)dw = \int_0^\infty \phi(t\sqrt{w/n})\sqrt{\frac{w}{n}} f_W(w)dw \\
&= \frac{1}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)2^{(n+1)/2}n^{1/2}}\int_0^\infty w^{(n+1)/2-1}\exp\left\{-w\frac{1}{2}\left(1+\frac{t^2}{n}\right)\right\}dw \\
&= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)n^{1/2}}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad t \in \mathbb{R}.
\end{aligned}
$$

Portanto, $T \sim t_n(0,1)$.

Resultado

Se $T \sim t_n(0,1)$ e $Y = T^2$, então $Y \sim \mathcal{F}_{1,n}$.

Prova

$$
\begin{aligned}
F_Y(y) &= \Pr(Y \leqslant y) = \Pr(T^2 \leqslant y) = \Pr(-\sqrt{y} \leqslant T \leqslant \sqrt{y}) \\
&= \Pr(T \leqslant \sqrt{y}) - \Pr(T \leqslant -\sqrt{y}) = F_T(\sqrt{y}) - F_T(-\sqrt{y}) \\
f_Y(y) &= \frac{\partial F_Y(y)}{\partial y} = f_T(\sqrt{y})\frac{1}{2\sqrt{y}} + f_T(-\sqrt{y})\frac{1}{2\sqrt{y}} \\
&= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)n^{1/2}} y^{1/2-1}\left(1+\frac{y}{n}\right)^{-(n+1)/2}, \quad y \geqslant 0.
\end{aligned}
$$

Portanto, $Y \sim \mathcal{F}_{1,m}$.

Distribuição t-Student multivariada

Seja $\boldsymbol{X}$ um vetor aleatório p-dimensional com distribuição t-Student multivariada com (parâmetros) **vetor de localização μ**, **matriz de escala $\boldsymbol{\Sigma}$** e **graus de liberdade ν**. Então, a função de densidade de probabilidade é dada por

$$f(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi^{p/2}\nu^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\left(1+\frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{-(\nu+p)/2},$$

para $\boldsymbol{x}\in\mathbb{R}^p$, com $\boldsymbol{\mu}\in\mathbb{R}^p$, $\boldsymbol{\Sigma}$ uma matriz $p\times p$ simétrica e positiva definida, e $\nu>0$. Temos que

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}) &= \boldsymbol{\mu}, && \text{se} \quad \nu>1 \\ \mathrm{Var}(\boldsymbol{X}) &= \frac{\nu}{\nu-2}\boldsymbol{\Sigma}, && \text{se} \quad \nu>2. \end{aligned}$$

Denotaremos a distribuição t-Student multivariada por $t_\nu(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Distribuição Wishart

Seja $\boldsymbol{X}$ uma matriz aleatória $p \times p$ com distribuição Wishart com (parâmetros) **matriz de escala $\boldsymbol{V}$** e **graus de liberdade ν**. Então, a função de densidade de probabilidade é dada por

$$f(\boldsymbol{X}|\boldsymbol{V},\nu) = \frac{|\boldsymbol{X}|^{(\nu-p-1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})\right\}}{2^{(\nu p)/2}|\boldsymbol{V}|^{\nu/2}\pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\left(\frac{\nu}{2}+\frac{(1-j)}{2}\right)},$$

para $\boldsymbol{X}$ no espaço das matrizes $p \times p$, simétricas e positiva definida. Além disso, $\boldsymbol{V}$ é uma matriz $p \times p$ simétrica e positiva definida, e $\nu \geqslant p$. Temos que

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}) &= \nu\boldsymbol{V} \\ \mathrm{Var}(X_{jk}) &= \nu(v_{jk}^2 + v_{jj}v_{kk}), \quad j,k = 1,2,\ldots,p. \end{aligned}$$

Denotaremos a distribuição Wishart por $\mathcal{W}(\boldsymbol{V},\boldsymbol{\nu})$.

Resultado

Seja $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ uma amostra aleatória de $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tal que $n \geqslant p$. Então,

$$\sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\top \sim \mathcal{W}(\boldsymbol{\Sigma}, n).$$

Resultado

Se $\boldsymbol{X} \sim \mathcal{W}(\boldsymbol{V}, \nu)$ e $\boldsymbol{Y} = \boldsymbol{X}^{-1}$, então $\boldsymbol{Y} \sim \mathcal{IW}(\boldsymbol{U}, \nu)$ - chamada de **inversa Wishart** com $\boldsymbol{U} = \boldsymbol{V}^{-1}$.

A função de densidade de probabilidade de $\boldsymbol{Y}$ é dada por

$$f(\boldsymbol{Y}|\boldsymbol{U}, \nu) = \frac{|\boldsymbol{U}|^{\nu/2} |\boldsymbol{Y}|^{-(\nu+p+1)/2} \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{U}\boldsymbol{Y}^{-1})\right\}}{2^{(\nu p)/2} \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{\nu}{2} + \frac{(1-j)}{2}\right)},$$

Temos que

$$\text{E}(\boldsymbol{Y}) = \frac{\boldsymbol{U}}{\nu - p - 1}, \quad \text{para} \quad \nu > p + 1.$$

Normal multivariada: inferência sobre o vetor de médias

Considere o problema univariado no qual se tem uma amostra aleatória de tamanho n, $(X_1, X_2, ..., X_n)$, da distribuição $\mathcal{N}(\mu, \sigma^2)$, em que ambos os parâmetros - média e variância - são desconhecidos.

A estatística dada por

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}}$$

tem distribuição t-Student com $n - 1$ graus de liberdade.

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ e $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ são as estatísticas utilizadas.

Podemos tomar o quadrado da estatística t na forma

$$T^2 = \sqrt{n}(\overline{X} - \mu)^\top \left(S^2\right)^{-1} \sqrt{n}(\overline{X} - \mu).$$

Neste caso, a distribuição T^2 é uma F-Snedecor com 1 e $n-1$ graus de liberdade.

Notação: $T^2 \sim \mathcal{F}_{1,n-1}$.

Esta última expressão nos sugere a seguinte versão multivariada desta estatística:

$$T^2 = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}),$$

em que

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \quad \text{e} \quad \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^\top,$$

com $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ sendo uma amostra aleatória da $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

A estatística T^2 é chamada Estatística T^2 de Hotelling em homenagem a Harold Hotelling.

Proposição

Sejam $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ uma amostra aleatória da distribuição $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ e $T^2 = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu})$. Então,

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

Natureza da estatística T^2:

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

em que

$$\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}) \quad \text{e} \quad \sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^\top$$

são independentes.

Teste de hipóteses

Considere o problema de testar as hipóteses

$$\begin{cases} \mathrm{H}_0: & \boldsymbol{\mu} = \boldsymbol{\mu}_0 \\ \mathrm{H}_1: & \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \end{cases}$$

quando se tem uma amostra aleatória da distribuição normal multivariada.

A estatística de teste neste caso será dada por

$$T^2 = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}).$$

Sob H_0, a estatística T^2 tem distribuição $\dfrac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}$ de acordo com a Proposição anterior. Assim, ao nível de significância α, rejeitamos H_0 se

$$T^2_{obs} > \frac{(n-1)p}{n-p}F_{p,n-p,1-\alpha} \quad \text{para} \quad T^2_{obs} = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}),$$

com $F_{p,n-p,1-\alpha}$ representando o quantil de $100(1-\alpha)\%$ da distribuição $\mathcal{F}_{p,n-p}$.

Notação: $\boldsymbol{S}$ representa a matriz aleatória e também sua realização (o valor observado). Em geral, omitiremos $\{obs\}$ de T^2_{obs}, isto é, utilizaremos só T^2.

Exemplo

A transpiração de 20 mulheres saudáveis foi analisada produzindo X_1 - taxa de suor, X_2 - conteúdo de sódio e X_3 - conteúdo de potássio. Deseja-se testar, ao nível de significância de 10%, as hipóteses

$$\begin{cases} H_0: & \boldsymbol{\mu} = (4, 50, 10)^\top \\ H_1: & \boldsymbol{\mu} \neq (4, 50, 10)^\top. \end{cases}$$

Lembrem-se antes de verificar a normalidade dos dados!

Após os cálculos, obteve-se $T^2 \simeq 9,74$ e $RC = \{T^2 : T^2 > 8,17\}$.

Logo, ao nível de significância de 10%, rejeitamos a hipótese nula.

Qual é o valor p deste teste? Calcule $T^2(n-p)/(p(n-1))$ e obtenha a probabilidade da cauda superior da distribuição $\mathcal{F}_{p,n-p}$ associada a este valor.

Temos, $p = 0,0649$. Portanto, para qualquer nível de significância menor que 6,5%, H_0 não seria rejeitada.

Ver arquivo exemplo_07.r

Invariância da estatística T^2

A estatística T^2 é invariante sob transformações de localização e escala.

Defina $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X} + \boldsymbol{d}$, com $\boldsymbol{X}$ $(p \times 1)$, $\boldsymbol{Y}$ $(p \times 1)$ vetores aleatórios, $\boldsymbol{d}$ vetor $(p \times 1)$ de constantes fixadas e $\boldsymbol{C}$ matriz $(p \times p)$ não-singular de constantes fixadas.

Então, $\overline{\boldsymbol{Y}} = \boldsymbol{C}\overline{\boldsymbol{X}} + \boldsymbol{d}$ e $\boldsymbol{S}_{\boldsymbol{Y}} = \boldsymbol{C}\boldsymbol{S}_{\boldsymbol{X}}\boldsymbol{C}^\top$. Além disso, $\boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{C}\boldsymbol{\mu}_{\boldsymbol{X}} + \boldsymbol{d}$.

Assim,

$$
\begin{aligned}
T^2_{\boldsymbol{Y}} &= n(\overline{\boldsymbol{Y}} - \boldsymbol{\mu}_{\boldsymbol{Y}})^\top \boldsymbol{S}_{\boldsymbol{Y}}^{-1}(\overline{\boldsymbol{Y}} - \boldsymbol{\mu}_{\boldsymbol{Y}}) \\
&= n(\boldsymbol{C}\overline{\boldsymbol{X}} - \boldsymbol{C}\boldsymbol{\mu}_{\boldsymbol{X}})^\top \left(\boldsymbol{C}\boldsymbol{S}_{\boldsymbol{X}}\boldsymbol{C}^\top\right)^{-1}(\boldsymbol{C}\overline{\boldsymbol{X}} - \boldsymbol{C}\boldsymbol{\mu}_{\boldsymbol{X}}) \\
&= n(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top \underbrace{\boldsymbol{C}^\top(\boldsymbol{C}^\top)^{-1}}_{\boldsymbol{I}_p} \boldsymbol{S}_{\boldsymbol{X}}^{-1} \underbrace{\boldsymbol{C}^{-1}\boldsymbol{C}}_{\boldsymbol{I}_p}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_{\boldsymbol{X}}) \\
&= n(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top \boldsymbol{S}_{\boldsymbol{X}}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_{\boldsymbol{X}}) \\
&= T^2_{\boldsymbol{X}}.
\end{aligned}
$$

A estatística T^2 e o teste de razão de verossimilhanças

Uma metodologia muito usada na construção de testes de hipóteses é conhecida como teste de razão de verossimilhanças **(teste RV)**.

Em linhas gerais, se temos uma amostra aleatória de uma distribuição que depende de um vetor de parâmetros $\boldsymbol{\theta}$ cuja densidade é $f_n(\boldsymbol{x}|\boldsymbol{\theta})$ e desejamos testar

$$\begin{cases} \mathrm{H}_0: & \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \\ \mathrm{H}_1: & \boldsymbol{\theta} \notin \boldsymbol{\Theta}_0, \end{cases}$$

a estatística do teste RV é dada por

$$\Lambda(\boldsymbol{x}) = \frac{\max_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}|\boldsymbol{x})}{\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{x})}$$

com $L(\boldsymbol{\theta}|\boldsymbol{x})$ a função de verossimilhança.

Observe que a estatística $\Lambda(\boldsymbol{x})$ é um número entre 0 e 1.

Se o máximo sob H_0 for o máximo global, teremos $\Lambda(\boldsymbol{x}) = 1$.

Caso contrário, $\Lambda(\boldsymbol{x}) < 1$ e como $\Lambda(\boldsymbol{x})$ representa uma razão entre quantidades positivas, segue que $0 < \Lambda(\boldsymbol{x}) \leqslant 1$.

Assim, é razoável dizer que a hipótese nula será rejeitada para valores pequenos de $\Lambda(\boldsymbol{x})$ tal que as regiões críticas nos testes RV são da forma $\Lambda(\boldsymbol{x}) \leqslant c$.

Para obter o valor de c é necessário conhecer a distribuição amostral de $\Lambda(\boldsymbol{x})$. Esta distribuição nem sempre é fácil de ser obtida e muitas vezes são necessários métodos aproximados para avaliar o valor de c para um dado nível de significância.

Distribuição assintótica da estatística Λ

Temos que

$$-2 \ln \Lambda \overset{a}{\sim} \chi^2_{\nu - \nu_0}$$

em que ν é a dimensão do espaço de parâmetros e ν_0 é a dimensão do sub-espaço correspondente à hipótese nula.

Calcularemos a estatística $\Lambda(\boldsymbol{x})$ no contexto do teste das hipóteses

$$\begin{cases} \mathrm{H}_0 : & \boldsymbol{\mu} = \boldsymbol{\mu}_0 \\ \mathrm{H}_1 : & \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \end{cases}$$

quando se tem uma amostra aleatória da distribuição normal multivariada.

O máximo global da função de verossimilhança é obtido quando

$$\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{X}} \quad \text{e} \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^\top = \frac{n-1}{n}\boldsymbol{S}$$

e dado por

$$L(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = (2\pi)^{-np/2}|\widehat{\boldsymbol{\Sigma}}|^{-n/2}\mathrm{e}^{-np/2}.$$

Portanto, já temos o denominador da estatística $\Lambda(\boldsymbol{x})$, neste caso.

Para obter o numerador, observe que sob H_0 há apenas um valor possível para $\boldsymbol{\mu}$ dado por $\boldsymbol{\mu}_0$. Neste caso, usando os resultados apresentados na seção de estimação por máxima verossimilhança, é fácil ver que a matriz que maximiza a verossimilhança sob H_0 é dada por

$$\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^\top.$$

Assim, o numerador de $\Lambda(\boldsymbol{x})$ é

$$L(\boldsymbol{\mu}_0, \widehat{\boldsymbol{\Sigma}}_0) = (2\pi)^{-np/2}|\widehat{\boldsymbol{\Sigma}}_0|^{-n/2}\mathrm{e}^{-np/2}.$$

Logo,

$$\Lambda = \left(\frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_0|} \right)^{n/2} = \left(\frac{\left| \sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^\top \right|}{\left| \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^\top \right|} \right)^{n/2}.$$

A estatística equivalente $\Lambda^{2/n} = |\widehat{\Sigma}|/|\widehat{\Sigma}_0|$ é chamada **lambda de Wilks**.

é fácil, por meio de artifícios algébricos, mostrar que as estatísticas T^2 e Λ para o teste aqui considerado satisfazem a relação

$$\Lambda^{2/n} = \left(1 + \frac{T^2}{n-1} \right)^{-1}$$

e, usando esta relação, chegamos a uma outra expressão para o cálculo da estatística T^2:

$$T^2 = \frac{(n-1)|\hat{\boldsymbol{\Sigma}}_0|}{|\hat{\boldsymbol{\Sigma}}|} - (n-1).$$

Ver arquivo exemplo_07.r

Regiões de confiança e intervalos simultâneos

Dizemos que $R(\boldsymbol{X})$ é uma região de $100(1-\alpha)\%$ de confiança para $\boldsymbol{\theta}$ se

$$\Pr\left(R(\boldsymbol{X}) \text{ compreender } \boldsymbol{\theta}\right) = 1-\alpha.$$

A região de confiança para o vetor de médias $\boldsymbol{\mu}$ quando se dispõe de uma amostra aleatória da distribuição $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ é dada por

$$n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{\mu}) \leqslant \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha},$$

em que $F_{p,n-p,1-\alpha}$ é o quantil acumulado de $100(1-\alpha)\%$ da distribuição $\mathcal{F}_{p,n-p}$.

Este resultado é obtido se utilizando a distribuição amostral da estatística T^2 apresentada anteriormente.

Observe que a região de confiança é dada pelo hiper-elipsóide de eixos determinados pelos autovetores da matriz de covariância amostral $\boldsymbol{S}$ e cujas medidas são proporcionais às raizes quadradas dos respectivos autovalores.

Para verificar se um dado vetor $\boldsymbol{\mu}_0$ pertence à região de confiança, basta calcular $n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)$ e comparar com $\dfrac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$.

Para $p \geqslant 4$ não é possível representar visualmente a região de confiança. No entanto, podemos calcular as medidas dos eixos do hiper-elipsóide de confiança centrado em $\overline{\boldsymbol{x}}$:

$$n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \leqslant c^2 = \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}.$$

Os semi-eixos têm medida

$$\sqrt{\lambda_j}\frac{c}{\sqrt{n}} = \sqrt{\lambda_j \frac{(n-1)p}{n(n-p)} F_{p,n-p,1-\alpha}}.$$

Exemplo

O departamento de controle de qualidade de uma fabricante de fornos de microondas foi cobrado pelo governo federal a monitorar a quantidade de radiação quando as portas dos microondas são fechadas. Observações da radiação emitida através das portas fechadas de $n = 42$ fornos selecionados ao acaso foram feitas. Medidas de radiação também foram feitas com as portas abertas dos 42 fornos selecionados.

Se verificarmos a suposição de normalidade, veremos que esta não é apropriada e uma transformação potência dos dados é buscada.

Sugere-se trabalhar com a potência $0,25$, ou seja, a raiz quarta da escala original da medida de radiação.

Pede-se construir uma elipse de 95% de confiança para o vetor de médias, considerando a escala dos dados transformados de modo que a suposição de normalidade seja razoável.

Ver arquivo exemplo_08.r

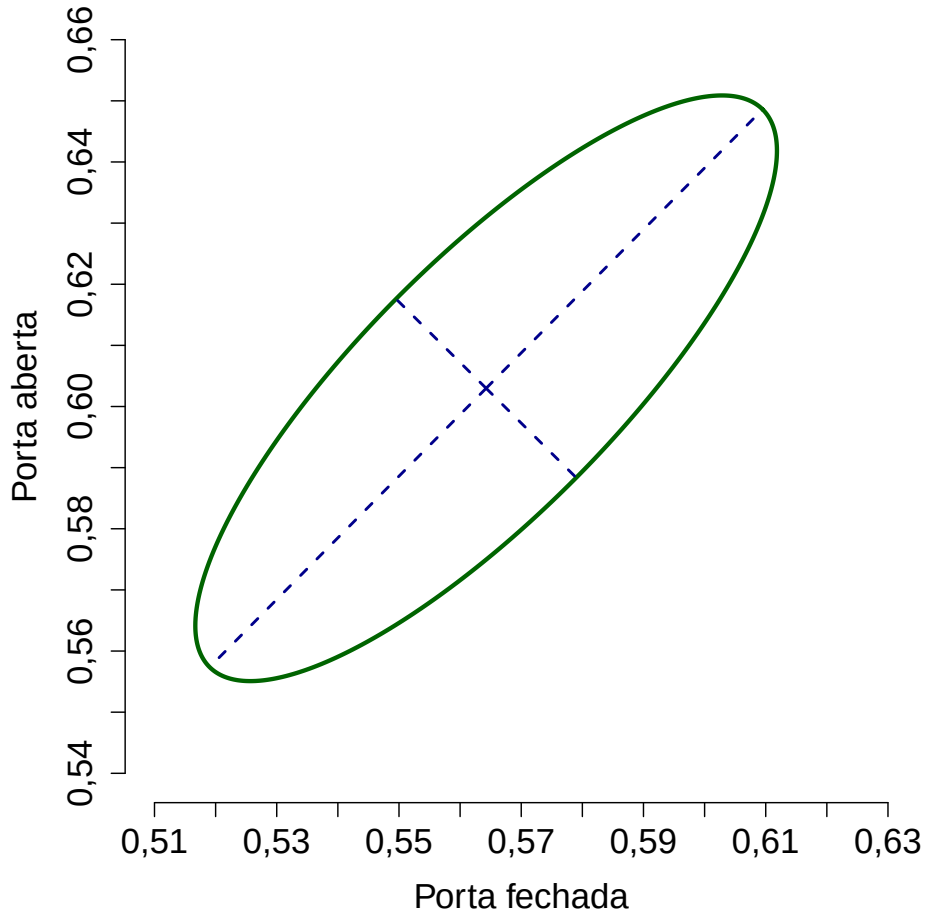

Figura 4: Região de 95% de confiança para o vetor de médias, μ, das radiações de microondas de portas abertas e fechadas, respectivamente.

Intervalos de confiança simultâneos para as componentes do vetor de médias

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

Vimos que se $\boldsymbol{a}$ é um vetor de constantes em $\mathbb{R}^p$, então

$$Z = \boldsymbol{a}^\top \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}).$$

Logo, se $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ é uma amostra aleatória da $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, segue que $Z_1, Z_2, \ldots, Z_n$, definidos por $Z_i = \boldsymbol{a}^\top \boldsymbol{X}_i$, $i = 1, \ldots, n$, é uma amostra aleatória da $\mathcal{N}(\underbrace{\boldsymbol{a}^\top \boldsymbol{\mu}}_{\mu_Z}, \underbrace{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}}_{\sigma^2_Z})$.

Da teoria normal univariada, temos que um intervalo de $100(1-\alpha)\%$ de confiança para $\mu_Z = \boldsymbol{a}^\top \boldsymbol{\mu}$ é dado por

$$\text{IC}_{100(1-\alpha)\%}(\mu_Z) = \left(\boldsymbol{a}^\top \overline{\boldsymbol{x}} - t_{n-1,1-\frac{\alpha}{2}} \sqrt{\frac{\boldsymbol{a}^\top \boldsymbol{S} \boldsymbol{a}}{n}};\ \boldsymbol{a}^\top \overline{\boldsymbol{x}} + t_{n-1,(1-\frac{\alpha}{2})} \sqrt{\frac{\boldsymbol{a}^\top \boldsymbol{S} \boldsymbol{a}}{n}} \right)$$

Claramente, poderíamos construir vários intervalos de confiança sobre combinações lineares dos componentes do vetor $\boldsymbol{\mu}$, cada um associado com um coeficiente de confiança $1 - \alpha$, escolhendo diferentes vetores de constantes $\boldsymbol{a}$. Porém, o coeficiente de confiança conjunto do conjunto de intervalos resultantes não seria mais $1 - \alpha$.

É desejável associar um coeficiente de confiança **coletivo** de $1 - \alpha$ aos intervalos de confiança que podem ser gerados para todas as escolhas de $\boldsymbol{a}$.

Naturalmente, um preço deverá ser pago pela conveniência de uma confiança simultânea grande para todos os intervalos: intervalos que têm amplitudes maiores (mais largos, menos precisos) do que os intervalos apresentados anteriormente via a distribuição amostral t-Student com $n - 1$ graus de liberdade.

Dado o conjunto de dados observados $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ e um $\boldsymbol{a}$ particular

$$|t| = \frac{\left|\sqrt{n}(\boldsymbol{a}^\top \overline{\boldsymbol{x}} - \boldsymbol{a}^\top \boldsymbol{\mu})\right|}{\sqrt{\boldsymbol{a}^\top \boldsymbol{S}\boldsymbol{a}}} \leqslant t_{n-1,1-\frac{\alpha}{2}}$$

ou, equivalentemente,

$$t^2 = \frac{n(\boldsymbol{a}^\top \overline{\boldsymbol{x}} - \boldsymbol{a}^\top \boldsymbol{\mu})^2}{\boldsymbol{a}^\top \boldsymbol{S}\boldsymbol{a}} \leqslant t^2_{n-1,1-\frac{\alpha}{2}}.$$

Uma região de confiança simultânea é dada para o conjunto de valores $\boldsymbol{a}^\top \boldsymbol{\mu}$ tais que t^2 é relativamente pequeno para todas as escolhas de $\boldsymbol{a}$.

Parece razoável esperar que o valor

$$t^2_{n-1,1-\frac{\alpha}{2}}$$

seja substituído por um valor maior, c^2, quando afirmações são feitas para muitas escolhas de $\boldsymbol{a}$.

Considerando os valores de $\boldsymbol{a}$ para os quais $t^2 \leqslant c^2$, somos naturalmente levados a

$$\max_{\boldsymbol{a}} t^2(\boldsymbol{a}) = \max_{\boldsymbol{a}} \frac{n(\boldsymbol{a}^\top \overline{\boldsymbol{x}} - \boldsymbol{a}^\top \boldsymbol{\mu})^2}{\boldsymbol{a}^\top \boldsymbol{S}\boldsymbol{a}}.$$

Usando os resultados sobre desigualdades, percebemos que o máximo ocorrerá para

$$\boldsymbol{a} \propto \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}).$$

Ora, isto nos levará à estatística T^2.

Por conveniência, costuma se referir a esses intervalos como intervalos-T^2.

Em particular, tomando os vetores $\boldsymbol{a}$'s como os vetores da base canônica do $\mathbb{R}^p$, obtém-se

$$\text{IC}_{100(1-\alpha)\%}(\mu_j) = \left(\overline{x}_j - \xi_{n,p,1-\alpha}\sqrt{\frac{s_{jj}}{n}};\ \overline{x}_j + \xi_{n,p,1-\alpha}\sqrt{\frac{s_{jj}}{n}} \right), \quad j = 1, 2, \ldots, p,$$

em que

$$\xi_{n,p,1-\alpha} = \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p,1-\alpha}}.$$

Agora os intervalos-T^2 coletivamente têm nível de confiança $1-\alpha$.

Observe que também podemos construir intervalos de confiança para relações estruturais entre os componentes do vetor $\boldsymbol{\mu}$ como, por exemplo, intervalos para as diferenças entre os componentes de $\boldsymbol{\mu}$.

Fazendo

$$\boldsymbol{a}^\top = (0,\ldots,0,\underbrace{1}_{j\text{-ésima entrada}},0,\ldots,0,\underbrace{-1}_{k\text{-ésima entrada}},0,\ldots,0),$$

teremos $\boldsymbol{a}^\top\boldsymbol{\mu} = \mu_j - \mu_k$, $\boldsymbol{a}^\top\overline{\boldsymbol{x}} = \overline{x}_j - \overline{x}_k$ e $\boldsymbol{a}^\top\boldsymbol{S}\boldsymbol{a} = s_{jj} + s_{kk} - 2s_{jk}$.

O intervalo para a diferença $\mu_j - \mu_k$ será dado por

$$(\overline{x}_j - \overline{x}_k) \pm \sqrt{\frac{p(n-1)}{n-p}F_{p,n-p,1-\alpha}}\sqrt{\frac{s_{jj}+s_{kk}-2s_{jk}}{n}}.$$

Exemplo

Obtivemos uma elipse de 95% de confiança para o vetor $\boldsymbol{\mu}$ nos dados sobre radiação em microondas. Pede-se construir os intervalos-T^2 de 95% de confiança para os componentes individuais do vetor $\boldsymbol{\mu}$, identificando-os como as "sombras" da elipse de 95% de confiança sobre os eixos coordenados. Pede-se também construir os intervalos baseados na distribuição t-Student e compará-los com os correspondentes intervalos T^2.

Ver arquivo exemplo_09.r

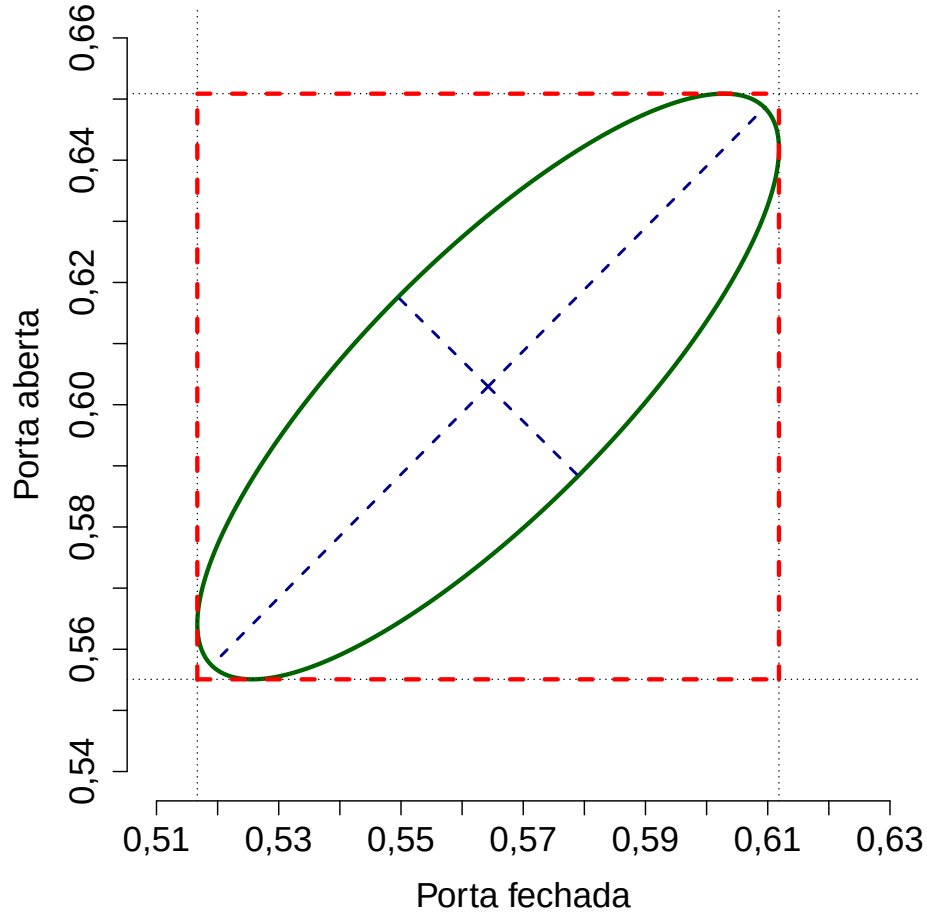

Figura 5: Intervalos T^2 (vermelho) e região (verde) de 95% de confiança para o vetor de médias, $\boldsymbol{\mu}$, das radiações de microondas de portas abertas e fechadas, respectivamente.

Uma comparação entre os intervalos-T^2 e os intervalos separados

A tabela a seguir mostra uma comparação entre os comprimentos dos intervalos de confiança separados e os intervalos "simultâneos" T^2 para alguns valores selecionados de p, n e $\alpha = 0,05$.

n	$t_{n-1}(0,975)$	$\xi_{n;p;0,95}$	
		$p = 4$	$p = 10$
15	2,145	4,14	11,52
25	2,064	3,60	6,39
50	2,010	3,31	5,05
100	1,970	3,19	4,61
∞	1,960	3,08	4,28

Os valores nas duas últimas colunas da tabela correspondem a

$$\xi_{n;p;0,95} = \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p}(0,95)}.$$

A comparação feita é imprópria, pois o nível de confiança associado a qualquer coleção de intervalos-T^2, para p fixado, é 0,95, e o nível global associado com uma coleção de intervalos separados via distribuição t-Student deve ser menor do que 0,95.

Uma outra abordagem, conhecida como **método de Bonferroni de comparações múltiplas**, será considerada. O método leva este nome devido à desigualdade de Bonferroni.

Seja $A_1, A_2, \ldots, A_m$ uma coleção de eventos em um espaço de probabilidade tais que $\Pr(A_\ell) = 1 - \alpha_\ell$, $\ell = 1, \ldots, m$. Então,

$$\begin{aligned}
\Pr\left(\cap_{\ell=1}^{m} A_\ell\right) &= 1 - P([\cap_{\ell=1}^{m} A_\ell]^c) \\
&= 1 - P(\cup_{\ell=1}^{m} A_\ell^c) \\
&\geqslant 1 - \sum_{\ell=1}^{m} \Pr(A_\ell^c) \\
&= 1 - \sum_{\ell=1}^{m} \alpha_\ell \\
&= 1 - (\alpha_1 + \alpha_2 + \cdots + \alpha_m).
\end{aligned}$$

A desigualdade de Bonferroni permite ao investigador controlar a taxa de erro

$$\alpha_1 + \alpha_2 + \cdots + \alpha_m,$$

sem olhar a estrutura de correlação por trás dos intervalos de confiança.

Assim, se o problema envolve a construção de m intervalos importantes, a ideia é fazer

$$\alpha_\ell = \alpha/m, \quad \text{para} \quad \ell = 1, \ldots, m,$$

e tomar os intervalos separados dados por

$$\boldsymbol{a}^\top \overline{\boldsymbol{x}} \pm t_{n-1,1-\frac{\alpha}{2m}} \sqrt{\frac{\boldsymbol{a}^\top \boldsymbol{S}\boldsymbol{a}}{n}}.$$

Observe que agora vale que o coeficiente coletivo de confiança é pelo menos

$$1 - \left(\underbrace{\frac{\alpha}{m} + \frac{\alpha}{m} + \cdots + \frac{\alpha}{m}}_{m \text{ termos}} \right) = 1 - \alpha.$$

Portanto, com um coeficiente de confiança global de pelo menos $1-\alpha$, podemos construir os seguintes $m = p$ intervalos para os componentes do vetor $\boldsymbol{\mu}$:

$$\text{IC}_{100(1-\alpha)\%}(\mu_j) = \left(\overline{x}_j - t_{n-1,1-\frac{\alpha}{2p}} \sqrt{\frac{s_{jj}}{n}};\ \overline{x}_j + t_{n-1,1-\frac{\alpha}{2p}} \sqrt{\frac{s_{jj}}{n}} \right),\ j = 1, 2, \ldots, p.$$

Então, esses intervalos podem, de forma mais apropriada, ser comparados aos intervalos-T^2.

Exemplo

Usando novamente os dados sobre radiação em fornos de microondas, pede-se comparar os intervalos-T^2 para os componentes do vetor de médias com os intervalos via Bonferroni.

Ver arquivo exemplo_10.r

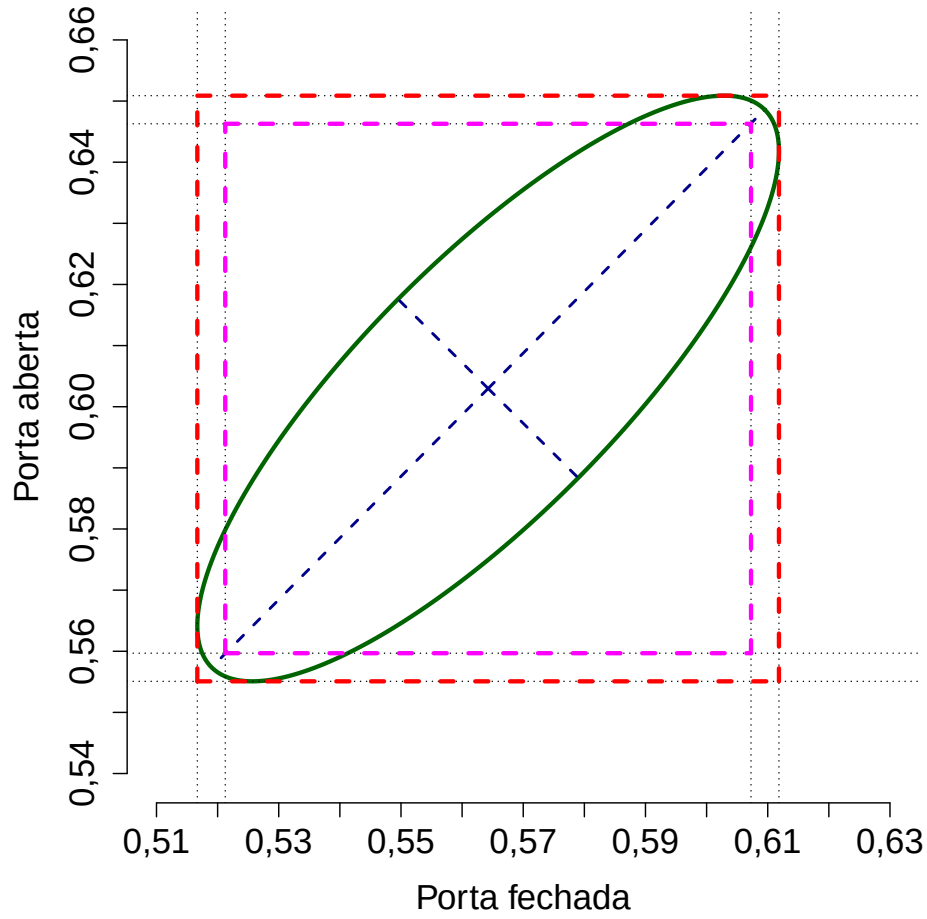

Figura 6: Intervalos T^2 (vermelho), via Bonferroni (magenta) e região (verde) de 95% de confiança para o vetor de médias, $\boldsymbol{\mu}$, das radiações de microondas de portas abertas e fechadas, respectivamente.

A tabela a seguir ilustra uma comparação entre os comprimentos dos intervalos via Bonferroni e T^2 para alguns valores selecionados de p, $m = p$, n e $\alpha = 0,05$. As entradas nas três colunas referentes aos diferentes valores de p selecionados representam a razão entre o comprimento do intervalo via Bonferroni e o comprimento do intervalo-T^2.

n	p		
	2	4	10
15	0,88	0,69	0,29
25	0,90	0,75	0,48
50	0,91	0,78	0,58
100	0,91	0,80	0,62
∞	0,91	0,81	0,66

Podemos ver desta tabela que **os intervalos via Bonferroni produzem intervalos mais estreitos quando** $m = p$.

Devido à facilidade de aplicação e aos resultados mais eficientes em termos de estimação, geralmente **é preferível usar os intervalos simultâneos via Bonferroni**.

Inferências sobre um vetor de médias para grandes amostras

Quando o tamanho da amostra é grande, testes de hipóteses e regiões de confiança para μ podem ser construídos mesmo que a população subjacente não seja normal.

De fato, desvios fortes de uma população normal podem ser superados para tamanhos amostrais grandes.

Ambos, testes de hipóteses e intervalos de confiança simultâneos, terão níveis nominais aproximados.

As vantagens associadas com grandes amostras podem ser parcialmente compensadas por uma perda de informação amostral causada pelo uso somente das estatísticas sumário $\overline{\boldsymbol{X}}$ e $\boldsymbol{S}$.

Por outro lado, como $(\overline{\boldsymbol{X}}, \boldsymbol{S})$ é uma estatística suficiente para populações normais, quanto mais próximas da normal multivariada forem as distribuições das populações, mais eficientemente a informação amostral será utilizada ao fazer inferências.

Todas as inferências sobre μ quando se tem grandes amostras são baseadas na distribuição de qui-quadrado.

Proposição

Seja $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ uma amostra aleatória observada de uma população com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias positiva definida $\boldsymbol{\Sigma}$. Quando $n - p$ é grande, a hipótese $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ é rejeitada em favor de $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, ao nível de significância α, se

$$n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \chi^2_{p,1-\alpha}.$$

Comparando este teste com o obtido via teoria normal, no início destas notas, vemos que a estatística de teste é a mesma, o que muda é o valor crítico.

Um exame mais minucioso revela, porém, que ambos os testes produzirão os mesmos resultados em situações nas quais o teste χ^2 é apropriado (amostras realmente grandes).

De fato, $\dfrac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}$ e $\chi^2_{p,1-\alpha}$ são aproximadamente iguais para $n >> p$.

Proposição

Seja $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ uma amostra aleatória observada de uma população com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias positiva definida $\boldsymbol{\Sigma}$. Se $n - p$ é grande,

$$\boldsymbol{a}^\top \overline{\boldsymbol{x}} \pm \sqrt{\chi^2_{p,1-\alpha}} \sqrt{\frac{\boldsymbol{a}^\top \boldsymbol{S} \boldsymbol{a}}{n}}$$

compreederá $\boldsymbol{a}^\top \boldsymbol{\mu}$, para todo $\boldsymbol{a}$ com probabilidade aproximadamente $1 - \alpha$.

Consequentemente, podemos fazer afirmações simultâneas para os p componentes do vetor de médias dadas por

$$\overline{x}_j \pm \sqrt{\chi^2_{p,1-\alpha}} \sqrt{\frac{s_{jj}}{n}}, \qquad j = 1, 2, \ldots, p.$$

Observação: elipses de confiança para pares de componentes também podem ser facilmente construídas.

A questão de quão grande deve ser o tamanho da amostra não é simples de ser respondida.

Em uma ou duas dimensões, tamanhos amostrais em torno de 30 a 50 podem geralmente ser considerados grandes.

A medida que o número de características torna-se maior, certamente tamanhos amostrais maiores serão exigidos para que as distribuições assintóticas forneçam boas aproximações das verdadeiras distribuições das várias estatísticas de teste.

Na falta de estudos definitivos, propõe-se que $n - p$ deve ser grande, reconhecendo que o caso real pode ser mais complicado do que isso.

Uma aplicação com $p = 2$ e $n = 50$ é muito diferente de uma aplicação com $p = 52$ e $n = 100$ apesar de ambas apresentarem $n - p = 48$.

Deve-se realizar as mesmas verificações exigidas para os métodos baseados na normal.

Apesar de pequenos desvios da normalidade não causarem quaisquer dificuldades para *n* grande, desvios extremos podem causar problemas.

Especificamente, a taxa de erro verdadeira pode estar bem afastada do nível nominal α.

Se, com base nos *qq-plots* e outros esquemas de investigação pontos extremos (*outliers*) e outras formas de desvios extremos aparecem, ações corretivas apropriadas, incluindo transformações, são desejáveis.

Análise de variância multivariada (MANOVA)

Parte 1: Comparações emparelhadas

Começaremos com uma breve revisão deste problema no caso univariado. O problema aqui pode ser descrito da seguinte forma.

Seja $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ uma amostra aleatória de tamanho n de uma população bivariada.

Aqui X e Y podem representar medidas antes e depois que um certo "tratamento" foi aplicado sobre a unidade experimental, por exemplo.

O objetivo aqui é comparar se não há diferenças entre tratamentos ou se não há efeito de tratamento, quando uma das observações pode ser considerada controle.

É interessante notar que neste problema, apesar de usarmos um teste t-Student, a suposição de normalidade bivariada não é necessária.

Basta supor $D_i = (Y_i - X_i) \sim \mathcal{N}(\theta, \psi^2), \qquad i = 1, 2, \ldots, n.$

Sejam $\mathrm{E}(X_i) = \mu_1$, $\mathrm{Var}(X_i) = \sigma_1^2$, $\mathrm{E}(Y_i) = \mu_2$, $\mathrm{Var}(Y_i) = \sigma_2^2$ e $\mathrm{Cov}(X_i, Y_i) = \sigma_{12}$.

Para unidades experimentais diferentes se tem $\mathrm{Cov}(X_i, Y_r) = 0, i \neq r$, pois as diferentes observações são independentes.

Segue que, $\theta = \mathrm{E}(D_i) = \mu_2 - \mu_1$ e $\psi^2 = \mathrm{Var}(D_i) = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$.

Como $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ é uma amostra aleatória, segue que $D_1, D_2, \ldots, D_n$ é uma amostra aleatória de tamanho n de uma população com média θ e variância ψ^2. Logo,

$$D_i \overset{ind}{\sim} \mathcal{N}(\theta, \psi^2), \quad i = 1, 2, \ldots, n \quad \text{e} \quad t = \frac{\sqrt{n}(\overline{D} - \theta)}{S_D} \sim t_{n-1},$$

com $\overline{D} = \dfrac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)$ e $S_D^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$.

Se desejamos testar as hipóteses $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$, rejeitaremos H_0, ao nível de significância α se

$$| t | = \frac{| \overline{d} |}{s_D/\sqrt{n}} \geqslant t_{n-1,1-\alpha/2}.$$

Equivalentemente, um intervalo de $100(1-\alpha)\%$ de confiança para θ é dado por

$$\text{IC}_{100(1-\alpha)\%}(\theta) = \left(\overline{d} - t_{n-1,1-\alpha/2} \frac{s_D}{\sqrt{n}};\ \overline{d} + t_{n-1,1-\alpha/2} \frac{s_D}{\sqrt{n}} \right).$$

O que muda no caso multivariado?

Neste caso, os elementos dos pares ordenados que representam a amostra aleatória, em vez de medidas escalares serão vetores em $\mathbb{R}^p$.

Notação adicional para a extensão multivariada é necessária.

Assim, suponha que observamos uma amostra $((\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n))$ de $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, em que $\boldsymbol{X}_i$ e $\boldsymbol{Y}_i$ são vetores aleatórios em $\mathbb{R}^p$.

Neste caso, podemos definir os vetores aleatórios diferenças

$$\boldsymbol{D}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i, \qquad i = 1, 2, \ldots, n.$$

Temos que

$$\mathrm{E}(\boldsymbol{D}_i) = \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} = \begin{bmatrix} \mu_{2,1} - \mu_{1,1} \\ \mu_{2,2} - \mu_{1,2} \\ \vdots \\ \mu_{2,p} - \mu_{1,p} \end{bmatrix},$$

em que $\theta_j = \mathrm{E}(D_{ji})$, para $j = 1, 2, \ldots, p$, $\quad \boldsymbol{D}_i^\top = (D_{1i}, D_{2i}, ..., D_{pi})$, e

$$\mathrm{Var}(\boldsymbol{D}_i) = \boldsymbol{\Psi}.$$

Se além disso,

$$\boldsymbol{D}_i \overset{ind}{\sim} \mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Psi}),$$

então, segue que

$$T^2 = n(\overline{\boldsymbol{D}} - \boldsymbol{\theta})^\top \boldsymbol{S}_{\boldsymbol{D}}^{-1} (\overline{\boldsymbol{D}} - \boldsymbol{\theta}) \sim \frac{(n-1)p}{n-p} \mathcal{F}_{p,n-p},$$

em que $\overline{\boldsymbol{D}} = \dfrac{1}{n} \sum_{i=1}^{n} (\boldsymbol{Y}_i - \boldsymbol{X}_i)$ e $\boldsymbol{S}_{\boldsymbol{D}} = \dfrac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{D}_i - \overline{\boldsymbol{D}})(\boldsymbol{D}_i - \overline{\boldsymbol{D}})^\top$.

Logo, no teste das hipóteses $H_0 : \boldsymbol{\theta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\theta} \neq \mathbf{0}$, rejeitaremos H_0, ao nível de significância α se

$$T^2 = n\overline{\boldsymbol{d}}^\top \boldsymbol{S}_{\boldsymbol{D}}^{-1}\overline{\boldsymbol{d}} > \frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}.$$

Uma região de $100(1-\alpha)\%$ de confiança para $\boldsymbol{\theta}$ é dada pelo hiper-elipsóide

$$(\overline{\boldsymbol{d}} - \boldsymbol{\theta})^\top \boldsymbol{S}_{\boldsymbol{D}}^{-1}(\overline{\boldsymbol{d}} - \boldsymbol{\theta}) \leqslant \frac{(n-1)p}{n(n-p)} F_{p,n-p,1-\alpha}.$$

Intervalos-T^2 simultâneos de $100(1-\alpha)\%$ de confiança para os componentes do vetor $\boldsymbol{\theta}$ são dados por

$$\text{IC}_{100(1-\alpha)\%}(\theta_j) = \left(\overline{d}_j - \xi_{n,p,1-\alpha}\sqrt{\frac{s_{Dj}^2}{n}};\ \overline{d}_j + \xi_{n,p,1-\alpha}\sqrt{\frac{s_{Dj}^2}{n}} \right),\ j = 1, 2, \ldots, p,$$

em que

$$\xi_{n,p,1-\alpha} = \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha}}.$$

Para $n - p$ grande,

$$\frac{(n-1)p}{n-p} F_{p,n-p,1-\alpha} \simeq \chi^2_{p,1-\alpha}$$

e a suposição de normalidade não é necessária.

Os intervalos de Bonferroni simultâneos de $100(1-\alpha)\%$ de confiança para os componentes do vetor de médias são dados por

$$\text{IC}_{100(1-\alpha)\%}(\theta_j) = \left(\overline{d}_j - t_{n-1,1-\frac{\alpha}{2p}} \sqrt{\frac{s^2_{Dj}}{n}}, \overline{d}_j + t_{n-1,1-\frac{\alpha}{2p}} \sqrt{\frac{s^2_{Dj}}{n}} \right),$$

para $j = 1, 2, \ldots, p$.

Exemplo

As plantas de tratamento de esgoto municipais devem monitorar suas descargas em rios e afluentes regularmente por exigência de lei. Preocupações com a fidedignidade dos dados de um destes programas de auto-monitoramento levaram a um estudo, no qual amostras de efluentes foram divididas e enviadas a dois laboratórios para teste. Metade de cada amostra foi enviada ao Laboratório de Higiene (LH) do estado de Wisconsin, e a outra metade para um laboratório comercial privado (LC) rotineiramente usado no programa de monitoramento. Medidas do oxigênio bioquímico demandado (obd) e sólidos suspensos (ss) foram obtidas, para $n = 11$ amostras divididas, dos dois laboratórios.

- As análises dos dois laboratórios são compatíveis?
- Se há diferenças, qual é a sua natureza?

Ver arquivo exemplo_11.r

Ao nível de significância de 5%, a hipótese nula de que não há diferença no vetor de médias deve ser rejeitada.

Porém, ao construir os intervalos-T^2 simultâneos de 95% de confiança para os componentes do vetor de médias, verifica-se que o zero pertence a ambos:

$$\text{IC}_{95\%}(\xi_1) = (-22,45;\ 3,73) \quad \text{e} \quad \text{IC}_{95\%}(\xi_2) = (-5,70;\ 32,25).$$

O que concluir?

A evidência aponta para diferenças reais.

O ponto $\boldsymbol{\theta} = \mathbf{0}$ está fora da região de 95% de confiança (elipse).

O coeficiente de confiança simultâneo de 95% se aplica ao conjunto completo de intervalos que poderiam ser construídos para todas as escolhas possíveis de combinações lineares $\boldsymbol{a}^\top \boldsymbol{\theta}$ da forma

$$a_1\theta_1 + a_2\theta_2.$$

Os intervalos particulares apresentados correspondem às escolhas $\boldsymbol{a}^\top = (a_1, a_2)^\top = (1, 0)$ e $\boldsymbol{a}^\top = (a_1, a_2)^\top = (0, 1)$ e contêm o zero.

Outras escolhas de a_1 e a_2 produzirão intervalos simultâneos que provavelmente não contêm o zero.

Se $H_0 : \boldsymbol{\theta} = \mathbf{0}$ não tivesse sido rejeitada, então todos os intervalos-T^2 simultâneos conteriam o zero.

Os intervalos simultâneos via Bonferroni de 95% também cobrem o zero, isto é,

$$\text{IC}_{95\%}(\xi_1) = (-20,57;\ 1,85) \quad \text{e} \quad \text{IC}_{95\%}(\xi_2) = (-2,97;\ 29,52).$$

Neste exemplo, o experimentador dividiu a amostra primeiro agitando a solução e depois colocando-a em duas garrafas para a análise química.

Esta foi uma boa estratégia, pois uma divisão simples da amostra em duas partes obtidas pela primeira metade do topo em uma garrafa e o resto na outra poderia resultar em sólidos suspensos na metade inferior devido à forma como os líquidos foram colocados nas garrafas.

Os dois laboratórios, então, não estariam necessariamente trabalhando com unidades experimentais similares, e as conclusões não pertenceriam à competência do laboratório, técnicas de medição, etc.

Sempre que um investigador puder controlar a designação de tratamentos às unidades experimentais, o emparelhamento apropriado de unidades experimentais ou a aleatorização da designação de tratamentos às unidades, permitirão uma análise estatística dos resultados mais apropriada.

Tratamentos em experimentos de medidas repetidas

Outra generalização da estatística t-Student emparelhada surge em situações nas quais q tratamentos são comparados com respeito a uma única variável resposta.

Cada unidade experimental recebe cada tratamento uma vez sobre sucessivos períodos de tempo. A i-ésima observação é

$$\boldsymbol{X}_i = \begin{bmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{qi} \end{bmatrix}, \quad i = 1, 2, ..., n.$$

Aqui X_{ji} é a resposta do j-ésimo tratamento, $j = 1, 2, ..., q$, sobre a i-ésima unidade experimental, $i = 1, 2, ..., n$.

O nome medidas repetidas deriva do fato de que todos os tratamentos são administrados a cada unidade.

Para propósitos de comparação, podemos considerar os seguintes contrastes dos componentes do vetor de médias $\boldsymbol{\mu}$:

$$\begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_q \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} = \boldsymbol{C}_1 \boldsymbol{\mu}$$

ou

$$\begin{bmatrix} \mu_2 - \mu_1 \\ \mu_3 - \mu_2 \\ \vdots \\ \mu_q - \mu_{q-1} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} = \boldsymbol{C}_2 \boldsymbol{\mu}.$$

Ambas $\boldsymbol{C}_1$ e $\boldsymbol{C}_2$ são chamadas matrizes de contraste, pois suas $q-1$ linhas são linearmente independentes e cada uma delas é um vetor de contraste.

A natureza do planejamento elimina muito da influência da variação unidade a unidade sobre as comparações de tratamento. O experimentador deve aleatorizar a ordem na qual os tratamentos são apresentados a cada unidade experimental.

Quando as médias dos diferentes tratamentos são iguais, $\boldsymbol{C}_1\boldsymbol{\mu} = \boldsymbol{C}_2\boldsymbol{\mu} = \mathbf{0}$.

Em resumo, a hipótese de que não há diferença entre tratamentos se torna $\boldsymbol{C}\boldsymbol{\mu} = \mathbf{0}$ para qualquer escolha adequada de $\boldsymbol{C}$.

Consequentemente, baseados nos contrastes observados $\boldsymbol{C}\boldsymbol{x}_i$ nas observações, teremos média amostral $\boldsymbol{C}\overline{\boldsymbol{x}}$ e matriz de covariâncias amostral $\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top$. Para testar

$$\begin{cases} H_0: & \boldsymbol{C}\boldsymbol{\mu} = \mathbf{0} \\ H_1: & \boldsymbol{C}\boldsymbol{\mu} \neq \mathbf{0} \end{cases},$$

usaremos a estatística[3]

$$T^2 = n(\boldsymbol{C}\overline{\boldsymbol{X}})^\top(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\overline{\boldsymbol{X}}).$$

[3] $\boldsymbol{C}$ não é quadrada! Logo, não simplificamos a expressão de T^2.

Teste para a igualdade de tratamentos em experimentos de medidas repetidas

Sejam

$$\boldsymbol{X}_i \overset{ind}{\sim} \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, 2, \ldots, n,$$

e $\boldsymbol{C}$ a matriz de contraste $(q-1) \times q$. Para

$$\begin{cases} H_0: & \boldsymbol{C\mu} = \boldsymbol{0} \\ H_1: & \boldsymbol{C\mu} \neq \boldsymbol{0} \end{cases},$$

rejeitaremos H_0, a um nível de significância α, se

$$T^2 = n(\boldsymbol{C}\overline{\boldsymbol{x}})^\top (\boldsymbol{CSC}^\top)^{-1} (\boldsymbol{C}\overline{\boldsymbol{x}}) > \frac{(n-1)(q-1)}{n-q+1} F_{q-1,n-q+1,1-\alpha},$$

em que $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ e $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top$.

Região de $100(1-\alpha)\%$ de confiança para $\boldsymbol{C\mu}$:

$$n(\boldsymbol{C\overline{x}}-\boldsymbol{C\mu})^\top(\boldsymbol{CSC}^\top)^{-1}(\boldsymbol{C\overline{x}}-\boldsymbol{C\mu}) \leqslant \frac{(n-1)(q-1)}{n-q+1}F_{q-1,n-q+1,1-\alpha}$$

Intervalos-T^2 simultâneos de $100(1-\alpha)\%$ de confiança para contrastes isolados $\boldsymbol{c}^\top\boldsymbol{\mu}$:

$$\left(\boldsymbol{c}^\top\overline{\boldsymbol{x}}-\xi_{n,q-1,1-\alpha}\sqrt{\frac{\boldsymbol{c}^\top\boldsymbol{Sc}}{n}};\ \boldsymbol{c}^\top\overline{\boldsymbol{x}}+\xi_{n,q-1,1-\alpha}\sqrt{\frac{\boldsymbol{c}^\top\boldsymbol{Sc}}{n}}\right),$$

em que

$$\xi_{n,q-1,1-\alpha}=\sqrt{\frac{(n-1)(q-1)}{n-q+1}F_{q-1,n-q+1,1-\alpha}}.$$

Os intervalos de Bonferroni simultâneos de $100(1-\alpha)\%$ de confiança para contrastes isolados $\boldsymbol{c}^\top\boldsymbol{\mu}$:

$$\left(\boldsymbol{c}^\top\overline{\boldsymbol{x}}-t_{n-1,1-\frac{\alpha}{2(q-1)}}\sqrt{\frac{\boldsymbol{c}^\top\boldsymbol{Sc}}{n}};\ \boldsymbol{c}^\top\overline{\boldsymbol{x}}+t_{n-1,1-\frac{\alpha}{2(q-1)}}\sqrt{\frac{\boldsymbol{c}^\top\boldsymbol{Sc}}{n}}\right).$$

Exemplo

Anestésicos melhorados são muitas vezes desenvolvidos primeiramente estudando seus efeitos em animais. Em um estudo, 19 cães receberam inicialmente a droga "pentobarbitol". Depois, cada cão recebeu dióxido de carbono (CO_2) em cada um de dois níveis de pressão (alta e baixa). Depois, halotano (H) foi adicionado e a administração de CO_2 foi repetida. A resposta, milisegundos entre batimentos cardíacos, foi medida para as quatro combinações de tratamento, a saber,

Tratamento 1: H ausente, CO_2 alta pressão;

Tratamento 2: H ausente, CO_2 baixa pressão;

Tratamento 3: H presente, CO_2 alta pressão; e

Tratamento 4: H presente, CO_2 baixa pressão.

Deseja-se testar os efeitos do anestésico quanto à pressão do CO_2 e quanto à presença de halotano.

Ver arquivo exemplo_12.r

Há três contrastes de tratamento que podem ser de interesse do experimentador.

Sejam μ_1, μ_2, μ_3 e μ_4 as correspondentes respostas médias para os Tratamentos 1, 2, 3 e 4. Então,

- $(\mu_3 + \mu_4) - (\mu_1 + \mu_2)$: contraste do Halotano representado pela diferença entre presença e ausência de H;
- $(\mu_1 + \mu_3) - (\mu_2 + \mu_4)$: contraste entre pressão alta e pressão baixa de CO_2; e
- $(\mu_1 + \mu_4) - (\mu_2 + \mu_3)$: contraste representando a interação entre pressão de CO_2 e H.

Rejeitamos H_0 porque $T^2 = 116,02$ é maior que o valor crítico $\xi^2_{19;\,3;\,0,95} = 10,93$.

Como H_0 é rejeitada, faz sentido estudar os intervalos-T^2 simultâneos de 95% dos efeitos estudados.

Intervalo para o efeito do Halotano: (135,65 ; 282,98);

Intervalo para o efeito do CO_2: (-114,73 ; -5,38); e

Intervalo para o efeito de interação H$\times$$CO_2$: (-78,73 ; 53,15).

Portanto, concluímos que a interação é desprezível, que a presença do Halotano aumenta o intervalo entre batimentos cardíacos e que na pressão baixa de CO_2 o intervalo entre batimentos cardíacos é maior.

Cuidados com estas interpretações simples dos resultados devem ser considerados. O efeito-H aparente pode ser devido à tendência do tempo. Ideal: a ordem temporal de todos os tratamentos deve ser determinada ao acaso.

O teste estudado é apropriado quando a matriz de covariâncias $\mathbf{\Sigma}$ não pode assumir qualquer estrutura especial. Se é razoável supor que ela tem uma estrutura particular, testes construídos para matrizes de covariâncias com estruturas especiais têm maior poder do que o teste apresentado aqui.

Parte 2: Comparação de vetores de médias para duas populações - amostras independentes

A estatística T^2 também é apropriada para comparar respostas de um conjunto de observações experimentais (**população 1**) com outro conjunto independente de observações experimentais (**população 2**). Isto pode ser feito sem explicitamente controlar a variabilidade unidade a unidade, como no caso das comparações emparelhadas.

Se possível, as unidades experimentais devem ser aleatoriamente designadas aos conjuntos de condições experimentais. A aleatorização irá, em alguma extensão, abrandar a variabilidade unidade a unidade na comparação subsequente de tratamentos. Apesar da perda em precisão relativa às comparações emparelhadas, as inferências no caso de duas populações são, ordinariamente, aplicáveis a uma coleção mais geral de unidades experimentais simplesmente porque a homogeneidade unitária não é mais exigida.

Considere uma amostra aleatória de tamanho n_1 da **população 1** e uma amostra aleatória de tamanho n_2, independente da primeira amostra, da **população 2**.

Deseja-se fazer inferência sobre os vetores de médias das duas populações, por exemplo,

$$\begin{cases} H_0: & \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0} \\ H_1: & \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \mathbf{0} \end{cases}$$

Suposições básicas:

(a) $(\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1})$ é uma amostra aleatória de uma população p-variada com vetor de médias $\boldsymbol{\mu}_1$ e matriz de covariâncias $\boldsymbol{\Sigma}_1$.

(b) $(\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2})$ é uma amostra aleatória de uma população p-variada com vetor de médias $\boldsymbol{\mu}_2$ e matriz de covariâncias $\boldsymbol{\Sigma}_2$.

(c) As duas amostras são independentes.

Suposições adicionais (tamanhos amostrais pequenos):

(d) Ambas as populações são normais multivariadas.

(e) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

A suposição de igualdade das covariâncias é bem mais forte do que a sua contrapartida univariada. Aqui estamos assumindo que vários pares de covariâncias são aproximadamente iguais.

Quando $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, temos que $\boldsymbol{S}_1$ e $\boldsymbol{S}_2$, sob a suposição de normalidade, têm distribuições independentes de Wishart com $n_1 - 1$ e $n_2 - 1$ graus de liberdade, respectivamente, e parâmetro de escala $\boldsymbol{\Sigma}$, em que $\mathrm{E}\left((n_\ell - 1)\boldsymbol{S}_\ell\right) = (n_\ell - 1)\boldsymbol{\Sigma}$, para $\ell = 1, 2$.

Assim, **podemos definir um estimador combinado para $\boldsymbol{\Sigma}$ dado por**

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

Para testar as hipóteses

$$\begin{cases} H_0: & \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0 \\ H_1: & \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0, \end{cases}$$

usaremos a estatística T^2 correspondente à diferença entre as médias amostrais, a saber,

$$T^2 = \left(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2 - \boldsymbol{\delta}_0\right)^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{S}_c\right]^{-1} \left(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2 - \boldsymbol{\delta}_0\right).$$

Além disso,

$$T^2 \sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} \mathcal{F}_{p, n_1+n_2-p-1}.$$

Rejeitamos a H_0 se $T^2_{obs} > c^2$ em que

$$c^2 = \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1+n_2-p-1, 1-\alpha}.$$

Intervalos-T^2 simultâneos

Faça

$$\xi_{n_1+n_2-1,p,1-\alpha} = \sqrt{\left[\frac{(n_1+n_2-2)p}{n_1+n_2-p-1}\right] F_{p,n_1+n_2-p-1,1-\alpha}}.$$

Então, os intervalos simultâneos de $100(1-\alpha)\%$ de confiança para $\boldsymbol{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ são dados por

$$\boldsymbol{a}^\top(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \pm \xi_{n_1+n_2-1,p,1-\alpha}\sqrt{\boldsymbol{a}^\top\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\boldsymbol{S}_c\boldsymbol{a}}$$

Em particular, os intervalos de $100(1-\alpha)\%$ de confiança para $\mu_{1j} - \mu_{2j}$ serão dados por

$$\overline{x}_{1j} - \overline{x}_{2j} \pm \xi_{n_1+n_2-1,p,1-\alpha}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_{jj,c}}, \quad j = 1, 2, \ldots, p.$$

A situação na qual $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Quando isto ocorre, não somos capazes de obter uma medida de "distância" T^2, cuja distribuição não dependa das matrizes de covariâncias desconhecidas.

Sugere-se que qualquer discrepância da ordem de $\sigma_{1,ii} = 4\sigma_{2,ii}$, ou vice-versa, é provavelmente séria. Uma transformação pode melhorar a situação quando as variâncias marginais são bem diferentes.

Porém, para n_1 e n_2 grandes, podemos evitar complexidades devido a uma desigualdade das covariâncias.

Proposição

Suponha que $n_1 - p$ e $n_2 - p$ sejam grandes. Então, um hiper-elipsóide aproximado de $100(1-\alpha)\%$ de confiança para $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ é dado para todo $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ que sastisfaça

$$[\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^\top \left[\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right]^{-1} [\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \leqslant \chi^2_{p,1-\alpha}.$$

Também, intervalos de $100(1-\alpha)\%$ de confiança para as combinações $\boldsymbol{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ são dados por

$$\boldsymbol{a}^\top(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \pm \sqrt{\chi^2_{p,1-\alpha}}\sqrt{\boldsymbol{a}^\top\left[\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right]\boldsymbol{a}}.$$

Uma aproximação da distribuição de T^2 para populações normais, quando os tamanhos amostrais não são grandes

Pode-se testar $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ quando as matrizes de covariâncias são desiguais, mesmo para tamanhos amostrais moderados, desde que as populações sejam normais multivariadas. Esta situação costuma ser chamada de problema de Behrens-Fisher. O resultado exige que $n_1 > p$ e $n_2 > p$.

A abordagem depende de uma aproximação da distribuição da estatística

$$T^2 = \left[\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^\top \left[\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right]^{-1} \left[\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right].$$

Observe que é a mesma estatística usada no resultado anterior. No entanto, em vez de usar a distribuição aproximada de qui-quadrado para obter os valores críticos para testar H_0, a aproximação recomendada para amostras moderadas a pequenas é dada por

$$T^2 \overset{a}{\sim} \frac{\nu p}{\nu - p + 1}\mathcal{F}_{p,\nu-p+1},$$

com

$$\nu = \frac{p + p^2}{\sum_{i=1}^{2}\frac{1}{n_i}\left\{\text{tr}\left[\left(\frac{1}{n_i}\boldsymbol{S}_i\left(\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right)^{-1}\right)^2\right] + \left(\text{tr}\left[\frac{1}{n_i}\boldsymbol{S}_i\left(\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right)^{-1}\right]\right)^2\right\}}.$$

em que $\min\{n_1, n_2\} \leqslant \nu \leqslant n_1 + n_2$.

Esta aproximação reduz-se à solução de Welch para o problema de Behrens-Fisher univariado ($p = 1$).

MANOVA a um fator

Dados

$$
\begin{array}{lll}
\text{Amostra 1:} & \boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1} & \text{da pop. 1} \\
\text{Amostra 2:} & \boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2} & \text{da pop. 2} \\
\vdots & \vdots & \\
\text{Amostra } g\text{:} & \boldsymbol{X}_{g1}, \boldsymbol{X}_{g2}, \ldots, \boldsymbol{X}_{gn_g} & \text{da pop. } g
\end{array}
$$

Suposições:

- $(\boldsymbol{X}_{k1}, \boldsymbol{X}_{k2}, \ldots, \boldsymbol{X}_{kn_k})$ é uma amostra de tamanho n_k de uma população com vetor de médias $\boldsymbol{\mu}_k$, $k = 1, 2, \ldots, g$;
- todas as populações têm matrizes de covariâncias comum $\boldsymbol{\Sigma}$;
- cada população é normal p-variada; e
- as amostras são independentes.

Primeiro, faremos uma rápida revisão do modelo ANOVA a um fator univariado.

Neste caso, tem-se g amostras aleatórias independentes $X_{k1}, X_{k2}, \ldots, X_{kn_k}$ de tamanhos n_k de distribuições $\mathcal{N}(\mu_k, \sigma^2)$, $k = 1, 2, \ldots, g$.

As hipóteses de interesse são

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_g \\ H_1: & \text{pelo menos uma média é diferente das demais.} \end{cases}$$

O modelo ANOVA é definido por

$$X_{ki} = \underbrace{\mu_k}_{\substack{\text{média da } k\text{-ésima amostra}}} + \overbrace{\varepsilon_{ki}}^{\substack{\text{erro aleatório } i\text{-ésima observação da } k\text{-ésima amostra}}},$$

para $i = 1, 2, \ldots, n_k$, e $k = 1, 2, \ldots, g$.

Em geral, adota-se a reparametrização

$$\mu_k = \underbrace{\mu}_{\text{média global}} + \underbrace{\tau_k}_{\text{efeito do } k\text{-ésimo tratamento}}$$

e, assim, as hipóteses equivalentes são

$$\begin{cases} H_0: & \tau_1 = \tau_2 = \cdots = \tau_g = 0 \\ H_1: & \text{pelo menos um deles é não nulo.} \end{cases}$$

Assim, o modelo é

$$X_{ki} = \mu + \tau_k + \varepsilon_{ki}, \quad i = 1, 2, \ldots, n_k, \quad k = 1, 2, \ldots, g.$$

Suposição: $\varepsilon_{ki} \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$.

Para definir de modo único os parâmetros do modelo e suas estimativas de mínimos quadrados é usual impor a restrição $\sum_{k=1}^{g} n_k \tau_k = 0$.

Se as amostras são balanceadas, isto é, se $n_1 = n_2 = \cdots = n_g$, então a restrição se simplifica para $\sum_{k=1}^{g} \tau_k = 0$.

O procedimento básico adotado é a decomposição da soma de quadrados totais:

$$\begin{aligned}
\underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(X_{ki} - \overline{X}_{..})^2}_{\text{SQ total}} &= \sum_{k=1}^{g}\sum_{i=1}^{n_k}(X_{ki} - \overline{X}_{k.} + \overline{X}_{k.} - \overline{X}_{..})^2 \\
&= \underbrace{\sum_{k=1}^{g} n_k(\overline{X}_{k.} - \overline{X}_{..})^2}_{\text{SQ tratamentos}} \\
&+ \underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(X_{ki} - \overline{X}_{k.})^2}_{\text{SQ devida ao erro aleatório}}.
\end{aligned}$$

Os estimadores de mínimos quadrados de μ e τ_k são dados por

$$\widehat{\mu} = \overline{X}_{..} = \frac{1}{n}\sum_{k=1}^{g}\sum_{i=1}^{n_k} X_{ki}, \quad \text{com} \quad n = \sum_{k=1}^{g} n_k,$$

$$\widehat{\tau}_k = \overline{X}_{k.} - \overline{X}_{..}, \quad \text{com} \quad \overline{X}_{k.} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki}.$$

Um estimador não tendencioso para σ^2 é dado por

$$\widetilde{\sigma}^2 = \frac{SQ_{Erro}}{n-g} = \frac{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(X_{ki} - \overline{X}_{k.})^2}{n-g}, \quad \text{com} \quad n = \sum_{k=1}^{g} n_k.$$

Tabela 1: ANOVA a um fator com g níveis.

Fonte de Variação	g.l.	Soma de Quadrados	Quadrado Médio	Razão F
Tratamento	$g-1$	$\sum_{k=1}^{g} n_k(\overline{x}_{k.}-\overline{x}_{..})^2$	$SQ_{Trat}/(g-1)$	$\frac{QM_{Trat}}{QM_{Erro}}$
Erro	$n-g$	$\sum_{k=1}^{g}\sum_{i=1}^{n_k}(x_{ki}-\overline{x}_{k.})^2$	$SQ_{Erro}/(n-g)$	
Total	$n-1$	$\sum_{k=1}^{g}\sum_{i=1}^{n_k}(x_{ki}-\overline{x}_{..})^2$		

em que $n=\sum_{k=1}^{g} n_k$.

Se H_0 é verdadeira, $F \sim \mathcal{F}_{g-1,n-g}$.

Logo, rejeitaremos H_0 ao nível de significância α se $F > F_{g-1,n-g,1-\alpha}$.

Observe que isto é equivalente a dizer que o teste rejeita H_0 para valores grandes da razão SQ_{Trat}/SQ_{Erro}, equivalentemente, valores grandes de $1 + SQ_{Trat}/SQ_{Erro}$ ou, equivalentemente, valores pequenos da recíproca, dada por

$$\frac{SQ_{Erro}}{SQ_{Trat} + SQ_{Erro}}.$$

Adiante, veremos que uma estatística de teste no caso multivariado, com suas devidas adaptações, tem forma similar a esta última.

MANOVA a um fator (continuação)

Modelo:

$$\boldsymbol{X}_{ki} = \boldsymbol{\mu} + \boldsymbol{\tau}_k + \varepsilon_{ki}, \quad k = 1, 2, \ldots, g, \quad i = 1, 2, \ldots, n_k.$$

Suposição:

$$\varepsilon_{ki} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

Restrição:

$$\sum_{k=1}^{g} n_k \boldsymbol{\tau}_k = \mathbf{0}$$

Soma de Quadrados Total:

$$SQ_{Tot} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (\boldsymbol{X}_{ki} - \overline{\boldsymbol{X}}_{..})(\boldsymbol{X}_{ki} - \overline{\boldsymbol{X}}_{..})^\top.$$

O vetor de médias amostral é dado por

$$\overline{\boldsymbol{X}}_{..} = \frac{1}{n} \sum_{k=1}^{g} \sum_{i=1}^{n_k} \boldsymbol{X}_{ki} \quad \text{com} \quad n = \sum_{k=1}^{g} n_k.$$

Agora, SQ_{Tot} é uma matriz $p \times p$.

Os elementos de sua diagonal correspondem às somas de quadrados totais para cada medida isolada.

Fora da diagonal temos as somas dos produtos cruzados dos desvios, caracterizando a estrutura de dependência entre as p medidas estudadas.

Decomposição de SQ_{Tot}:

$$
\begin{aligned}
SQ_{Tot} &= \sum_{k=1}^{g} \sum_{i=1}^{n_k} (\boldsymbol{X}_{ki} - \overline{\boldsymbol{X}}_{..})(\boldsymbol{X}_{ki} - \overline{\boldsymbol{X}}_{..})^{\top} \\
&= \underbrace{\sum_{k=1}^{g} n_k (\overline{\boldsymbol{X}}_{k.} - \overline{\boldsymbol{X}}_{..})(\overline{\boldsymbol{X}}_{k.} - \overline{\boldsymbol{X}}_{..})^{\top}}_{SQ_{Trat}} \\
&+ \underbrace{\sum_{k=1}^{g} \sum_{i=1}^{n_k} (\boldsymbol{X}_{ki} - \overline{\boldsymbol{X}}_{k.})(\boldsymbol{X}_{ki} - \overline{\boldsymbol{X}}_{k.})^{\top}}_{SQ_{Res}} \\
&= \mathrm{SQ}_{Trat} + \mathrm{SQ}_{Res} = \boldsymbol{B} + \boldsymbol{W}.
\end{aligned}
$$

No caso univariado, a decomposição, válidas as suposições, fornece estatísticas independentes com distribuições de qui-quadrado, parâmetro de escala σ^2.

No caso multivariado, o resultado é similar. A decomposição fornece estatísticas independentes com distribuição Wishart, parâmetro de escala $\mathbf{\Sigma}$.

Observe que quanto ao número de graus de liberdade, eles são exatamente os mesmos do caso univariado.

Hipóteses:

$$\begin{cases} H_0: & \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g = \mathbf{0} \\ H_1: & \text{pelo menos um deles é não nulo.} \end{cases}$$

Tabela 2: Tabela MANOVA a um fator com g níveis.

Fonte de Variação	g.l.	Matriz de Soma de Quadrados
Tratamento	$g-1$	$\boldsymbol{B} = \sum_{k=1}^{g} n_k(\overline{\boldsymbol{X}}_{k.} - \overline{\boldsymbol{X}}_{..})(\overline{\boldsymbol{X}}_{k.} - \overline{\boldsymbol{X}}_{..})^\top$
Erro	$n-g$	$\boldsymbol{W} = \sum_{k=1}^{g} \sum_{i=1}^{n_k}(\boldsymbol{X}_{ki} - \overline{\boldsymbol{X}}_{k.})(\boldsymbol{X}_{ki} - \overline{\boldsymbol{X}}_{k.})^\top$
Total	$n-1$	$\boldsymbol{B}+\boldsymbol{W} = \sum_{k=1}^{g} \sum_{i=1}^{n_k}(\boldsymbol{X}_{ki} - \overline{\boldsymbol{X}}_{..})(\boldsymbol{X}_{ki} - \overline{\boldsymbol{X}}_{..})^\top$

em que $n = \sum_{k=1}^{g} n_k$.

As diagonais principais de cada uma destas matrizes de somas de quadrados e produtos cruzados fornecem as somas de quadrados da ANOVA para cada medida separadamente.

Uma possível estatística de teste envolve variâncias generalizadas. Seja

$$\Lambda^{\star} = \frac{|\boldsymbol{W}|}{|\boldsymbol{B} + \boldsymbol{W}|}.$$

Rejeitaremos H_0 se $\Lambda^{\star}$ for um valor pequeno.

A estatística $\Lambda^{\star}$ foi originalmente proposta por Wilks e corresponde a uma forma equivalente do teste $\mathcal{F}$ da hipótese de ausência de efeito de tratamento do caso univariado.

A estatística lambda de Wilks tem a vantagem de ser conveniente e estar relacionada ao critério da razão de verossimilhança.

A distribuição exata de $\Lambda^\star$ pode ser derivada para os casos especiais apresentados a seguir.

Casos especiais:

1. $p = 1, g \geqslant 2, \left(\frac{n-g}{g-1}\right)\left(\frac{1-\Lambda^\star}{\Lambda^\star}\right) \sim \mathcal{F}_{g-1,n-g}.$
2. $p = 2, g \geqslant 2, \left(\frac{n-g-1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^\star}}{\sqrt{\Lambda^\star}}\right) \sim \mathcal{F}_{2(g-1),2(n-g-1)}.$
3. $p \geqslant 1, g = 2, \left(\frac{n-p-1}{p}\right)\left(\frac{1-\Lambda^\star}{\Lambda^\star}\right) \sim \mathcal{F}_{p,n-p-1}.$
4. $p \geqslant 1, g = 3, \left(\frac{n-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda^\star}}{\sqrt{\Lambda^\star}}\right) \sim \mathcal{F}_{2p,2(n-p-2)}.$

Para outros casos e tamanhos amostrais grandes, uma modificação de $\Lambda^\star$ devido à Bartlett pode ser usada para testar H_0.

Bartlett mostrou que se H_0 é verdadeira e n é grande, então

$$-\left(n-1-\frac{(p+g)}{2}\right)\ln\Lambda^{\star} \overset{a}{\sim} \chi^2_{p(g-1)}$$

Consequentemente, se n é grande, rejeitamos H_0 ao nível de significância α, se

$$-\left(n-1-\frac{(p+g)}{2}\right)\ln\Lambda^{\star} > \chi^2_{p(g-1),1-\alpha}.$$

Exemplo

O Departamento de Saúde e Serviços Sociais de Wisconsin reembolsa casas de repouso para idosos no estado por serviços fornecidos. O departamento desenvolve um conjunto de fórmulas para as taxas de cada facilidade, baseado em fatores tais como nível de cuidados, taxa média de salários, e taxa média de salário no estado.

As casas de repouso podem ser classificadas com relação à propriedade - privadas, sem fins lucrativos e públicas - e também com relação à certificação - especializadas em enfermagem, unidades de cuidados intermediários, ou uma combinação dos dois.

Um objetivo de um estudo foi investigar os efeitos de propriedade e certificação (ou ambos) sobre os custos. Quatro custos, calculados por paciente-dia, medidos em horas-paciente, foram usados:

X_1: custo de mão de obra de enfermagem;

X_2: custo de nutricionista;

X_3: custo de trabalho de manutenção e funcionamento; e

X_4: custo de limpeza e lavanderia.

Um total de $n = 516$ observações sobre cada um dos $p = 4$ custos foram inicialmente separados pelo tipo de propriedade.

Tabela 3: Estatísticas sumário para cada um dos $g = 3$ grupos.

Grupo k	n_k	Vetor de médias amostral
$k = 1$, privado	$n_1 = 271$	$\overline{\boldsymbol{x}}_1^\top = (2,066;\ 0,480;\ 0,082;\ 0,360)$
$k = 2$, não-lucrativo	$n_2 = 138$	$\overline{\boldsymbol{x}}_2^\top = (2,167;\ 0,596;\ 0,124;\ 0,418)$
$k = 3$, público	$n_3 = 107$	$\overline{\boldsymbol{x}}_3^\top = (2,273;\ 0,521;\ 0,125;\ 0,383)$

$$\boldsymbol{S}_1 = \begin{bmatrix} 0,291 & -0,001 & 0,002 & 0,010 \\ -0,001 & 0,011 & 0,000 & 0,003 \\ 0,002 & 0,000 & 0,001 & 0,000 \\ 0,010 & 0,003 & 0,000 & 0,010 \end{bmatrix},$$

$$\boldsymbol{S}_2 = \begin{bmatrix} 0,561 & 0,011 & 0,001 & 0,037 \\ 0,011 & 0,025 & 0,004 & 0,007 \\ 0,001 & 0,004 & 0,005 & 0,002 \\ 0,037 & 0,007 & 0,002 & 0,019 \end{bmatrix} \quad \text{e}$$

$$\boldsymbol{S}_3 = \begin{bmatrix} 0,261 & 0,030 & 0,003 & 0,018 \\ 0,030 & 0,017 & 0,000 & 0,006 \\ 0,003 & 0,000 & 0,004 & 0,001 \\ 0,018 & 0,006 & 0,001 & 0,013 \end{bmatrix}.$$

Como os $\boldsymbol{S}_k$'s parecem similares, é razoável supor matriz de covariâncias iguais.

Calculando as matrizes da tabela MANOVA obtemos

$$\boldsymbol{W} = \begin{bmatrix} 182,962 & 4,408 & 1,695 & 9,581 \\ 4,408 & 8,200 & 0,633 & 2,428 \\ 1,695 & 0,633 & 1,484 & 0,394 \\ 9,581 & 2,428 & 0,394 & 6,538 \end{bmatrix},$$

com 513 graus de liberdade e

$$\boldsymbol{B} = \begin{bmatrix} 3,475 & 1,111 & 0,821 & 0,584 \\ 1,111 & 1,225 & 0,453 & 0,610 \\ 0,821 & 0,453 & 0,235 & 0,230 \\ 0,584 & 0,610 & 0,230 & 0,304 \end{bmatrix},$$

com 2 graus de liberdade.

Neste caso temos $\Lambda^\star = \dfrac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|} \simeq 0,771$.

Como $p = 4$ e $g = 3$, observe que estamos sob o caso especial 4:

$$p \geqslant 1, \quad g = 3, \quad \left(\frac{n-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda^\star}}{\sqrt{\Lambda^\star}}\right) \sim \mathcal{F}_{2p,2(n-p-2)}.$$

Temos que $\left(\dfrac{n-p-2}{p}\right)\left(\dfrac{1-\sqrt{\Lambda^\star}}{\sqrt{\Lambda^\star}}\right) \simeq 17,67$.

Ao nível de significância de 1% temos $F_{8;\,1000;\,0,99} \simeq \chi^2_{8;\,0,99}/8 = 2,51$.

Logo, rejeitamos H_0, a hipótese de que não há diferença no custo médio entre os diferentes tipos de propriedades.

Vale comparar este resultado com o procedimento sugerido por Bartlett.

Temos que $$-\left(n-1-\frac{(p+g)}{2}\right)\ln\Lambda^{\star}\simeq 132,76.$$

Ao nível de significância de 1% tem-se $\chi^2_{8;\,0,99}=20,09$, o que novamente leva à rejeição de H_0.

Este resultado é consistente com o resultado baseado na estatística $\mathcal{F}$.

Intervalos simultâneos de confiança para efeitos de tratamento

Quando H_0 é rejeitada, os efeitos que levaram à rejeição são de interesse. Para comparações pareadas, a abordagem de Bonferroni pode ser usada para construir intervalos simultâneos de confiança para os componentes dos vetores diferença $\boldsymbol{\tau}_k - \boldsymbol{\tau}_\ell$.

Seja τ_{kj} o j-ésimo componente de $\boldsymbol{\tau}_k$.

Como $\boldsymbol{\tau}_k$ é estimado por $\widehat{\boldsymbol{\tau}}_k = \overline{\boldsymbol{X}}_{k.} - \overline{\boldsymbol{X}}_{..}$ temos que o j-ésimo componente de $\boldsymbol{\tau}_k$ é estimado por $\overline{X}_{k.(j)} - \overline{X}_{..(j)}$.

Assim, $\tau_{kj} - \tau_{\ell j}$ é estimado por

$$\widehat{\tau}_{kj} - \widehat{\tau}_{\ell j} = \overline{X}_{k.(j)} - \overline{X}_{\ell.(j)}$$

que é uma diferença entre médias de duas amostras independentes.

O teste t-Student para duas amostras independentes é válido para um nível de significância apropriado.

$$\text{Var}(\widehat{\tau}_{kj} - \widehat{\tau}_{\ell j}) = \left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\sigma_{jj}$$

com σ_{jj} o j-ésimo elemento da diagonal de $\mathbf{\Sigma}$.

Como sugerido, σ_{jj} é estimado pelo j-ésimo elemento da diagonal da matriz $\boldsymbol{W}$ dividido pelos respectivos graus de liberdade, a saber, $\widehat{\sigma}_{jj} = \dfrac{1}{n-g} w_{jj}$.

Resta identificar o número de intervalos. Como são g tratamentos para p medidas, segue que ao todo teremos $m = p\begin{pmatrix} g \\ 2 \end{pmatrix} = pg(g-1)/2$ intervalos.

Logo, com confiança pelo menos $1-\alpha$ os intervalos simultâneos para $\tau_{kj} - \tau_{\ell j}$ são dados por

$$(\overline{x}_{k.(j)} - \overline{x}_{\ell.(j)}) \pm t_{n-g,1-\alpha/(pg(g-1))}\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\frac{w_{jj}}{n-g}}.$$

Exemplo

Vimos no exemplo anterior - referente aos dados sobre casas de repouso para idosos de Wisconsin - que há diferenças de custos conforme o tipo de propriedade do estabelecimentos: privado, não lucrativo e público. Podemos, então calcular os intervalos simultâneos para estimar as magnitudes das diferenças. Temos que

$$\begin{aligned}
\widehat{\boldsymbol{\tau}}_1^\top &= (-0,070;\ -0,039;\ -0,020;\ -0.020) \\
\widehat{\boldsymbol{\tau}}_2^\top &= (\ \ 0,031;\ \ 0,076;\ \ 0,022;\ \ 0,038) \\
\widehat{\boldsymbol{\tau}}_3^\top &= (\ \ 0,137;\ \ 0,002;\ \ 0,023;\ \ 0,003)
\end{aligned}$$

e

$$\boldsymbol{W} = \begin{bmatrix} 182,962 & 4,408 & 1,695 & 9,581 \\ 4,408 & 8,200 & 0,633 & 2,428 \\ 1,695 & 0,633 & 1,484 & 0,394 \\ 9,581 & 2,428 & 0,394 & 6,538 \end{bmatrix},$$

com 513 graus de liberdade.

Tabela 4: Intervalos simultâneos de pelo menos 95% de confiança.

Grupos	Comp.	Dif. médias	Limite Inf.	Limite Sup.
1,2	1	-0,10	-0,28	0,08
	2	-0,12	**-0,15**	**-0,08**
	3	-0,04	**-0,06**	**-0,01**
	4	-0,06	**-0,09**	**-0,02**
1,3	1	-0,21	**-0,40**	**-0,01**
	2	-0,04	-0,08	0,00
	3	-0,04	**-0.06**	**-0,03**
	4	-0,02	-0,06	0,01
2,3	1	-0,11	-0,33	0,12
	2	0,08	**0,03**	**0,12**
	3	0,00	-0,02	0,02
	4	0,04	-0,01	0,08

Foram destacados os intervalos que excluíram o zero.

Com relação ao custo mão de obra de enfermagem: as casas de repouso particulares parecem ter um custo mais baixo do que as públicas.

Com relação ao custo de nutricionista: as casas particulares parecem ter custo menor do que as sem fins lucrativos e, as sem fins lucrativos mostram um custo maior do que as públicas.

Com relação ao custo de manutenção e operação: as particulares mostram ter um custo menor tanto em comparação com as sem fins lucrativos, como com as públicas.

Com relação ao custo de limpeza e lavanderia: as particulares parecem ter custo menor do que as sem fins lucrativos.

Teste para a igualdade das matrizes de covariâncias

Uma das suposições feitas no modelo MANOVA foi a de que todas as populações têm variâncias iguais. Portanto, deve-se verificar se esta suposição é razoável ou não.

Um teste comumente usado é o teste *M* de Box.

Com g populações, tem-se

$$\begin{cases} H_0: & \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma} \\ H_1: & \text{pelo menos duas matrizes de covariâncias não são iguais.} \end{cases}$$

Supondo populações normais multivariadas, uma estatística de razão de verossimilhança é dada por

$$\Lambda = \prod_{k=1}^{g} \left(\frac{|\boldsymbol{S}_k|}{|\boldsymbol{S}_c|} \right)^{(n_k-1)/2} \quad \text{com} \quad \boldsymbol{S}_c = \frac{1}{n-g} \sum_{k=1}^{g} (n_k - 1)\boldsymbol{S}_k.$$

O teste de Box é baseado na aproximação de qui-quadrado da distribuição amostral de $-2 \ln \Lambda$.

Faça $M = -2 \ln \Lambda$, então

$$M = (n - g) \ln |\boldsymbol{S}_c| - \sum_{k=1}^{g} [(n_k - 1) \ln |\boldsymbol{S}_k|].$$

Se H_0 é verdadeira, as matrizes de covariâncias devem ser bem parecidas e, consequentemente, também serão similares aos elementos em $\boldsymbol{S}_c$. Nesse caso, Λ estará próximo de 1 e, M, próximo de 0.

Teste *M* de Box para a igualdade de matrizes de covariância

Faça

$$u = \left[\sum_{k=1}^{g} \frac{1}{n_k - 1} - \frac{1}{\sum_{k=1}^{g}(n_k - 1)}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(g-1)}\right]$$

em que p é o número de variáveis e g é o número de populações. Então,

$$C = (1-u)M \overset{a}{\sim} \chi^2_\nu, \quad \text{com} \quad \nu = g\frac{1}{2}p(p+1) - \frac{1}{2}p(p+1) = \frac{1}{2}p(p+1)(g-1).$$

Logo, ao nível de significância α rejeitamos H_0 se

$$C > \chi^2_{\nu,1-\alpha}.$$

A aproximação de χ^2 de Box funciona melhor se cada $n_k > 20$ e se p e g não são superiores a 5.

O teste *M* é sensível a algumas formas de não-normalidade. Porém, com amostras razoavelmente grandes, os testes MANOVA para efeitos de tratamento são mais robustos à não-normalidade, isto é, podemos prosseguir com a MANOVA usual, mesmo que o teste *M* rejeite H_0.

Perspectivas e estratégia na análise de modelos multivariados

Quando se tem várias características, é importante controlar a probabilidade global de tomar uma decisão incorreta. Isto é particularmente importante quando testamos a igualdade entre dois ou mais tratamentos.

Um único teste multivariado com seu p-valor associado é preferível a realizar um grande número de testes univariados.

A saída do teste multivariado nos dirá se vale ou não a pena olhar de forma mais minuciosa variável por variável e analisar grupo por grupo.

Um único teste multivariado é recomendável em vez de, digamos, p testes univariados, pois como o próximo exemplo demonstrará, testes univariados ignoram informações importantes e podem levar a conclusões erradas.

Exemplo

Suponha que tenhamos coletado medidas sobre duas variáveis X_1 e X_2 para 10 unidades expermentais, selecionadas ao acaso de cada uma de duas populações (dois grupos). Os dados, hipotéticos são listados a seguir.

Grupo 1		Grupo 2	
X_1	X_2	X_1	X_2
5,0	3,0	4,6	4,9
4,5	3,2	4,9	5,9
6,0	3,5	4,0	4,1
6,0	4,6	3,8	5,4
6,2	5,6	6,2	6,1
6,9	5,2	5,0	7,0
6,8	6,0	5,3	4,7
5,3	5,5	7,1	6,6
6,6	7,3	5,8	7,8
7,3	6,5	6,8	8,0

É possível ver pelos diagramas de ponto de cada variável e o diagrama de dispersão das duas, que há boa sobreposição para os dois grupos em questão.

O teste de comparação dos dois vetores de médias, ao nível de significância de 10%, resulta em $T^2 = 17,29$ com valor crítico ($\alpha = 0,1$) dado por $12,94$, tal que a hipótese nula, de igualdade entre os dois vetores de médias deve ser rejeitada. O valor-p deste deste é aproximadamente $0,0033$.

Se realizarmos as ANOVAS separadamente para cada variável, teremos para a variável X_1 a razão $F = 2,46$ com valor-p superior a 0,10, implicando a não rejeição de que as médias são iguais para esta componente.

O mesmo sucede com a variável X_2 com $F = 2,68$ e, assim, também não rejeitamos a hipótese de que estas médias são iguais.

O teste T^2 leva em conta a correlação positiva entre as duas medidas para cada grupo - informação que é simplesmente ignorada pelos testes univariados.

Estes exemplos demonstram a eficácia de um teste multivariado relativo às suas contrapartidas univariadas. No exemplo sobre o tratamento de esgotos (efluentes) também percebemos exatamente essa aparente incoerência.

Três outros testes multivariados também são comuns

1. TRAÇO DE LAWLEY-HOTELLING: $\text{tr}(\boldsymbol{B}\boldsymbol{W}^{-1})$.
2. TRAÇO DE PILLAI: $\text{tr}([\boldsymbol{B}+\boldsymbol{W}]^{-1})$.
3. MAIOR RAIZ QUADRADA DE ROY: maior autovalor de $\boldsymbol{W}(\boldsymbol{B}+\boldsymbol{W})^{-1}$.

Todos estes testes são aproximadamente equivalentes quando se tem grandes amostras.

Quando, e somente quando, os testes multivariados apontam diferenças, deve-se olhar os componentes separadamente.

No `R`, pacote `stats`, há a função `manova` que calcula todas estas estatísticas e também as matrizes de somas de quadrados para um ou mais fatores.

Estratégia para a comparação multivariada de tratamentos

1. Identifique primeiro a presença pontos aberrantes (*outliers*).
2. Realize um teste multivariado.
3. Se o teste multivariado apontou diferenças, calcule os intervalos simultâneos de Bonferroni que interessam.

Regressão linear

Sejam r covariáveis, $z_1, z_2, \ldots, z_r$, relacionadas a uma variável resposta y.

O **modelo de regressão linear múltiplo** (univariado) é dado pela equação:

$$\underbrace{y}_{\text{resposta}} = \underbrace{\beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r}_{\text{média; parte estrutural}} + \underbrace{\varepsilon}_{\text{erro; parte aleatória}}$$

O modelo é dito linear, pois a parte estrutural é linear nos parâmetros β_j, $j = 1, 2, \ldots, r$.

Se dispomos de n observações independentes, então

$$y_i = \beta_0 + \beta_1 z_{i1} + \cdots + \beta_r z_{ir} + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Suposições:

S1: $\mathrm{E}(\varepsilon_i) = 0, \forall i = 1, 2, \ldots, n.$

S2: $\mathrm{Var}(\varepsilon_i) = \sigma^2, \forall i = 1, 2, \ldots, n$ (homocedasticidade).

S3: $\mathrm{Cov}(\varepsilon_i, \varepsilon_k) = 0, \forall i \neq k, \ i, k \in \{1, 2, ..., n\}.$

Na notação matricial, tem-se

$$\underbrace{\boldsymbol{y}}_{n\times 1} = \underbrace{\boldsymbol{Z}}_{n\times(r+1)}\ \underbrace{\boldsymbol{\beta}}_{(r+1)\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{n\times 1},$$

com

S1: $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$;

S2 e S3: $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$;

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad Z = \begin{bmatrix} 1 & z_{11} & z_{12} & \dots & z_{1r} \\ 1 & z_{21} & z_{22} & \dots & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \dots & z_{nr} \end{bmatrix};$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix}; \quad \text{e} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Observe que ainda não fizemos nenhuma suposição acerca da distribuição dos erros.

Para efeito de obter os estimadores de mínimos quadrados, de fato, não é necessária nenhuma suposição sobre a distribuição da parte aleatória. Porém, para fins de inferência, será necessário.

Estimadores de Mínimos Quadrados

Suponha que a matriz $\boldsymbol{Z}$ seja de posto completo tal que suas colunas formam um conjunto de vetores linearmente independentes. Neste caso, a matriz $\boldsymbol{Z}^\top \boldsymbol{Z}$ é não-singular e o estimador de mínimos quadrados do vetor $\boldsymbol{\beta}$ é dado por

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}.$$

Os valores ajustados são, então, dados por

$$\widehat{\boldsymbol{y}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}, \quad \text{em que} \quad \boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top,$$

e os resíduos

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = [\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}, \quad \text{em que} \quad \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{H},$$

satisfazendo as seguintes relações (somente quando houver a constante β_0 no modelo)

$$\boldsymbol{Z}^\top \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{0} \quad \text{e} \quad \widehat{\boldsymbol{y}}^\top \widehat{\boldsymbol{\varepsilon}} = 0.$$

Temos que $\boldsymbol{H}$ e $\boldsymbol{P}$ são matrizes idempotentes ($\boldsymbol{H} = \boldsymbol{HH}$ e $\boldsymbol{H} = \boldsymbol{H}^\top$).

A soma de quadrados de resíduos é

$$SQ_{Res} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y}^\top \boldsymbol{P}\boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{Z}\widehat{\boldsymbol{\beta}}.$$

Observe que

$$\sum_{i=1}^{n} y_i^2 = \boldsymbol{y}^\top \boldsymbol{y} = (\boldsymbol{y} - \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{y}})^\top (\boldsymbol{y} - \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{y}}) = \widehat{\boldsymbol{y}}^\top \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}.$$

O coeficiente de determinação R^2

$$R^2 = 1 - \frac{\widehat{\varepsilon}^\top \widehat{\varepsilon}}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}.$$

O R^2 fornece a proporção da variação total dos y_i's que é "explicada" pelas covariáveis.

Por um lado, o R^2 seria igual 1 se a equação do modelo se ajustasse perfeitamente aos dados.

Por outro lado, o R^2 seria zero se $\widehat{\beta}_0 = \overline{y}$ e os demais seriam todos nulos. Neste caso, as covariáveis não exerceriam nenhuma influência sobre a resposta.

O R^2 deve ser olhado com cuidado na verificação do modelo, pois valores altos de R^2 não necessariamente implicam que o modelo ajustado é bom. Além dessa medida é fundamental realizar uma análise dos resíduos.

Ademais, um R^2 não tão elevado, para um modelo ajustado cuja análise de resíduos foi boa, pode ser considerado.

Considere o modelo

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{\beta}$ é uma combinação linear das linhas da matriz $\boldsymbol{Z}$ com coeficientes $\beta_0, \beta_1, \ldots, \beta_r$.

A medida que $\boldsymbol{\beta}$ varia, $\boldsymbol{Z}\boldsymbol{\beta}$ gera o "plano modelo" de todas as combinações lineares das colunas de $\boldsymbol{Z}$.

Geralmente, o vetor observado $\boldsymbol{y}$ não pertencerá ao plano modelo devido ao erro aleatório. Isto é, $\boldsymbol{y}$ não é uma combinação linear das colunas de $\boldsymbol{Z}$.

Uma vez que as observações se tornam disponíveis, a solução de mínimos quadrados é obtida a partir do vetor desvio dado por

$$\underbrace{\boldsymbol{y}}_{\text{vetor de observação}} - \underbrace{\boldsymbol{Z}\boldsymbol{\beta}.}_{\text{vetor no plano modelo}}$$

O quadrado do módulo deste vetor é

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^{\top}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}).$$

$S(\widehat{\beta})$ é tão pequeno quanto possível, quando $\widehat{\beta}$ é selecionado tal que $\boldsymbol{Z}\widehat{\beta}$ é o ponto no plano modelo mais próximo do vetor de observação.

De fato, $\boldsymbol{Z}\widehat{\beta}$ é a projeção ortogonal do vetor de observação $\boldsymbol{y}$ sobre o plano modelo.

O vetor $\widehat{\varepsilon}$ é ortogonal ao plano modelo.

Como $\widehat{\boldsymbol{y}} = \boldsymbol{Z}\widehat{\beta} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$, dizemos que $\boldsymbol{H}$ é a matriz de projeção ortogonal do vetor de observações sobre o plano modelo $\boldsymbol{Z}\beta$.

Similarmente, $\boldsymbol{P}$ é a matriz de projeção de $\boldsymbol{y}$ sobre o plano perpendicular ao modelo.

Inferências sobre os parâmetros

Sob o modelo $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ com $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ e $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, o estimador de mínimos quadrados de $\boldsymbol{\beta}$ é $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}$. Então,

$$\begin{aligned}
\mathrm{E}(\widehat{\boldsymbol{\beta}}) &= \boldsymbol{\beta}, \\
\mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= \sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}, \\
\mathrm{E}(\widehat{\boldsymbol{\varepsilon}}) &= \boldsymbol{0}, \quad \text{e} \\
\mathrm{Var}(\widehat{\boldsymbol{\varepsilon}}) &= \sigma^2\boldsymbol{P}.
\end{aligned}$$

Observe que apesar dos erros serem supostos independentes, os resíduos não o serão necessariamente, pois a matriz $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{H}$ pode não ser diagonal.

Temos que

$$\mathrm{E}(\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}) = (n - r - 1)\sigma^2,$$

tal que $s^2 = \dfrac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n - r - 1} = \dfrac{\boldsymbol{y}^\top\boldsymbol{P}\boldsymbol{y}}{n - r - 1}$ é um estimador não tendencioso de σ^2.

Além disso, $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\varepsilon}}$ são não correlacionados.

Para ir adiante nas inferências se impõe a necessidade de se fazer suposições sobre a distribuição dos erros.

A suposição adicional aqui é $\varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$.

Com esta suposição adicional, é possível provar que $\widehat{\boldsymbol{\beta}}$ é o estimador de máxima verossimilhança de $\boldsymbol{\beta}$. Além disso,

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{r+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1})$$

e

$$\frac{(n-r-1)s^2}{\sigma^2} = \frac{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}}{\sigma^2} \sim \chi^2_{n-r-1}.$$

Podemos então construir uma região de confiança conjunta para $\boldsymbol{\beta}$ usando

$$\frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{Z}^\top \boldsymbol{Z} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})}{s^2} \sim (r+1) F_{r+1, n-r-1}.$$

Intervalos simultâneos de $100(1-\alpha)\%$ para os β_i's

$$\text{IC}_{100(1-\alpha)\%}(\beta_i) = \widehat{\beta}_i \pm \sqrt{\widehat{\text{Var}}(\widehat{\beta}_i)}\sqrt{(r+1)F_{r+1,n-r-1,1-\alpha}}, \quad i = 0, 1, \ldots, r,$$

em que $\widehat{\text{Var}}(\widehat{\beta}_i)$ é o i-ésimo elemento da diagonal de $s^2(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}$.

é comum ignorar a simultaneidade dos intervalos se trabalhando com os intervalos separados que são mais estreitos.

$$\text{IC}_{100(1-\alpha)\%}(\beta_i) = \widehat{\beta}_i \pm \sqrt{\widehat{\text{Var}}(\widehat{\beta}_i)}t_{n-r-1,1-\alpha/2}.$$

Neste caso, deve-se atentar para o fato de que aqui a confiança não é simultânea para os $r+1$ intervalos.

Uma abordagem que é interessante de ser usada aqui é a abordagem de Bonferroni.

Inferências a partir da função de regressão estimada

Estimação da média de uma observação nova $\mathrm{E}(y_0|\boldsymbol{z}_0)$, com $\boldsymbol{z}_0^\top = (1, z_{01}, \ldots, z_{0r})$.

A média é dada por $\mathrm{E}(y_0|\boldsymbol{z}_0) = \beta_0 + \beta_1 z_{01} + \cdots + \beta_r z_{0r} = \boldsymbol{z}_0^\top \boldsymbol{\beta}$.

Logo, $\widehat{\mathrm{E}(y_0|\boldsymbol{z}_0)} = \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}$.

Temos que

$$\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{z}_0^\top \boldsymbol{\beta}, \sigma^2 \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0).$$

O intervalo de 100(1-α)% de confiança para $\boldsymbol{z}_0^\top \boldsymbol{\beta}$ é dado por

$$IC_{100(1-\alpha)\%}(\boldsymbol{z}_0^\top \boldsymbol{\beta}) = \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}} \pm t_{n-r-1,1-\alpha/2} \sqrt{s^2 \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0}.$$

Previsão de uma nova observação, quando $\boldsymbol{z} = \boldsymbol{z}_0$

O intervalo de previsão 100(1-α)% de confiança para y_0 é dado por

$$IC_{100(1-\alpha)\%}(y_0) = \boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}} \pm t_{n-r-1,1-\alpha/2} \sqrt{s^2(1 + \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0)}.$$

Alguns outros tópicos que devem ser revisados pelo aluno:

- Verificação do ajuste e outros aspectos da regressão.
- Análise de resíduos.
- Pontos de alavanca e de influência.
- Seleção de covariáveis.
- Comparação de modelos.

Regressão linear multivariada

Suponha que a variável resposta $\boldsymbol{y}$ é p-variada e que $z_1, z_2, \ldots, z_r$ representam r covariáveis escalares.

O **modelo de regressão linear múltiplo multivariado** é dado por:

$$\begin{aligned}
y_{1,i} &= \beta_{01} + \beta_{11} z_{1,i} + \cdots + \beta_{r1} z_{r,i} + \varepsilon_{1,i} \\
y_{2,i} &= \beta_{02} + \beta_{12} z_{1,i} + \cdots + \beta_{r2} z_{r,i} + \varepsilon_{2,i} \\
\vdots \quad &\vdots \qquad\qquad\qquad\qquad \vdots \\
y_{p,i} &= \beta_{0p} + \beta_{1p} z_{1,i} + \cdots + \beta_{rp} z_{r,i} + \varepsilon_{p,i}
\end{aligned}$$

para $i = 1, 2, \ldots, n$, sendo $\boldsymbol{\varepsilon}_i^\top = (\varepsilon_{1,i}, \varepsilon_{2,i}, \ldots, \varepsilon_{p,i})$.

Suposições:

- $\mathrm{E}(\boldsymbol{\varepsilon}_i) = \mathbf{0}$; e
- $\mathrm{Var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Sigma}$, uma matriz $p \times p$ simétrica e positiva definida.

Portanto, os termos de erro associados com diferentes componentes do vetor resposta podem ser correlacionados.

Na notação matricial, tem-se

$$\underbrace{\boldsymbol{Y}}_{n\times p} = \underbrace{\boldsymbol{Z}}_{n\times(r+1)}\ \underbrace{\boldsymbol{\beta}}_{(r+1)\times p} + \underbrace{\boldsymbol{\varepsilon}}_{n\times p}$$

com

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} = \left[\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p\right],$$

$$\boldsymbol{Z} = \begin{bmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1r} \\ 1 & z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \cdots & z_{nr} \end{bmatrix} = \left[\mathbf{1}, \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_r\right],$$

$$
\begin{aligned}
\boldsymbol{\beta} &= \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rp} \end{bmatrix} = \left[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p\right], \quad \text{e} \\
\boldsymbol{\varepsilon} &= \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{np} \end{bmatrix} = \left[\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_p\right].
\end{aligned}
$$

Suposições do modelo $\underbrace{\boldsymbol{Y}}_{n\times p} = \underbrace{\boldsymbol{Z}}_{n\times(r+1)}\ \underbrace{\boldsymbol{\beta}}_{(r+1)\times p} + \underbrace{\boldsymbol{\varepsilon}}_{n\times p}$:

- $\mathrm{E}(\boldsymbol{\varepsilon}_j) = \mathbf{0}$, e
- $\mathrm{Cov}(\boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_k) = \sigma_{jk}\boldsymbol{I}_n$, para $j, k = 1, 2, \ldots, p$.

As p medidas sobre a i-ésima observação têm matriz de covariância dada por $\boldsymbol{\Sigma}$, mas medidas provenientes observações diferentes são não correlacionadas.

Temos que $\boldsymbol{\beta}$ e $\boldsymbol{\Sigma}$ são desconhecidos.

Observe que a j-ésima coluna da matriz resposta segue o modelo linear univariado múltiplo dado por

$$\boldsymbol{y}_j = \boldsymbol{Z}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \quad \text{para} \quad j = 1, 2, \ldots, p.$$

com $\mathrm{Var}(\boldsymbol{\varepsilon}_j) = \sigma_{jj}\boldsymbol{I}_n$.

Estimação por Mínimos Quadrados:

$$\boldsymbol{B} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{Y}.$$

Matriz de somas de quadrados e produtos cruzados:

$$(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{B})^\top (\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{B}).$$

Valores ajustados:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{Z}\boldsymbol{B} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{Y}$$

Resíduos:

$$\widehat{\varepsilon} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{B} = [\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top] \boldsymbol{Y}.$$

Propriedades:

$$\boldsymbol{Z}^\top \widehat{\varepsilon} = \underbrace{\boldsymbol{0}}_{(r+1)\times p} \quad \text{e} \quad \widehat{\boldsymbol{Y}}^\top \widehat{\varepsilon} = \underbrace{\boldsymbol{0}}_{p\times p};$$

$$\boldsymbol{Y}^\top \boldsymbol{Y} = \widehat{\boldsymbol{Y}}^\top \widehat{\boldsymbol{Y}} + \widehat{\varepsilon}^\top \widehat{\varepsilon}; \quad \text{e}$$

$$\widehat{\varepsilon}^\top \widehat{\varepsilon} = \boldsymbol{Y}^\top \boldsymbol{Y} - \boldsymbol{B}^\top (\boldsymbol{Z}^\top \boldsymbol{Z}) \boldsymbol{B}.$$

Para o estimador de mínimos quadrados $\boldsymbol{B}$ e com a matriz $\boldsymbol{Z}$ de posto completo, tem-se

$$\begin{aligned} \mathrm{E}(\boldsymbol{B}) &= \boldsymbol{\beta}, \\ \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_j, \widehat{\boldsymbol{\beta}}_k) &= \sigma_{jk}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}, \\ \mathrm{E}(\widehat{\boldsymbol{\varepsilon}}) &= \underbrace{\mathbf{0}}_{\text{matriz nula}}, \quad \text{e} \\ \mathrm{E}\left(\frac{1}{n-r-1}\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}\right) &= \boldsymbol{\Sigma}. \end{aligned}$$

$\boldsymbol{B}$ e $\widehat{\boldsymbol{\varepsilon}}$ são não correlacionados.

Estimador não tendencioso de $\boldsymbol{\Sigma}$:

$$\boldsymbol{S} = \frac{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}}{n-r-1}.$$

Inferências a partir da função de regressão estimada

Suponha que o modelo $\boldsymbol{Y} = \boldsymbol{Z}\beta + \varepsilon$, com erros normais tenha sido ajustado. Se o ajuste for considerado bom, ele poderá ser usado para fins de previsão.

Um problema é prever a média correspondente a um vetor de covariáveis $\boldsymbol{z}_0$.

Inferências sobre a média podem ser feitas se usando os mesmos resultados estudados no caso univariado.

$$\boldsymbol{B}^\top \boldsymbol{z}_0 \sim \mathcal{N}_p(\beta^\top \boldsymbol{z}_0, \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0 \boldsymbol{\Sigma})$$

e

$$n\widehat{\boldsymbol{\Sigma}} \sim \mathcal{W}_p(n - r - 1, \boldsymbol{\Sigma}), \quad \text{independentemente.}$$

Intervalos-T^2 simultâneos de 100(1-α)% de confiança para $\mathrm{E}(y_k|\boldsymbol{z}_0) = \boldsymbol{z}_0^\top \beta_k$:

$$\boldsymbol{z}_0^\top \widehat{\beta}_k \pm \sqrt{\frac{p(n-r-1)}{n-r-p} F_{p,n-r-p,1-\alpha}} \sqrt{\boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0 \left(\frac{n}{n-r-1} \widehat{\sigma}_{kk} \right)},$$

para $k = 1, 2, \ldots, p$.

O outro problema está voltado para a previsão de uma nova resposta $\boldsymbol{y}_0$ dado o vetor de covariáveis $\boldsymbol{z}_0$.

Agora,

$$\boldsymbol{y}_0 \sim \mathcal{N}_p(\boldsymbol{\beta}^\top \boldsymbol{z}_0, (1 + \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0)\boldsymbol{\Sigma}).$$

E, assim, intervalos simultâneos de previsão de 100(1-α)% para a k-ésima componente de $\boldsymbol{y}_0$, $y_{0,k}$, $k = 1, 2, \ldots, p$, são dados por

$$\boldsymbol{z}_0^\top \widehat{\boldsymbol{\beta}}_k \pm \sqrt{\frac{p(n-r-1)}{n-r-p} F_{p,n-r-p,1-\alpha}} \sqrt{(1 + \boldsymbol{z}_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{z}_0) \left(\frac{n}{n-r-1} \widehat{\sigma}_{kk} \right)},$$

O conceito de regressão linear

O modelo de regressão linear clássico diz respeito à associação entre uma variável dependente (ou resposta) y e uma coleção de covariáveis $z_1, z_2, \ldots, z_r$.

O modelo que consideramos trata y como uma variável aleatória cuja média depende dos valores fixados das covariáveis.

A média é uma função linear dos coeficientes da regressão $\beta_0, \beta_1, \ldots, \beta_r$.

O modelo de regressão linear também surge em uma configuração diferente.

Suponha que todas as variáveis $y, z_1, z_2, \ldots, z_r$ são variáveis aleatórias e têm uma distribuição conjunta com média $\boldsymbol{\mu}$ de dimensão $(r+1) \times 1$ e variância $\boldsymbol{\Sigma}$ de dimensão $(r+1) \times (r+1)$.

Particionando $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ de forma apropriada temos

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_z \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{yy} & \boldsymbol{\Sigma}_{yz} \\ \boldsymbol{\Sigma}_{zy} & \boldsymbol{\Sigma}_{zz} \end{bmatrix},$$

com

$$\boldsymbol{\Sigma}_{yz} = [\sigma_{yz_1}, \sigma_{yz_2}, \ldots, \sigma_{yz_r}].$$

$\boldsymbol{\Sigma}_{zz}$ é suposta ser não singular. Se isso não acontecer, significa que pelo menos uma das covariáveis é combinação linear das demais, de modo que esta covariável é redundante e pode ser eliminada do problema.

Preditor Linear:

$$\widehat{\beta}_0 + \widehat{\beta}_1 z_1 + \cdots + \widehat{\beta}_r z_r = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}, \quad \text{com} \quad \widehat{\beta}^\top = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_r).$$

Para um preditor dessa forma, o erro de previsão é dado por:

$$y - \widehat{\beta}_0 - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}.$$

Como este erro é aleatório, costuma-se escolher $\widehat{\beta}_0$ e $\widehat{\boldsymbol{\beta}}$ de modo a minimizar o erro quadrático médio

$$\text{EQM} = \text{E}\left([y - \widehat{\beta}_0 - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}]^2 \right).$$

O EQM depende da distribuição conjunta de y e $\boldsymbol{z}$ somente através de $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$.

é possível expressar o preditor linear ótimo em função dessas quantidades.

Resultado

O preditor linear $\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{z}$ com coeficientes $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{zy}$, $\beta_0 = \mu_y - \boldsymbol{\beta}^\top \boldsymbol{\mu}_z$ tem EQM mínimo entre todos os preditores lineares de y. Além disso,

$$\mathrm{E}\left([y - \widehat{\beta}_0 - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}]^2\right) = \sigma_{yy} - \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{zy}.$$

Também $\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{z}$ é o preditor linear de maior correlação com y dada por

$$\sqrt{\frac{\boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{zy}}{\sigma_{yy}}}.$$

Esse coeficiente de correlação é chamado coeficiente de correlação múltiplo da população e será denotado por $\rho_{y(z)}$.

$$\rho_{y(z)} = \sqrt{\frac{\boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{zy}}{\sigma_{yy}}}.$$

Observe que este coeficiente de correlação múltiplo, diferente dos outros varia entre 0 e 1.

O quadrado do coeficiente de correlação múltiplo populacional é chamado **coeficiente de determinação** populacional.

O coeficiente de determinação populacional tem uma interpretação importante. O EQM em usar $\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{z}$ para prever y é

$$\sigma_{yy} - \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{zy} = \sigma_{yy} - \sigma_{yy}\rho_{y(z)}^2 = \sigma_{yy}(1 - \rho_{y(z)}^2).$$

Se $\rho_{y(z)} = 0$, $\boldsymbol{z}$ não acrescenta nenhum poder de previsão. No outro extremo, y pode ser perfeitamente explicada por $\boldsymbol{z}$.

Previsão de várias variáveis

A extensão destes resultados para a previsão de várias respostas $y_1, y_2, \ldots, y_p$ é quase imediata.

Suponha

$$\begin{pmatrix} \boldsymbol{y}_{p\times 1} \\ \boldsymbol{z}_{r\times 1} \end{pmatrix} \sim \mathcal{N}_{p+r}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

com

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_y} \\ \boldsymbol{\mu_z} \end{pmatrix} \quad \text{e} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_{yy}} & \boldsymbol{\Sigma_{yz}} \\ \boldsymbol{\Sigma_{zy}} & \boldsymbol{\Sigma_{zz}} \end{bmatrix}.$$

Vimos que

$$\mathrm{E}(\boldsymbol{y}|z_1, z_2, \ldots, z_r] = \boldsymbol{\mu_y} + \boldsymbol{\Sigma_{yz}}\boldsymbol{\Sigma_{zz}}^{-1}(\boldsymbol{z} - \boldsymbol{\mu_z}).$$

Este valor esperado condicional, considerado como uma função de $z_1, z_2, \ldots, z_r$, é chamado vetor de regressão multivariada de $\boldsymbol{y}$ em $\boldsymbol{z}$.

Ele se compõe de p regressões múltiplas univariadas.

Por exemplo, o primeiro componente do vetor de média condicional é:

$$\mu_{y_1} + \boldsymbol{\Sigma}_{y_1 \boldsymbol{z}} \boldsymbol{\Sigma}_{\boldsymbol{zz}}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_{\boldsymbol{z}}) = \mathrm{E}(y_1 | z_1, z_2, \ldots, z_r)$$

que minimiza o erro de previsão nesse componente.

A matriz $\beta_{p \times r} = \boldsymbol{\Sigma}_{\boldsymbol{yz}} \boldsymbol{\Sigma}_{\boldsymbol{zz}}^{-1}$ é chamada matriz de coeficientes de regressão.

O vetor de erro de previsão é:

$$\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{\Sigma}_{\boldsymbol{yz}} \boldsymbol{\Sigma}_{\boldsymbol{zz}}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_{\boldsymbol{z}}) \quad \text{e} \quad \mathrm{E}(\boldsymbol{v}\boldsymbol{v}^{\top}) = \boldsymbol{\Sigma}_{\boldsymbol{yy.z}} = \boldsymbol{\Sigma}_{\boldsymbol{yy}} - \boldsymbol{\Sigma}_{\boldsymbol{yz}} \boldsymbol{\Sigma}_{\boldsymbol{zz}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{zy}}.$$

Como $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ costumam ser desconhecidos, eles devem ser estimados.

Os estimadores de máxima verossimilhança da função de regressão são dados por:

$$\overline{\boldsymbol{y}} + \boldsymbol{S}_{yz}\boldsymbol{S}_{zz}^{-1}(\boldsymbol{z} - \overline{\boldsymbol{z}}), \quad \text{e}$$

$$\widehat{\boldsymbol{\Sigma}}_{yy.z} = \frac{n-1}{n}(\boldsymbol{S}_{yy} - \boldsymbol{S}_{yz}\boldsymbol{S}_{zz}^{-1}\boldsymbol{S}_{zy}).$$

Observação: a coleção $y_1, y_2, \ldots, y_p, z_1, z_2, \ldots, z_r$ leva as seguintes equações de previsão:

$$\begin{array}{rcl}
\widehat{y}_1 & = & \widehat{\beta}_{01} + \widehat{\beta}_{11}z_1 + \cdots + \widehat{\beta}_{r1}z_r \\
\widehat{y}_2 & = & \widehat{\beta}_{02} + \widehat{\beta}_{12}z_1 + \cdots + \widehat{\beta}_{r2}z_r \\
\vdots & \vdots & \vdots \\
\widehat{y}_p & = & \widehat{\beta}_{0p} + \widehat{\beta}_{1p}z_1 + \cdots + \widehat{\beta}_{rp}z_r.
\end{array}$$

1. Os mesmos valores $z_1, z_2, \ldots, z_r$ são usados para prever cada y_i.
2. Os $\widehat{\beta}_{jk}$ são estimativas das entradas (j, k) da matriz $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}$.

Coeficiente de correlação parcial

Considere $\begin{pmatrix} \boldsymbol{y}_{p\times 1} \\ \boldsymbol{z}_{r\times 1} \end{pmatrix} \sim \mathcal{N}_{p+r}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

O coeficiente de correlação parcial entre y_j e y_k, eliminando-se $\boldsymbol{z}$, é definido por

$$\rho_{y_j y_k . \boldsymbol{z}} = \frac{\sigma_{y_j y_k . \boldsymbol{z}}}{\sqrt{\sigma_{y_j y_j . \boldsymbol{z}} \sigma_{y_k y_k . \boldsymbol{z}}}}$$

com $\sigma_{y_j y_k . \boldsymbol{z}}$ a entrada (j,k) da matriz $\boldsymbol{\Sigma}_{\boldsymbol{yy}.\boldsymbol{z}} = \boldsymbol{\Sigma}_{\boldsymbol{yy}} - \boldsymbol{\Sigma}_{\boldsymbol{yz}} \boldsymbol{\Sigma}_{\boldsymbol{zz}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{zy}}$.

O coeficiente de correlação amostral correspondente é dado por

$$r_{y_j y_k .} = \frac{s_{y_j y_k . \boldsymbol{z}}}{\sqrt{s_{y_j y_j . \boldsymbol{z}} s_{y_k y_k . \boldsymbol{z}}}}$$

com $s_{y_j y_k . \boldsymbol{z}}$ a entrada (j,k) da matriz $\boldsymbol{S}_{\boldsymbol{yy}.\boldsymbol{z}} = \boldsymbol{S}_{\boldsymbol{yy}} - \boldsymbol{S}_{\boldsymbol{yz}} \boldsymbol{S}_{\boldsymbol{zz}}^{-1} \boldsymbol{S}_{\boldsymbol{zy}}$.

Análise em componentes principais

O propósito da **análise de componentes principais** (ACP) é substituir as variáveis originais por um número menor de variáveis que são funções 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.

Em geral, espera-se que os primeiras componentes chamados **componentes principais** (em número menor do que o de variáveis originais) compreendam a maior parte da variação total no conjunto de dados original tal que a dimensionalidade efetiva dos dados pode ser reduzida.

A **análise fatorial** (AF) tem um propósito similar, mas é baseada em um modelo estatístico próprio que especifica um dado número de variáveis subjacentes chamadas fatores. A AF é uma técnica voltada para a "explicação" da estrutura de covariâncias das variáveis, em vez de olhar apenas as variâncias.

Devido às similaridades das duas técnicas os dois métodos são algumas vezes confundidos. Porém, existem diferenças fundamentais que serão apresentadas ao longo deste curso.

Em muitas aplicações, os componentes principais ou fatores aparecem como o objetivo final da análise e, os pesquisadores tentam, então, interpretá-los de forma significativa. Porém, o principal benefício da ACP está na possibilidade de reduzir a dimensão do problema de forma a simplificar as análises subsequentes.

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 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) 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*.

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.

Por exemplo, construir o gráfico de dispersão dos escores dos dois primeiros componentes para cada observação é uma forma de tentar identificar conglomerados (agrupamentos).

A ACP transforma um conjunto de variáveis correlacionadas em um conjunto de variáveis não correlacionadas. Se as variáveis originais são aproximadamente não correlacionadas entre si, então não existe razão para realizarmos uma ACP.

Componentes principais populacionais

Algebricamente, CP 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 ao rotacionar o sistema original. Os novos eixos representam as direções com variabilidade máxima e fornecem descrição mais simples e parcimoniosa da estrutura de covariância.

Os CP 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.

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\top$ um vetor aleatório p-dimensional, com média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. Desejamos encontrar um novo conjunto de variáveis, $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_p)^\top$, 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 tomado 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 $\boldsymbol{a}_j = (a_{j1}, a_{j2}, \ldots, a_{jp})^\top$ 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 ao escolher $\boldsymbol{a}_1$ 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 ao escolher $\boldsymbol{a}_2$ 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.

Definição

Se $\boldsymbol{A}$ é uma matriz $p \times p$ e $\boldsymbol{v}$ é um vetor não nulo de tamanho p tal que

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}, \quad \text{com} \quad \lambda \geqslant 0,$$

então dizemos que λ é um autovalor correspondente a $\boldsymbol{v}$. Da teoria de decomposição espectral de $\boldsymbol{A}$, temos que $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top$ com $\boldsymbol{P} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ e $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$.

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:

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

e

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

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

Prova

Vimos que

$$\max_{\boldsymbol{a}\neq\mathbf{0}} \frac{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{a}} = \lambda_1.$$

O máximo é atingido quando $\boldsymbol{a} = \boldsymbol{v}_1$.

Mas $\boldsymbol{v}_1^\top \boldsymbol{v}_1 = 1$ e, assim,

$$\max_{\boldsymbol{a}\neq\mathbf{0}} \frac{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{a}} = \lambda_1 = \frac{\boldsymbol{v}_1^\top \boldsymbol{\Sigma} \boldsymbol{v}_1}{\boldsymbol{v}_1^\top \boldsymbol{v}_1} = \boldsymbol{v}_1^\top \boldsymbol{\Sigma} \boldsymbol{v}_1 = \text{Var}(Y_1).$$

Similarmente, vimos que

$$\max_{\boldsymbol{a}\perp \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k} \frac{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{a}} = \lambda_{k+1} \quad \text{para} \quad k = 1, 2, \ldots, p-1.$$

Para a escolha $\boldsymbol{a} = \boldsymbol{v}_{k+1}$ com $\boldsymbol{v}_{k+1}^\top \boldsymbol{v}_j = 0$, para $j = 1, 2, \ldots, k$ e $k = 1, 2, \ldots, p-1$,

$$\text{Var}(Y_{k+1}) = \lambda_{k+1} \quad \text{e} \quad \text{Cov}(Y_{k+1}, Y_j) = 0, \quad j = 1, 2, \ldots, k.$$

Proposição 2

Seja $\boldsymbol{X}$ um 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,

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

Prova

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

Logo,

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

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

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

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 $\boldsymbol{v}_j = (v_{j1}, v_{j2}, \ldots, v_{jp})^\top$ que definem os componentes 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

$$\rho(Y_j, X_k) = \frac{v_{jk}\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}, \quad \text{para} \quad j, k = 1, 2, \ldots, p,$$

são os coeficientes de correlação entre Y_j e X_k.

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

Prova

Faça $\boldsymbol{e}_k = (0, \ldots, 0, \underbrace{1}_{k\text{-ésima entrada}}, 0, \ldots, 0)^\top$ tal que $X_k = \boldsymbol{e}_k^\top \boldsymbol{X}$ e

$$\begin{aligned}
\text{Cov}(X_k, Y_j) &= \text{Cov}(\boldsymbol{e}_k^\top \boldsymbol{X}, \boldsymbol{v}_j^\top \boldsymbol{X}) = \boldsymbol{e}_k^\top \boldsymbol{\Sigma} \boldsymbol{v}_j = \boldsymbol{e}_k^\top \boldsymbol{v}_j \lambda_j = v_{jk}\lambda_j \Rightarrow \\
\rho(Y_j, X_k) &= \frac{\text{Cov}(Y_j, X_k)}{\sqrt{\text{Var}(Y_j)}\sqrt{\text{Var}(X_k)}} = \frac{v_{jk}\lambda_j}{\sqrt{\lambda_j}\sqrt{\sigma_{kk}}} = \frac{v_{jk}\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}.
\end{aligned}$$

Exemplo

Suponha um vetor aleatório de dimensão 3 cuja matriz de covariâncias é dada por

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

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

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

Portanto, os componentes principais são:

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

Ver arquivo exemplo_13.r

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 5: 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

ACP considerando uma normal multivariada

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

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

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 $(1/\lambda_1, \boldsymbol{v}_1), (1/\lambda_2, \boldsymbol{v}_2), \ldots, (1/\lambda_p, \boldsymbol{v}_p)$. Note que $\boldsymbol{\Sigma}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}$.

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 $\boldsymbol{v}_j$.

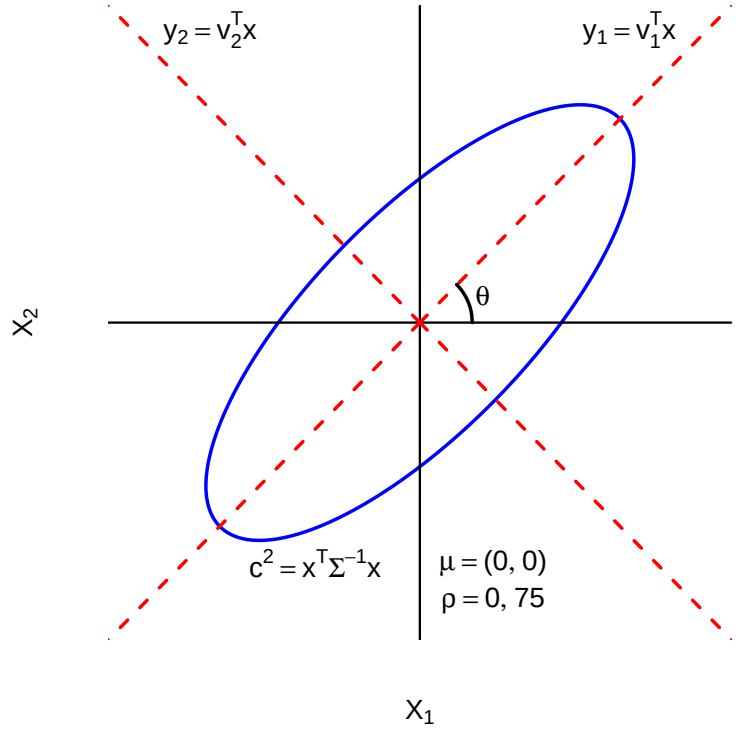

Figura 7: Elipse de densidade constante e os componentes principais para um vetor aleatório normal bivariado com $\boldsymbol{\mu} = \mathbf{0}$ e $\rho = 0,75$. (Ver arquivo exemplo_14.r)

É possível ver que os CP 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

$$\boldsymbol{Z} = \left[\boldsymbol{\Psi}^{1/2}\right]^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \quad \text{com} \quad \boldsymbol{\Psi}^{1/2} = \text{diag}(\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \ldots, \sqrt{\sigma_{pp}}).$$

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 δ_j/p, para $j = 1, 2, \ldots, p$, em que δ_j é 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

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

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

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

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

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

Ver arquivo exemplo_15.r

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.

Discussão

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.

ACP para matrizes de covariância com estruturas especiais

Existem certos padrões das matrizes de covariâncias e correlações cujos componentes principais podem ser expressos de forma simples. Por exemplo, suponha que

$$\boldsymbol{\Sigma} = \text{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp}).$$

Fazendo $\boldsymbol{e}_j = (0, \ldots, 0, \underbrace{1}_{j\text{-ésima posição}}, 0, \ldots, 0)^\top$ é fácil ver que $\boldsymbol{\Sigma}\boldsymbol{e}_j = \sigma_{jj}\boldsymbol{e}_j$ e concluímos que os pares $(\sigma_{jj}, \boldsymbol{e}_j)$, $j = 1, 2, \ldots, p$ são pares de autovalores e autovetores de $\boldsymbol{\Sigma}$. Como a combinação linear $\boldsymbol{e}_j^\top \boldsymbol{X} = X_j$, o conjunto de componentes principais é exatamente o conjunto de variáveis originais não correlacionadas, exceto que agora apresentadas em ordem decrescente de variância.

Para uma matriz de covariâncias diagonal, nada se ganha em usar ACP. De outro ponto de vista, se $\boldsymbol{X}$ tem distribuição $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, os contornos de densidade constante são elipsóides cujos eixos já estão na direção da variação máxima. Consequentemente, não há necessidade de rotacionar o sistema de coordenadas originais.

Padronização não altera substancialmente a situação para $\boldsymbol{\Sigma}$ diagonal. Neste caso $\boldsymbol{\mathcal{R}} = \boldsymbol{I}_p$. Claramente, $\boldsymbol{\mathcal{R}}\boldsymbol{e}_j = 1 \times \boldsymbol{e}_j$, tal que o autovalor 1 é de multiplicidade p e

$$\boldsymbol{e}_j = (0, \ldots, 0, \underbrace{1}_{j\text{-ésima posição}}, 0, \ldots, 0)^\top, \quad j = 1, 2, \ldots, p,$$

é uma escolha conveniente para os autovetores. Consequentemente, os componentes principais obtidos por $\boldsymbol{\mathcal{R}}$ são também as variáveis originais, agora padronizadas, $Z_1, Z_2, \ldots, Z_p$. No caso de autovalores iguais e sob normalidade, os contornos de densidade constante serão esferóides.

Outra matriz de covariância padrão, que em geral descreve correspondência entre certas variáveis biológicas tem a seguinte forma geral

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix}.$$

A matriz de correlações correspondente é

$$\boldsymbol{\mathcal{R}} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

Observe que com este padrão as variáveis $X_1, X_2, \ldots, X_p$ são igualmente correlacionadas.

Não é difícil mostrar que os p autovalores correspondentes à matriz de correlações podem ser divididos em dois grupos.

Quando $\rho > 0$, o maior autovalor é $\lambda_1 = 1 + (p-1)\rho$ com autovetor associado

$$\boldsymbol{v}_1 = \left(\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}} \right)^\top .$$

Os $p-1$ autovalores restantes são $\lambda_2 = \lambda_3 = \cdots = \lambda_p = 1 - \rho$ e uma escolha apropriada para os autovetores é

$$\begin{aligned}
\boldsymbol{v}_2 &= \left(\frac{1}{\sqrt{1 \times 2}}, \frac{-1}{\sqrt{1 \times 2}}, 0, \ldots, 0 \right)^\top \\
\boldsymbol{v}_3 &= \left(\frac{1}{\sqrt{2 \times 3}}, \frac{1}{\sqrt{2 \times 3}}, \frac{-2}{\sqrt{2 \times 3}}, 0, \ldots, 0 \right)^\top \\
\vdots \quad \vdots & \qquad\qquad \vdots \\
\boldsymbol{v}_j &= \left(\frac{1}{\sqrt{(j-1) \times j}}, \ldots, \frac{1}{\sqrt{(j-1) \times j}}, \frac{-(j-1)}{\sqrt{(j-1) \times j}}, 0, \ldots, 0 \right)^\top \\
\vdots \quad \vdots & \qquad\qquad \vdots \\
\boldsymbol{v}_p &= \left(\frac{1}{\sqrt{(p-1) \times p}}, \ldots, \frac{1}{\sqrt{(p-1) \times p}}, \frac{-(p-1)}{\sqrt{(p-1) \times p}} \right)^\top .
\end{aligned}$$

O primeiro componente principal é

$$Y_1 = \boldsymbol{v}_1^\top \boldsymbol{Z} = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} Z_i$$

que é proporcional à soma das p variáveis padronizadas. Ele pode ser olhado como um "índice" com pesos iguais para todas as variáveis. A proporção da variação total explicada por este componente é dada por

$$\frac{\lambda_1}{p} = \frac{1 + (p-1)\rho}{p} = \rho + \frac{1-\rho}{p}.$$

Se ρ estiver próximo de 1 ou p for grande temos que $\dfrac{\lambda_1}{p} \simeq \rho$. Por exemplo, se $\rho = 0,80$ e $p = 5$, este componentes explicará cerca de 84% da variação total do dados.

Quando ρ está perto de 1, os $p - 1$ autovalores restantes contribuem relativamente muito pouco e podem ser desprezados.

Se o vetor aleatório $(Z_1, Z_2, \ldots, Z_p)$ tem distribuição normal multivariada com correlações iguais duas a duas, então o elipsóide de densidade constante terá uma forma de charuto.

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

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.

O valor do j-ésimo CP para esta observação é $\widehat{y}_j = \widehat{\boldsymbol{v}}_j^\top \boldsymbol{x}$.

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

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

Denotaremos

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

Exemplo

Um censo forneceu informações sobre 5 variáveis socioeconômicas em 61 setores censitários na área de Madison, Wisconsin.

As variáveis observadas foram as seguintes:

- 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?

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

Tabela 6: Médias amostrais das variáveis.

Ver exemplo_16.r

	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

Tabela 7: Matriz de covariâncias 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

Tabela 8: Matriz de correlações amostrais das variáveis.

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

Os **desvios padrões** são dados por:

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

	$\widehat{\boldsymbol{v}}_1$	$\widehat{\boldsymbol{v}}_2$	$\widehat{\boldsymbol{v}}_3$	$\widehat{\boldsymbol{v}}_4$	$\widehat{\boldsymbol{v}}_5$
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

Tabela 9: Rotação das variáveis.

	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

Tabela 10: Importância das componentes.

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.

	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

Tabela 11: Correlações entre os componentes e as variáveis

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 (continuação)

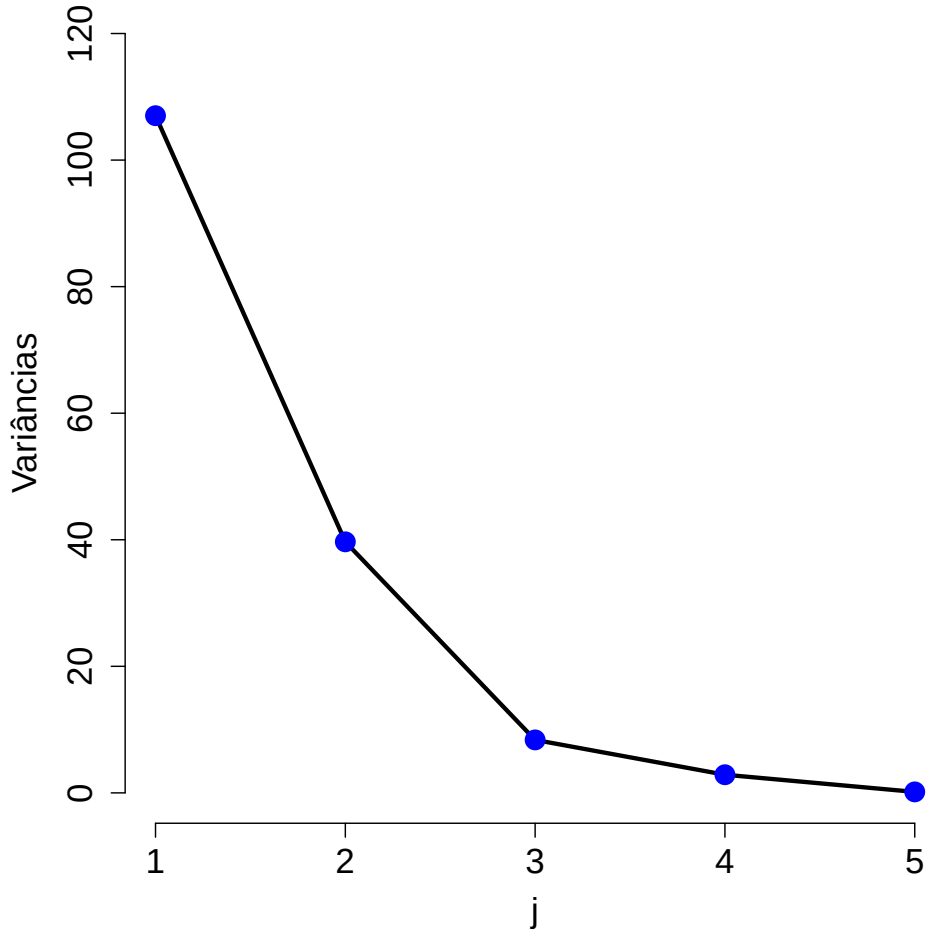

Figura 8: Autovalores ordenados de forma não crescente.

Componentes principais usando as variáveis padronizadas

Os componentes principais podem ser obtidos via matriz de correlações amostrais $\boldsymbol{R}$ 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

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.

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 se mover junto em resposta à conjuntura econômica.

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

Tabela 12: Médias amostrais das variáveis.

Ver exemplo_17.r

	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

Tabela 13: Matriz de correlações amostrais das variáveis.

	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

Tabela 14: Importância das componentes.

	$\widehat{\boldsymbol{v}}_1$	$\widehat{\boldsymbol{v}}_2$	$\widehat{\boldsymbol{v}}_3$	$\widehat{\boldsymbol{v}}_4$	$\widehat{\boldsymbol{v}}_5$
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

Tabela 15: Rotação das variáveis.

- 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, constatamos 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 e as variáveis é dada a seguir.

	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

Tabela 16: Correlações entre os componentes e as variáveis.

A seguir mostramos o ***scree plot***.

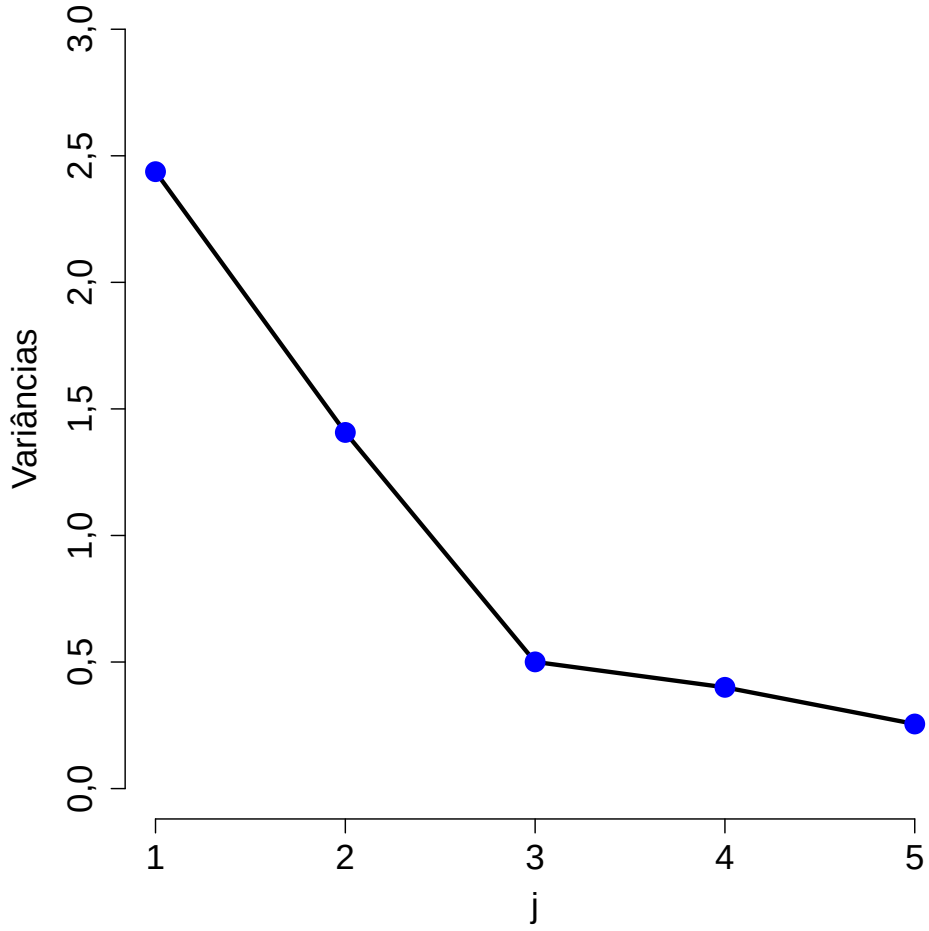

Figura 9: Autovalores ordenados de forma não crescente.

Gráficos dos componentes principais

Gráficos dos CP podem revelar observações suspeitas bem como desvios da suposição de normalidade. Como os CP são combinações lineares das variáveis originais, é razoável esperar que eles se comportem 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 são aproximadamente normais, quando eles são usados como dados de entrada em análises posteriores.

Os últimos CP 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.

Assim, as magnitudes dos últimos CP 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 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. Eles ajudam a identificar observações suspeitas.

Exemplo (continuação)

Usando os dados sobre ações, podemos construir os gráficos de dispersão dos dois primeiros CP e respectivos *qq-plots* para normalidade, e para os dois últimos CP para identificar a presença de observações suspeitas.

Ver exemplo_18.r

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 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 $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^\top$. 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) $\sqrt{n}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \overset{a}{\sim} \mathcal{N}_p(\mathbf{0}, 2\boldsymbol{\Lambda}^2)$.

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

(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 $\widehat{\lambda}_j$ são independentemente distribuídos, isto é

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

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

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

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 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 $\frac{1}{n}\widehat{\boldsymbol{V}}_j$ em que os elementos de $\widehat{\boldsymbol{V}}_j$ são obtidos a partir de $\boldsymbol{V}_j$ substituindo suas entradas por suas estimativas $\widehat{\lambda}_j$ e $\widehat{\boldsymbol{v}}_j$.

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

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

Assim, os IC simultâneos são dados por

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

Teste para estrutura de correlações iguais

$$\mathrm{H}_0: \ \boldsymbol{\mathcal{R}} = \boldsymbol{\mathcal{R}}_0 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \quad \text{versus} \quad \mathrm{H}_1: \ \boldsymbol{\mathcal{R}} \neq \boldsymbol{\mathcal{R}}_0.$$

É possível construir um teste destas hipóteses usando a estatística da razão de verossimilhança. Porém, apresentaremos um teste equivalente, devido à Lawley, que é construído a partir dos elementos fora da diagonal de $\boldsymbol{R}$.

Faça

$$\begin{aligned} \bar{r}_k &= \frac{1}{p-1}\sum_{i=1, i\neq k}^{p} r_{ik}, \quad k = 1, 2, \ldots, p, \\ \bar{r} &= \frac{2}{p(p-1)}\sum\sum_{i<k} r_{ik}, \quad \text{e} \\ \widehat{\gamma} &= \frac{(p-1)^2[1-(1-\bar{r})^2]}{p-(p-2)(1-\bar{r})^2}. \end{aligned}$$

É evidente que $\overline{r}_k$ é a média dos elementos fora da diagonal na k-ésima coluna de $\boldsymbol{R}$ e $\overline{r}$ é a média global dos elementos fora da diagonal de $\boldsymbol{R}$.

Testes aproximados para grandes amostras rejeitam H_0 ao nível de significância α se

$$T > \chi^2_{(p+1)(p-2)/2,1-\alpha} \quad \text{com}$$

$$T = \frac{n-1}{(1-\overline{r})^2}\left[\sum\sum_{i<k}(r_{ik}-\overline{r})^2 - \widehat{\gamma}\sum_{k=1}^{p}(\overline{r}_k-\overline{r})^2\right].$$

Exemplo

Desejamos fazer um teste para a estrutura de equicorrelação. Para isto, temos

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0,7501 & 0,6329 & 0,6363 \\ 0,7501 & 1 & 0,6925 & 0,7386 \\ 0,6329 & 0,6925 & 1 & 0,6625 \\ 0,6363 & 0,7386 & 0,6625 & 1 \end{bmatrix},$$

$\overline{r}_1 = 0,6731$, $\overline{r}_2 = 0,7271$, $\overline{r}_3 = 0,66626$, $\overline{r}_4 = 0,6791$, $\overline{r} = 0,6855$ e $\widehat{\gamma} = 2,1329$.

Fazendo os cálculos, obtemos $T \simeq 11,4$.

Temos também $\chi^2_{5;\ 0,95} \simeq 11,07$.

Logo, rejeitamos H_0 ao nível de significância de 5%.

Supondo população normal multivariada, existe um teste assintótico para verificar se todas as variáveis são independentes ($\boldsymbol{\mathcal{R}} = \boldsymbol{I}_p$).

Análise fatorial

O propósito fundamental da **análise fatorial (AF)** é descrever, se possível, as relações de covariâncias entre muitas variáveis em função de umas poucas, porém não observáveis, quantidades aleatórias subjacentes que são chamadas **fatores**.

O modelo de AF é motivado pelo seguinte argumento.

Suponha que as variáveis podem ser agrupadas por suas correlações. Isto é, suponha que todas as variáveis dentro de um grupo particular sejam altamente correlacionadas entre si, mas apresentam correlações relativamente pequenas com variáveis de outros grupos.

Então, é razoável propor um único constructo (fator) subjacente para cada grupo, que resume as correlações observadas no grupo.

AF pode ser considerada uma extensão da ACP. Ambas podem ser vistas como voltadas para aproximar a matriz de covariâncias $\mathbf{\Sigma}$. Porém, a aproximação baseada no modelo AF é mais elaborada.

A questão primária na AF é se os dados são consistentes com uma estrutura prescrita.

Modelo fatorial ortogonal

Seja $\boldsymbol{X}$ vetor aleatório observável $p \times 1$, com $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X})$ e $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{X})$.

O modelo fatorial (MF) postula que $\boldsymbol{X}$ é linearmente dependente de $m < p$ variáveis aleatórias não observáveis $F_1, F_2, \ldots, F_m$ chamadas fatores comuns, e p fontes de variação adicionais $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ chamadas erros ou, algumas vezes, fatores específicos.

$$X_j - \mu_j = \ell_{j1} F_1 + \ell_{j2} F_2 + \cdots + \ell_{jm} F_m + \varepsilon_j, \quad \text{para} \quad j = 1, 2, \ldots, p,$$

ou, em notação matricial,

$$\underbrace{\boldsymbol{X} - \boldsymbol{\mu}}_{p \times 1} = \underbrace{\boldsymbol{L}}_{p \times m} \underbrace{\boldsymbol{F}}_{m \times 1} + \underbrace{\boldsymbol{\varepsilon}}_{p \times 1},$$

em que

$$\begin{aligned}
\boldsymbol{\varepsilon} &= \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_p \end{bmatrix}^\top \quad \text{(vetor aleatório)}, \\
\boldsymbol{F} &= \begin{bmatrix} F_1 & F_2 & \cdots & F_m \end{bmatrix}^\top \quad \text{(vetor aleatório)}, \\
\boldsymbol{L} &= \begin{bmatrix} \boldsymbol{\ell}_1 & \boldsymbol{\ell}_2 & \cdots & \boldsymbol{\ell}_m \end{bmatrix} \quad \text{e} \\
\boldsymbol{\ell}_k &= \begin{bmatrix} \ell_{1k} & \ell_{2k} & \cdots & \ell_{pk} \end{bmatrix}^\top, \quad k = 1, 2, \ldots, m.
\end{aligned}$$

Os coeficientes ℓ_{jk} são chamados cargas da j-ésima variável sobre o k-ésimo fator ($j = 1, 2, \ldots, p$ e $k = 1, 2, \ldots, m$).

O fator específico ε_j está associado à j-ésima variável (X_j).

Os p desvios $X_j - \mu_j$ são expressos em função de $p + m$ quantidades aleatórias: $F_1, F_2, \ldots, F_m$ e $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$, não observáveis.

Isto distingue o modelo AF do modelo de regressão linear múltipla multivariado no qual as variáveis explicativas são observáveis.

Com tantas quantidades não observáveis, uma verificação do modelo fatorial a partir das observações sobre $X_1, X_2, \ldots, X_p$ é desanimadora.

Porém, com algumas suposições adicionais sobre os vetores aleatórios $\boldsymbol{F}$ e ε, o modelo AF implicará em certas relações de covariâncias que podem ser verificadas.

Suposições adicionais:

1. $\mathrm{E}(\boldsymbol{F}) = \mathbf{0}$ e $\mathrm{Var}(\boldsymbol{F}) = \mathrm{E}(\boldsymbol{F}\boldsymbol{F}^\top) = \boldsymbol{I}_m$;
2. $\mathrm{E}(\varepsilon) = \mathbf{0}$ e $\mathrm{Var}(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon^\top) = \boldsymbol{\Psi} = \mathrm{diag}\{\psi_1, \psi_2, \ldots, \psi_p\}$; e
3. $\boldsymbol{F}$ e ε são independentes tal que $\mathrm{Cov}(\boldsymbol{F}, \varepsilon) = \mathbf{0}_{m\times p}$.

Com estas suposições o modelo apresentado é chamado **modelo fatorial ortogonal (MFO)**.

O MFO implica a seguinte estrutura de covariâncias para $\boldsymbol{X}$:

$$\begin{aligned}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top &= (\boldsymbol{L}\boldsymbol{F}+\varepsilon)(\boldsymbol{L}\boldsymbol{F}+\varepsilon)^\top \\ &= \boldsymbol{L}\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{L}^\top + \boldsymbol{L}\boldsymbol{F}\varepsilon^\top + \varepsilon\boldsymbol{F}^\top\boldsymbol{L}^\top + \varepsilon\varepsilon^\top.\end{aligned}$$

tal que

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{Var}(\boldsymbol{X}) = \mathrm{E}\left((\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top\right) \\ &= \boldsymbol{L}\mathrm{E}(\boldsymbol{F}\boldsymbol{F}^\top)\boldsymbol{L}^\top + \boldsymbol{L}\mathrm{E}(\boldsymbol{F}\varepsilon^\top) + \mathrm{E}(\varepsilon\boldsymbol{F}^\top)\boldsymbol{L}^\top + \mathrm{E}(\varepsilon\varepsilon^\top) \\ &= \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}.\end{aligned}$$

Também, é fácil verificar que,

$$\text{Cov}(\boldsymbol{X}, \boldsymbol{F}) = \text{E}\left((\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{F}^\top\right) = \boldsymbol{L}.$$

Logo, o MFO implica em $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}$ e $\text{Cov}(\boldsymbol{X}, \boldsymbol{F}) = \boldsymbol{L}$ tal que
$\text{Cov}(X_j, f_k) = \ell_{jk}$ e $\text{Cor}(X_j, f_k) = \dfrac{\ell_{jk}}{\sqrt{\sigma_{jj}}}$, $j = 1, 2, \ldots, p$ e $k = 1, 2, \ldots, m$.

O modelo $\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \varepsilon$ é linear nos fatores comuns. Se as p respostas são, de fato, relacionadas aos fatores subjacentes, mas a relação não é linear, tal como

$X_1 - \mu_1 = \ell_{11} F_1 F_3 + \varepsilon_1, \quad , X_2 - \mu_2 = \ell_{22} F_2 F_3 + \epsilon_3, \quad$ etc.,

então a estrutura de covariâncias dada por $\boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}$ pode não ser adequada.

A suposição de linearidade é fundamental na formulação do modelo fatorial usual.

- A porção da variância da j-ésima variável contribuída pelos m fatores é chamada j-ésima **comunalidade**.
- A porção da variância devida ao fator específico ε_j é chamada **variância específica**.

Denotaremos a j-ésima comunalidade por h_j^2 e a variância específica por ψ_j. Assim,

$$\underbrace{\sigma_{jj}}_{\text{Var}(X_j)} = \underbrace{\ell_{j1}^2 + \ell_{j2}^2 + \cdots + \ell_{jm}^2}_{\text{comunalidade}} + \underbrace{\psi_j}_{\text{var. específica}},$$

$$h_j^2 = \sum_{k=1}^{m} \ell_{jk}^2 \quad \text{e} \quad \sigma_{jj} = h_j^2 + \psi_j.$$

A j-ésima comunalidade é a soma dos quadrados das cargas da j-ésima variável sobre os m fatores, $j = 1, 2, \ldots, p$.

Exemplo

Aqui, faremos uma verificação da relação $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}$ para $m = 2$.

$$\boldsymbol{\Sigma} = \begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 4 & 7 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix},$$

ou, equivalentemente, $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}$ com

$$\boldsymbol{L}^\top = \begin{bmatrix} 4 & 7 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{bmatrix} \quad \text{e} \quad \Psi = \text{diag}(2, 4, 1, 3).$$

Variável	Comunalidade	Var. específica	Variância
X_1	16+1=17	2	19
X_2	49+4=53	4	57
X_3	1+36=37	1	38
X_4	1+64=65	3	68

Ver arquivo exemplo_19.r

O MFO assume que os $p(p+1)/2$ elementos distintos da matriz de covariâncias de $\boldsymbol{X}$, $\boldsymbol{\Sigma}$, podem ser representados pelas pm cargas das p variáveis sobre os m fatores e as p variâncias específicas.

Por exemplo, se $p = 10$ e $m = 2$ o MFO simplifica a estrutura de covariâncias com 55 parâmetros desconhecidos para a estimação de 30.

Se $m = p$, qualquer matriz de covariâncias $\boldsymbol{\Sigma}$ pode ser decomposta em $\boldsymbol{LL}^{\top}$, assim $\boldsymbol{\Psi}$ pode ser a matriz nula.

Porém, é quando m é pequeno relativamente a p que a AF terá utilidade.

Neste caso, o MFO fornece uma explicação simples das covariâncias de $\boldsymbol{X}$ com menos parâmetros dos que os $p(p+1)/2$ em $\boldsymbol{\Sigma}$.

Mas, muitas das matrizes $\boldsymbol{\Sigma}$ não podem ser decompostas na forma $\boldsymbol{LL}^\top + \boldsymbol{\Psi}$ quando m é bem menor do que p.

Exemplo

Suponha $p = 3$, $m = 1$ e $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0,9 & 0,7 \\ 0,9 & 1 & 0,4 \\ 0,7 & 0,4 & 1 \end{bmatrix}$.

Usando o MFO a estrutura imposta para $\boldsymbol{\Sigma}$ é $\underbrace{\boldsymbol{L}}_{3\times 1}\boldsymbol{L}^\top + \boldsymbol{\Psi}$ ou

$$\begin{aligned} \ell_{11}^2 + \psi_1 &= 1 & \qquad \ell_{11}\ell_{21} &= 0,9 \\ \ell_{21}^2 + \psi_2 &= 1 & \qquad \ell_{11}\ell_{31} &= 0,7 \\ \ell_{31}^2 + \psi_3 &= 1 & \qquad \ell_{21}\ell_{31} &= 0,4 \end{aligned}$$

Resolvendo o sistema encontramos uma solução única, porém inconsistente com as interpretações estatísticas dadas aos parâmetros. Por exemplo, conclui-se que $\text{Cov}(X_1, f_1) = \text{Cor}(X_1, f_1) = \ell_{11} = \pm 1,255$!!!

Observe que $\text{Var}(X_1) = 1$ e $\text{Var}(F_1) = 1$.

Ver arquivo exemplo_20.r

Além disso, também obtem-se $\psi_1 = -0,575$!

Assim, para este exemplo, com $m = 1$ é possível obter uma solução numérica única para a equação $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}$. No entanto, a solução não é apropriada.

Quando $m > 1$, há sempre uma ambiguidade inerente associada ao modelo fatorial. Para ver isto, seja $\boldsymbol{Q}$ uma matriz ortogonal $m \times m$ tal que $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}_m$. Então,

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}^\star \boldsymbol{F}^\star + \boldsymbol{\varepsilon}.$$

com $\boldsymbol{L}^\star = \boldsymbol{L}\boldsymbol{Q}$ e $\boldsymbol{F}^\star = \boldsymbol{Q}^\top \boldsymbol{F}$.

Como $\mathrm{E}(\boldsymbol{F}^\star) = \boldsymbol{Q}^\top \mathrm{E}(\boldsymbol{F}) = \boldsymbol{0}$ e $\mathrm{Var}(\boldsymbol{F}^\star) = \boldsymbol{Q}^\top \mathrm{Var}(\boldsymbol{F})\boldsymbol{Q} = \boldsymbol{Q}^\top \boldsymbol{I}_m \boldsymbol{Q} = \boldsymbol{I}_m$ é impossível, com base nas observações sobre $\boldsymbol{X}$, distinguir as cargas em $\boldsymbol{L}$ das cargas em $\boldsymbol{L}^\star$. Isto é, os fatores $\boldsymbol{F}$ e $\boldsymbol{F}^\star$, têm as mesmas propriedades estatísticas e, apesar das cargas em $\boldsymbol{L}^\star$ serem, em geral, diferentes das cargas em $\boldsymbol{L}$, ambas geram a mesma matriz de covariâncias $\boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi} = (\boldsymbol{L}\boldsymbol{Q})(\boldsymbol{L}\boldsymbol{Q})^\top + \boldsymbol{\Psi} = [\boldsymbol{L}^\star][\boldsymbol{L}^\star]^\top + \boldsymbol{\Psi}.$$

Esta ambiguidade fornece a justificativa para o uso da “rotação de fatores”, pois as transformações lineares correspondentes à matrizes ortogonais correspondem à rotações (e reflexões) no sistema de coordenadas $\boldsymbol{X}$.

A análise do MFO prossegue se impondo condições que permitem estimar univocamente $\boldsymbol{L}$ e $\boldsymbol{\Psi}$.

A matriz de cargas $\boldsymbol{L}$ é, então, rotacionada (pré-multiplicada por uma matriz ortogonal). A rotação é determinada por algum critério de facilitação da interpretação dos fatores.

Obtidas as cargas e as variâncias específicas, os fatores são identificados, e valores estimados para os fatores em si, chamados **escores dos fatores**, são calculados.

Métodos de estimação

Dadas as observações $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, p-variadas, a AF busca responder a seguinte questão.

O MFO, com um número pequeno de fatores, representa adequadamente os dados?

Em essência, lidamos com esse problema tentando verificar a estrutura de covariâncias dos dados.

$\boldsymbol{S}$ é um estimador de $\boldsymbol{\Sigma}$. Se os elementos fora da diagonal de $\boldsymbol{S}$ são pequenos (fora da diagonal de $\boldsymbol{R}$ estão próximos de 0), as variáveis são não correlacionadas e uma AF não será útil.

Nestas circunstâncias, os fatores específicos representarão o papel principal, enquanto que o objetivo central da AF é determinar uns poucos fatores comuns importantes.

Se $\boldsymbol{\Sigma}$ parece se desviar de uma matriz diagonal, então um MFO pode ser investigado, e o problema inicial será estimar as cargas ℓ_{jk} e as variâncias específicas ψ_j, $j = 1, 2, \ldots, p$ e $k = 1, 2, \ldots, m$.

Consideraremos dois métodos de estimação de parâmetros: o método dos **componentes principais** e o método da **máxima verossimilhança**.

A solução de cada método pode ser rotacionada de forma a facilitar a interpretação dos fatores comuns.

é sempre prudente tentar mais de um método de solução, se o MFO é adequado para o problema em mãos, as soluções deveriam ser consistentes umas com as outras.

Os métodos de estimação e rotação requerem cálculos iterativos. Vários programas estatísticos dispõem de rotinas com esse fim. No **R** está disponível a função **factanal**.

Método das componentes principais

Considere a decomposição espectral de $\boldsymbol{\Sigma}$ com $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$ tal que $\boldsymbol{\Sigma} = \sum_{j=1}^{p} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top$ e observe que podemos escrever $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top$ com

$$\boldsymbol{B} = \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{v}_1 & \sqrt{\lambda_2}\boldsymbol{v}_2 & \cdots & \sqrt{\lambda_p}\boldsymbol{v}_p \end{bmatrix}.$$

Isto ajusta a estrutura de covariâncias prescrita para o MFO quando $m = p$ e $\psi_j = 0\ \forall j$. As cargas da j-ésima coluna são dadas por $\sqrt{\lambda_j}\boldsymbol{v}_j$.

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top + \underbrace{\boldsymbol{\Psi}}_{=\mathbf{0}} = \boldsymbol{B}\boldsymbol{B}^\top.$$

Exceto pelo fator de escala $\sqrt{\lambda_j}$, as cargas do j-ésimo fator para as p variáveis são os coeficientes do j-ésimo componente principal da população.

Apesar da representação aqui considerada ser exata, ela não é particularmente útil, pois usa tantos fatores quantas são as variáveis e não permite qualquer variação nos fatores específicos dados por ε.

É desejável, no entanto, modelos que expliquem a estrutura de covariâncias em função de apenas uns poucos fatores comuns.

Uma abordagem possível é, quando os últimos $p - m$ autovalores são bem pequenos, desprezar a contribuição de $\lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top$, $j = m+1, m+2, \ldots, p$ para $\boldsymbol{\Sigma}$.

Desprezando esta contribuição, obtemos

$$\boldsymbol{\Sigma} \simeq \sum_{j=1}^{m} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top = \underbrace{\boldsymbol{L}}_{p \times m} \boldsymbol{L}^\top \quad \text{com}$$

$$\boldsymbol{L} = \begin{bmatrix} \ell_1 & \ell_2 & \cdots & \ell_m \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{v}_1 & \sqrt{\lambda_2}\boldsymbol{v}_2 & \cdots & \sqrt{\lambda_m}\boldsymbol{v}_m \end{bmatrix}.$$

Esta aproximação assume que os fatores específicos ε são de menor importância e também podem ser ignorados na decomposição da matriz $\boldsymbol{\Sigma}$.

Se os fatores específicos estão incluídos no modelo, suas variâncias podem ser tomadas como os elementos da diagonal da matriz

$$\boldsymbol{\Sigma} - \boldsymbol{L}\boldsymbol{L}^\top.$$

Incluindo os fatores específicos, temos que

$$\begin{aligned}\boldsymbol{\Sigma} &\simeq \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi} \\ &= \begin{bmatrix}\sqrt{\lambda_1}\boldsymbol{v}_1 & \sqrt{\lambda_2}\boldsymbol{v}_2 & \cdots & \sqrt{\lambda_m}\boldsymbol{v}_m\end{bmatrix} \begin{bmatrix}\sqrt{\lambda_1}\boldsymbol{v}_1^\top \\ \sqrt{\lambda_2}\boldsymbol{v}_2^\top \\ \vdots \\ \sqrt{\lambda_m}\boldsymbol{v}_m^\top\end{bmatrix} + \text{diag}(\psi_1, \psi_2, \ldots, \psi_p)\end{aligned}$$

em que $\psi_j = \sigma_{jj} - \sum_{k=1}^{m} \ell_{jk}^2$, para $j = 1, 2, \ldots, p$.

Na aplicação desta abordagem aos dados $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ observados é usual primeiro centrar as observações trabalhando com $\boldsymbol{x}_i - \overline{\boldsymbol{x}}$, $i = 1, 2, \ldots, n$. Observe que a matriz de covariâncias amostral $\boldsymbol{S}$ não se altera com esta transformação de localização.

Nos casos em que as unidades das variáveis são não comensuráveis é desejável trabalhar com as variáveis padronizadas decompondo, então, a matriz de correlações.

A padronização evita o problema de se ter uma variável com variância grande influenciando indevidamente a determinação das cargas sobre os fatores.

O uso da aproximação sugerida por $\boldsymbol{\Sigma} \simeq \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}$ acima é conhecido como solução via componentes principais (CP).

A AF via CP de $\boldsymbol{S}$ é obtida em função da decomposição espectral da matriz de covariâncias amostral $\boldsymbol{S}$. Sejam $(\widehat{\lambda}_j, \widehat{\boldsymbol{v}}_j)$ os pares de autovalores e autovetores de $\boldsymbol{S}$ com $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \cdots \geqslant \widehat{\lambda}_p \geqslant 0$.

Seja $m < p$ o número de fatores comuns. Então,

$$\begin{aligned} \widehat{\boldsymbol{L}} &= \left[\sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{v}}_1 \quad \sqrt{\widehat{\lambda}_2}\widehat{\boldsymbol{v}}_2 \quad \cdots \quad \sqrt{\widehat{\lambda}_m}\widehat{\boldsymbol{v}}_m\right] \quad \text{e} \\ \widehat{\boldsymbol{\Psi}} &= \text{diag}(\widehat{\psi}_1, \widehat{\psi}_2, \ldots, \widehat{\psi}_p), \end{aligned}$$

em que $\widehat{\psi}_j = s_{jj} - \sum_{k=1}^{m} \widehat{\ell}_{jk}^2$, isto é, $\widehat{\boldsymbol{\Psi}}$ matriz diagonal cujos elementos diagonais são os elementos diagonais da matriz $\boldsymbol{S} - \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top$.

Comunalidades: $\widehat{h}_j^2 = \sum_{k=1}^{m} \widehat{\ell}_{jk}^2$.

Observações:

1. A solução da AF via CP com base na matriz de correlações amostral $\boldsymbol{R}$ é obtida da mesma forma. Neste caso, denotaremos os autovalores por $\widehat{\delta}_j$ e os autovetores por $\widehat{\boldsymbol{w}}_j$, para $j = 1, 2, \ldots, p$.
2. Na AF via CP as cargas dos fatores não se alteram se aumentarmos o número de fatores.

Pela definição dos $\widehat{\psi}_j$, os elementos da diagonal de $\boldsymbol{S}$ são iguais aos elementos da diagonal de $\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top + \widehat{\boldsymbol{\Psi}}$. Porém, os elementos fora da diagonal de $\boldsymbol{S}$ não serão, em geral, reproduzidos por $\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top + \widehat{\boldsymbol{\Psi}}$.

Como, então, selecionar m?

Se m não é determinado a priori por considerações tais como teóricas, sua escolha pode ser baseada sobre os autovalores estimados, da mesma forma que em ACP.

Considere a matriz residual dada por $\boldsymbol{S} - \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^{\top} - \widehat{\boldsymbol{\Psi}}$, cujos elementos da diagonal são nulos. Se os outros elementos também forem próximos de zero, podemos, subjetivamente, considerar o MFO a *m* fatores apropriado.

Analiticamente temos:

Soma dos quadrados das entradas de
$\boldsymbol{S} - \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^{\top} - \widehat{\boldsymbol{\Psi}} \leqslant \widehat{\lambda}_{m+1}^2 + \widehat{\lambda}_{m+2}^2 + \cdots + \widehat{\lambda}_p^2$.

Consequentemente, um valor pequeno da soma dos quadrados dos autovalores estimados desprezados implica em um valor pequeno da soma de quadrados dos erros de aproximação.

Idealmente, as contribuições dos primeiros poucos fatores às variâncias amostrais das variáveis deve ser grande.

A contribuição para s_{jj} do primeiro fator é $\widehat{\ell}_{j1}^2$. A contribuição total para a variância amostral total $\text{tr}(\boldsymbol{S}) = \sum_{j=1}^{p} s_{jj}$ do primeiro fator comum é $\sum_{j=1}^{p} \widehat{\ell}_{j1}^2 = \left(\sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{v}}_1\right)^\top \left(\sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{v}}_1\right) = \widehat{\lambda}_1$, pois $\widehat{\boldsymbol{v}}_1$ tem norma 1.

Resumindo, a proporção da variação amostral total devida ao k-ésimo fator ($k = 1, 2, \ldots, m$) é dada por

$$\frac{\widehat{\lambda}_k}{\text{tr}(\boldsymbol{S})} \quad \text{se AF via} \quad \boldsymbol{S}$$

$$\frac{\widehat{\delta}_k}{p} \quad \text{se AF via} \quad \boldsymbol{R}.$$

Este critério costuma ser usado como um esquema heurístico para determinar o número de fatores comuns apropriado.

Note que $\widehat{\lambda}_k = \sqrt{\widehat{\lambda}_k}\boldsymbol{v}_k^\top \boldsymbol{v}_k \sqrt{\widehat{\lambda}_k} = \boldsymbol{\ell}_k^\top \boldsymbol{\ell}_k = \sum_{j=1}^{p} \widehat{\ell}_{jk}^2$ (soma de quadrados da k-ésima coluna de $\boldsymbol{L}$ - correspondente ao k-ésimo fator f_k).

O número de fatores comuns retido no modelo é crescente até que uma proporção adequada da variação amostral total tenha sido explicada.

Outra convenção muito usada nos pacotes estatísticos é fazer m igual ao número de autovalores maiores do que 1, se estamos com os dados padronizados ou igual ao número de autovalores **positivos** ("bem maiores que zero") de $\boldsymbol{S}$, se estamos trabalhando com os dados originais.

Essas "regras do polegar" não devem ser aplicadas indiscriminadamente. Por exemplo, $m = p$, se a regra para $\boldsymbol{S}$ for obedecida, já que esperamos que todos os autovalores de $\boldsymbol{S}$ sejam positivos para tamanhos amostrais grandes.

A melhor abordagem é reter poucos em vez de muitos fatores, assumindo que eles fornecem uma interpretação satisfatória dos dados e produzem um ajuste adequado para $\boldsymbol{S}$ ou $\boldsymbol{R}$.

Exemplo

Em um estudo sobre a preferências de consumo, uma amostra de consumidores foi solicitada a classificar vários atributos de um novo produto. As respostas, em uma escala diferencial semântica de 1 a 7 pontos, foram tabuladas e a matriz de correlações dos atributos foi calculada. Os atributos classificados foram sabor (X_1), preço (X_2), aroma (X_3), "bom para lanche" (X_4) e "fornece muita energia" (X_5).

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0,02 & 0,96 & 0,42 & 0,01 \\ 0,02 & 1 & 0,13 & 0,71 & 0,85 \\ 0,96 & 0,13 & 1 & 0,50 & 0,11 \\ 0,42 & 0,71 & 0,50 & 1 & 0,79 \\ 0,01 & 0,85 & 0,11 & 0,79 & 1 \end{bmatrix}$$

A partir das entradas de $\boldsymbol{R}$, verificamos que as variáveis 1 e 3 (gosto e aroma) e as variáveis 2 e 5 (preço e energia) formam grupos. A variável 4 (bom para lanche) está mais próxima do grupo (2,5) do que grupo (1,3).

Ver arquivo exemplo_21.r

Dados estes resultados e o pequeno número de variáveis, podemos esperar que as relações lineares aparentes entre as variáveis podem ser explicadas em função de, no máximo, 2 ou 3 fatores comuns.

Calculando a decomposição espectral de $\boldsymbol{R}$ obtemos $\widehat{\delta}_1 \simeq 2,853$ e $\widehat{\delta}_2 \simeq 1,806$ tal que ambos explicam cerca 93,2% da variação amostral total. Logo, a escolha $m = 2$ parece apropriada.

Na tabela a seguir são apresentadas as estimativas das cargas dos fatores sobre as cinco variáveis, das comunalidades e das variâncias.

Variável	$\widehat{\ell}_1$	$\widehat{\ell}_2$	$\widehat{h}_j^2$	$\widehat{\psi}_j$
Z_1	-0,560	0,816	0,979	0,021
Z_2	-0,777	-0,524	0,879	0,121
Z_3	-0,645	0,748	0,976	0,024
Z_4	-0,939	-0,105	0,893	0,107
Z_5	-0,798	-0,543	0,932	0,068

Lembre-se que $\widehat{\ell}_k = \sqrt{\widehat{\delta}_k}\boldsymbol{w}_k \Rightarrow \widehat{\delta}_k = \widehat{\ell}_k^{\top}\widehat{\ell}_k$, $k = 1, 2$.

$$\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^{\top} + \widehat{\boldsymbol{\Psi}} = \begin{bmatrix} 1,000 & 0,007 & 0,972 & 0,440 & 0,004 \\ 0,007 & 1,000 & 0,110 & 0,785 & 0,905 \\ 0,972 & 0,110 & 1,000 & 0,528 & 0,109 \\ 0,440 & 0,785 & 0,528 & 1,000 & 0,807 \\ 0,004 & 0,905 & 0,109 & 0,807 & 1,000 \end{bmatrix}$$

reproduz $\boldsymbol{R}$ adequadamente. A matriz residual correspondente a solução $m = 2$ é:

$$\boldsymbol{R} - \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^{\top} - \widehat{\boldsymbol{\Psi}} = \begin{bmatrix} 0,000 & 0,013 & -0,012 & -0,020 & 0,006 \\ 0,013 & 0,000 & 0,020 & -0,075 & -0,055 \\ -0,012 & 0,020 & 0,000 & -0,028 & 0,001 \\ -0,020 & -0,075 & -0,028 & 0,000 & -0,017 \\ 0,006 & -0,055 & 0,001 & -0,017 & 0,000 \end{bmatrix}$$

As comunalidades indicam que os dois fatores levam em conta a maior parte da variação de cada variável.

Interpretação: Uma rotação dos fatores frequentemente revela uma estrutura simples e auxilia na interpretação dos fatores. Voltaremos a este exemplo depois de tratarmos este tema.

Exemplo

Considere novamente os dados sobre preços de $p = 5$ ações durante 103 semanas consecutivas. Realizamos uma ACP sobre estes dados padronizados e verificamos que os dois primeiros componentes correspondiam a cerca de 77% da variação total. Fazendo $m = 1$ e $m = 2$, podemos facilmente obter a solução do MFO via CP.
Especificamente, as cargas dos fatores estimadas são os coeficientes dos componentes principais (autovetores de $\boldsymbol{R}$), corrigidas pela raiz quadrada dos correspondentes autovalores.

A tabela a seguir apresenta as estimativas das cargas dos fatores, das comunalidades, das variâncias específicas e a proporção total da variação explicada por cada fator para as soluções do modelo MFO com $m = 1$ e $m = 2$.

Ver arquivo exemplo_22.r

Variável	$m = 1$		$m = 2$		
	$\widehat{\ell}_1$	$\widehat{\psi}$	$\widehat{\ell}_1$	$\widehat{\ell}_2$	$\widehat{\psi}$
JP	-0,732	0,464	-0,732	0,437	0,273
Citi	-0,831	0,309	-0,831	0,280	0,230
Fargo	-0,726	0,473	-0,726	0,374	0,333
Royal	-0,605	0,634	-0,605	-0,694	0,153
Exxonl	-0,563	0,683	-0,563	-0,719	0,166
Prop. Acum.	0,487		0,487	0,769	

A matriz residual correspondente a solução $m = 2$ é:

$$\boldsymbol{R} - \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^{\top} - \widehat{\boldsymbol{\Psi}} = \begin{bmatrix} 0,000 & -0,099 & -0,185 & -0,025 & 0,056 \\ -0,099 & 0,000 & -0,134 & 0,014 & -0,054 \\ -0,185 & -0,134 & 0,000 & 0,003 & 0,006 \\ -0,025 & 0,014 & 0,003 & 0,000 & -0,156 \\ 0,056 & -0,054 & 0,006 & -0,156 & 0,000 \end{bmatrix}$$

A proporção da variação total explicada por dois fatores é bem maior do que a explicada quando se considera o modelo a um fator. No entanto, o modelo a dois fatores parece produzir correlações amostrais maiores. Este é particularmente o caso de r_{13}.

Parece bastante claro que o primeiro fator representa as condições econômicas gerais e pode ser chamado de fator de mercado. Todas as cargas das ações são altas sobre este fator, e as cargas são praticamente iguais.

O segundo fator contrasta as ações do setor bancário com as ações de petróleo. Assim, este fator parece diferenciar ações em diferentes ramos de atividade e pode ser chamado de fator de atividade.

Resumindo, as taxas de retorno parecem ser determinadas pelas condições gerais de mercado e atividades que são exclusivas para diferentes ramos de atividade, bem como um fator residual específico relacionado a cada empresa.

Método da máxima verossimilhança

$$\boldsymbol{F} \sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{I}_m), \quad \varepsilon \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi}) \quad \text{e} \quad \boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Temos,

$$\begin{aligned}
L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2} \\
&\times \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \left[\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top \right] \right) \right\} \\
&\times \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \left[n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \right] \right) \right\} \\
&= (2\pi)^{-(n-1)p/2} |\boldsymbol{\Sigma}|^{-(n-1)/2} \\
&\times \exp\left\{ -\frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \left[\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top \right] \right) \right\} \\
&\times (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{n}{2} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \right\}.
\end{aligned}$$

que depende de $\boldsymbol{L}$ e $\boldsymbol{\Psi}$ em função de $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}$. Este modelo ainda não está bem definido devido à multiplicidade de escolhas para $\boldsymbol{L}$ que são possíveis pelas transformações ortogonais da solução.

É desejável tornar $\boldsymbol{L}$ bem definida se impondo uma condição de unicidade que seja computacionalmente conveniente:

$$\boldsymbol{L}^{\top}\boldsymbol{\Psi}\boldsymbol{L} = \boldsymbol{\Delta} \quad \text{com} \quad \boldsymbol{\Delta} \quad \text{uma matriz diagonal.}$$

As estimativas de máxima verossimilhança de $\boldsymbol{L}$ e $\boldsymbol{\Psi}$ dadas por $\widetilde{\boldsymbol{L}}$ e $\widetilde{\boldsymbol{\Psi}}$ são obtidas por maximização numérica. Os pacotes estatísticos em geral dispõem de rotinas para calcular tais estimativas.

Proposição

Seja $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ uma amostra aleatória de $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}$ para o MFO com m fatores comuns. Seja $(\widetilde{\boldsymbol{L}}, \widetilde{\boldsymbol{\Psi}})$ a estimativa de máxima verossimilhança de $(\boldsymbol{L}, \boldsymbol{\Psi})$ sob a condição $\boldsymbol{L}^\top \boldsymbol{\Psi} \boldsymbol{L} = \boldsymbol{\Delta}$ matriz diagonal. Então, a estimativa de máxima verossimilhança das comunalidades são dadas por (soma de quadrados da j-ésima linha de $\boldsymbol{L}$)

$$\widetilde{h}_j^2 = \widetilde{\ell}_{j1}^2 + \widetilde{\ell}_{j2}^2 + \cdots + \widetilde{\ell}_{jm}^2 \quad \text{para} \quad j = 1, 2, \ldots, p.$$

Assim, a proporção da variação total devida ao k-ésimo fator é dada por

$$\frac{\widetilde{\lambda}_k}{\text{tr}(\boldsymbol{S})} = \frac{\widetilde{\boldsymbol{\ell}}_k^\top \widetilde{\boldsymbol{\ell}}_k}{\text{tr}(\boldsymbol{S})} = \frac{\sum_{j=1}^{p} \widetilde{\ell}_{jk}^2}{\text{tr}(\boldsymbol{S})}, \quad k = 1, 2, \ldots, m \quad \textbf{(Invariância do EMV)}.$$

Se usarmos as variáveis padronizadas $\boldsymbol{Z} = \boldsymbol{V}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$, temos

$$\begin{aligned}
\boldsymbol{\mathcal{R}} &= \boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2} \\
&= \boldsymbol{V}^{-1/2}(\boldsymbol{L}\boldsymbol{L}^{\top} + \boldsymbol{\Psi})\boldsymbol{V}^{-1/2} \\
&= (\boldsymbol{V}^{-1/2}\boldsymbol{L})(\boldsymbol{V}^{-1/2}\boldsymbol{L})^{\top} + \boldsymbol{V}^{-1/2}\boldsymbol{\Psi}\boldsymbol{V}^{-1/2} \\
&= \boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{L}_{\boldsymbol{Z}}^{\top} + \boldsymbol{\Psi}_{\boldsymbol{Z}}.
\end{aligned}$$

Dado $\widetilde{\boldsymbol{L}}_{\boldsymbol{Z}}$ e $\widetilde{\boldsymbol{\Psi}}_{\boldsymbol{Z}}$ via $\boldsymbol{R}$, obtemos

$$\begin{aligned}
\widetilde{\boldsymbol{L}} &= \widetilde{\boldsymbol{V}}^{1/2}\widetilde{\boldsymbol{L}}_{\boldsymbol{Z}} \\
\widetilde{\boldsymbol{\Psi}} &= \widetilde{\boldsymbol{V}}^{1/2}\widetilde{\boldsymbol{\Psi}}_{\boldsymbol{Z}}\widetilde{\boldsymbol{V}}^{1/2} \\
\widetilde{\ell}_{jk} &= \widetilde{\ell}_{\boldsymbol{Z},jk}\sqrt{\widetilde{\sigma}_{jk}} \\
\widetilde{\psi}_{j} &= \widetilde{\psi}_{\boldsymbol{Z},j}\widetilde{\sigma}_{jj},
\end{aligned}$$

com $\widetilde{\sigma}_{jj}$ o EMV de σ_{jj}, $j = 1, 2, \ldots, p$ e $k = 1, 2, \ldots, m$.

Logo, na análise fatorial via método da máxima verossimilhança há equivalência entre decompor $\boldsymbol{S}$ ou $\boldsymbol{R}$. O mesmo não se aplica a análise fatorial via componentes principais.

Exemplo

Para os dados do mercado de ações, obteremos a solução do MFO via máxima verossimilhança.

Variável	$m = 1$		$m = 2$		
	$\widetilde{\ell}_1$	$\widetilde{\psi}$	$\widetilde{\ell}_1$	$\widetilde{\ell}_2$	$\widetilde{\psi}$
JP	0,715	0,488	0,121	0,754	0,417
Citi	0,875	0,234	0,328	0,786	0,275
Fargo	0,663	0,560	0,188	0,650	0,542
Royal	0,346	0,880	0,997	-0,007	0,005
Exxonl	0,279	0,922	0,685	0,026	0,530
Prop. Acum.	0,383		0,324	0,646	

Ver arquivo exemplo_23.r

A matriz residual correspondente a solução $m = 2$ é:

$$\boldsymbol{R} - \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^{\top} - \widetilde{\boldsymbol{\Psi}} =$$

$$\begin{bmatrix}
0{,}00000011 & 0{,}00000750 & -0{,}00256422 & -0{,}00033256 & 0{,}05198222 \\
0{,}00000750 & 0{,}00000003 & 0{,}00160887 & 0{,}00021162 & -0{,}03307885 \\
-0{,}00256422 & 0{,}00160887 & 0{,}00000005 & -0{,}00000952 & 0{,}00055472 \\
-0{,}00033256 & 0{,}00021162 & -0{,}00000952 & -0{,}00000156 & 0{,}00012189 \\
0{,}05198222 & -0{,}03307885 & 0{,}00055472 & 0{,}00012189 & 0{,}00000027
\end{bmatrix}$$

Os elementos da matriz residual via máxima verossimilhança são bem menores do que os obtidos via método de componentes principais.

Portanto, a abordagem da máxima verossimilhança é mais apropriada.

A porcentagem da variação acumulada explicada pelos fatores é maior para os fatores via componentes principais do que para os fatores via máxima verossimilhança. Isto não é uma surpresa, pois o método componentes principais justamente tem como critério escolher o fator com maior proporção da variação total. As cargas obtidas via componentes principais estão relacionadas aos componentes principais que têm, por planejamento, uma propriedade de otimização da variância.

Verificando a solução via máxima verossimilhança, observamos que todas as cargas do fator 1 são positivas. Chamamos este fator de fator de mercado. A interpretação do segundo fator não é tão clara quanto a interpretação sugerida na solução via componentes principais. As ações dos bancos têm cargas grandes e positivas, enquanto que as ações de petróleo têm cargas desprezíveis.

A partir desta perspectiva, o segundo fator diferencia as ações de bancos das ações de petróleo e pode ser chamado fator de ramo de atividade. Alternativamente, o segundo fator pode ser simplesmente chamado “fator de banco”.

Os padrões das cargas iniciais dos fatores para a solução de máxima verossimilhança estão restritos pela condição de unicidade ($\widetilde{\boldsymbol{L}}^{\top}\widetilde{\boldsymbol{\Psi}}\widetilde{\boldsymbol{L}}$ é diagonal). Portanto, padrões úteis não serão revelados, em geral, até que os fatores sejam rotacionados.

Teste para o número de fatores comuns para grandes amostras

A suposição de normalidade leva diretamente a um teste da adequação do modelo a m fatores.

Suponha que m fatores comuns foram considerados no ajuste.

Neste caso $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}$. Assim, testar a adequação de m fatores é equivalente a testar

$$\begin{cases} H_0: & \boldsymbol{\Sigma} = \underbrace{\boldsymbol{L}}_{p\times m}\boldsymbol{L}^\top + \underbrace{\boldsymbol{\Psi}}_{\text{matriz diagonal}} \\ H_1: & \boldsymbol{\Sigma} \quad \text{é qualquer outra matriz positiva definida.} \end{cases}$$

Quando $\boldsymbol{\Sigma}$ não tem forma especial, o valor máximo da função de verossimilhança é atingido quando

$$\widetilde{\boldsymbol{\mu}} = \overline{\boldsymbol{x}} \quad \text{e} \quad \widetilde{\boldsymbol{\Sigma}} = \boldsymbol{S}_n = \frac{n-1}{n}\boldsymbol{S},$$

em que

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{S}_n|^{-n/2} \exp\left\{-\frac{np}{2}\right\}.$$

Sob H_0, o valor máximo da função de verossimilhança é atingido quando

$$\widetilde{\boldsymbol{\mu}} = \overline{\boldsymbol{x}} \quad \text{e} \quad \widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^\top + \widetilde{\boldsymbol{\Psi}},$$

em que

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = L(\boldsymbol{\mu}, \boldsymbol{L}, \boldsymbol{\Psi}) \propto |\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^\top + \widetilde{\boldsymbol{\Psi}}|^{-n/2} \exp\left\{-\frac{n}{2}\text{tr}\left(\left[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^\top + \widetilde{\boldsymbol{\Psi}}\right]^{-1}\boldsymbol{S}_n\right)\right\}.$$

Segue, após diversas simplificações que

$$-2\ln\boldsymbol{\Lambda} = n\ln\left(\frac{|\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^\top + \widetilde{\boldsymbol{\Psi}}|}{|\boldsymbol{S}_n|}\right) \quad \text{com} \quad \frac{(p-m)^2 - p - m}{2},$$

graus de liberdade.

Note que sob H_0,

$$\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{S}_n = \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^\top + \widetilde{\boldsymbol{\Psi}} \Rightarrow \text{tr}\left(\left[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^\top + \widetilde{\boldsymbol{\Psi}}\right]^{-1}\boldsymbol{S}_n\right) = \text{tr}\,(\boldsymbol{I}_p) = p.$$

Bartlett sugeriu a seguinte correção

$$(n-1-(2p+4m+5)/6)\ln\left(\frac{|\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^\top+\widetilde{\boldsymbol{\Psi}}|}{|\boldsymbol{S}_n|}\right)\overset{a}{\sim}\chi^2_\nu,$$

em que $\nu=\dfrac{(p-m)^2-p-m}{2}$ para n e $n-p$ grandes.

Como $\nu>0$, segue que $m<\dfrac{1}{2}(2p+1-\sqrt{8p+1})$ de modo a poder aplicar este teste.

Observe que na implementação deste teste para verificar se m é adequado para o númerro de fatores comuns, estamos comparando as variâncias generalizadas $|\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^\top+\widetilde{\boldsymbol{\Psi}}|$ e $|\boldsymbol{S}_n|$. Se n é grande e m é relativamente pequeno comparado a p, H_0 será comumente rejeitada, levando à retenção de mais fatores comuns. Porém, $\widetilde{\boldsymbol{\Sigma}}=\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^\top+\widetilde{\boldsymbol{\Psi}}$ pode ser bem parecida com $\boldsymbol{S}_n$ tal que adicionar mais fatores não fornecerá ganhos, mesmo que os novos fatores sejam “significantes”. Algum julgamento deve ser exercido na escolha de m (por exemplo a proporção acumulada da variabilidade total).

Exemplo (continuação)

Para os dados do mercado de ações, faremos o teste de adequação de $m = 2$ da solução do MFO via máxima verossimilhança.

Primeiro, note para $n = 103$ e $p = 5$ devemos ter m menor que

$$\frac{1}{2}(2p + 1 - \sqrt{8p + 1}) = 2,298.$$

Logo, podemos testar a adequação para $m = 1$ e $m = 2$.

Para $m = 1$, a estatística de qui-quadrado é 62,22 com 5 grau de liberdade. O valor p é 10^{-12}. Logo, rejeitamos H_0.

Para $m = 2$, a estatística de qui-quadrado é 2,00 com 1 grau de liberdade. O valor p é 0,16. Logo, não rejeitamos H_0.

Nota: os resultados apresentados acima diferem pouca coisa dos resultados obtidos via função factanal do R. Isto se deve aos cálculos numéricos e suas possíveis aproximações.

Rotação de fatores

Vimos que todas as cargas são obtidas das cargas iniciais por uma transformação ortogonal com a mesma habilidade em reproduzir a matriz de covariâncias (correlações).

Da álgebra matricial, sabemos que uma transformação ortogonal corresponde a uma rotação (ou reflexão) dos eixos coordenados em relação à origem. Por esta razão, a técnica de realizar uma transformação ortogonal das cargas para obter as cargas rotacionadas é chamada **rotação de fatores**.

Se $\widetilde{\boldsymbol{L}}_{p\times m}$ é a matriz das cargas estimadas, então $\widetilde{\boldsymbol{L}}^{\star} = \widetilde{\boldsymbol{L}}\boldsymbol{Q}$, com $\boldsymbol{Q}\boldsymbol{Q}^{\top} = \boldsymbol{Q}^{\top}\boldsymbol{Q} = \boldsymbol{I}_m$, é uma matriz de cargas rotacionadas.

Além disso, a matriz de covariâncias (correlações) estimada se mantém inalterada, pois

$$\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^{\top} + \widetilde{\boldsymbol{\Psi}} = \widetilde{\boldsymbol{L}}\boldsymbol{Q}\boldsymbol{Q}^{\top}\widetilde{\boldsymbol{L}}^{\top} + \widetilde{\boldsymbol{\Psi}} = \widetilde{\boldsymbol{L}}^{\star}\widetilde{\boldsymbol{L}}^{\star\top} + \widetilde{\boldsymbol{\Psi}}.$$

A matriz residual também não se altera, pois

$$\boldsymbol{S}_n - \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^\top - \widetilde{\boldsymbol{\Psi}} = \boldsymbol{S}_n - \widetilde{\boldsymbol{L}}^\star\widetilde{\boldsymbol{L}}^{\star\top} - \widetilde{\boldsymbol{\Psi}}$$

e as estimativas das comunalidades e variâncias específicas também não se alteram.

Do ponto de vista matemático, tanto faz usar $\widetilde{\boldsymbol{L}}$ como usar $\widetilde{\boldsymbol{L}}^\star$.

Como as cargas das variáveis sobre os fatores comuns podem não ser facilmente interpretáveis, é usual rotacioná-las até que uma “estrutura mais simples” seja alcançada.

O ideal seria poder obter um padrão de cargas tais que a carga de cada variável seja bem grande em um único fator e tenha valores pequenos a moderados nos demais fatores.

Nem sempre será possível obter tal “estrutura simples”.

Concentraremo-nos em métodos gráficos e analíticos para determinar a rotação ortogonal visando uma estrutura mais simples das cargas.

Quando $m = 2$ ou os fatores comuns são considerados 2 de cada vez, uma transformação para uma estrutura simples pode ser determinada graficamente.

Fatores comuns não correlacionados são olhados como vetores unitários no sistema de eixos coordenados ortogonais representando as variáveis originais.

Um gráfico de dispersão dos pares de cargas dos fatores produz p pontos, $(\widetilde{\ell}_{j1}, \widetilde{\ell}_{j2})$, $j = 1, 2, \ldots, p$, corrrespondentes a cada variável.

Os eixos coordenados podem ser visualmente rotacionados por um ângulo ϕ e as novas cargas rotacionadas $\widetilde{\ell}^{\star}_{jq}$, $j = 1, 2, \ldots, p$, $q = 1, 2$, são determinadas pelas relações ($\widetilde{\boldsymbol{L}}^{\star} = \widetilde{\boldsymbol{L}}\boldsymbol{Q}$)

$$\boldsymbol{Q} = \begin{bmatrix} \cos\phi & \text{sen}\phi \\ -\text{sen}\phi & \cos\phi \end{bmatrix} \quad \text{sentido horário}$$

$$\boldsymbol{Q} = \begin{bmatrix} \cos\phi & -\text{sen}\phi \\ \text{sen}\phi & \cos\phi \end{bmatrix} \quad \text{sentido anti-horário.}$$

Estas relações raramente são implementadas nas análises gráficas visuais a duas dimensões.

Nesta situação, conglomerados de variáveis são frequentemente percebidos no diagrama de dispersão das cargas e estes conglomerados nos permitem identificar os fatores comuns sem precisar calcular o ângulo ϕ da rotação.

Por outro lado, para $m > 2$, orientações não são facilmente visualizadas, e as magnitudes das cargas rotacionadas devem ser investigadas para encontrar uma interpretação com algum significado dos dados originais. A escolha de uma matriz ortogonal $\boldsymbol{Q}$ que satisfaz uma certa medida analítica de estrutura simples será brevemente considerada.

O critério **varimax** é uma medida analítica para a rotação de fatores.

Defina $\breve{\ell}^{\star}_{jk} = \dfrac{\widetilde{\ell}^{\star}_{jk}}{\widetilde{h}_j}$ como os **coeficientes rotacionados** corrigidos pelo inverso da raiz quadrada das comunalidades estimadas.

O procedimento varimax, também conhecido como **varimax normal**, seleciona a transformação ortogonal $\boldsymbol{Q}$ que maximiza

$$\eta(\breve{\boldsymbol{L}}^{\star}) = \frac{1}{p}\sum_{k=1}^{m}\left[\sum_{j=1}^{p}\breve{\ell}^{\star 4}_{jk} - \left(\sum_{j=1}^{p}\breve{\ell}^{\star 2}_{jk}\right)^2/p\right] \text{ com } \breve{\boldsymbol{L}}^{\star} = \left[\breve{\boldsymbol{\ell}}^{\star}_1 \quad \breve{\boldsymbol{\ell}}^{\star}_2 \quad \cdots \breve{\boldsymbol{\ell}}^{\star}_m\right].$$

Os coeficientes rotacionados $\breve{\ell}^{\star}_{jk}$ têm o efeito de dar as variáveis com comunalidades relativamente pequenas mais peso na determinação da estrutura simples.

Depois de determinar $\boldsymbol{Q}$, as cargas $\breve{\ell}^{\star}_{jk}$ são multiplicadas por $\widetilde{h}_j$ tal que as comunalidades estimadas sejam preservadas, isto é, $\widetilde{\ell}^{\star}_{jk} = \breve{\ell}^{\star}_{jk}\widetilde{h}_j$.

Uma interpretação simples para $\eta(\breve{\boldsymbol{L}}^{\star})$ é que ela é proporcional a soma das variâncias dos quadrados das cargas corrigidas pelo inverso da raiz quadrada das comunalidades para o k-ésimo fator.

Efetivamente, maximizar $\eta(\check{\boldsymbol{L}}^{\star})$ corresponde a espalhar os quadrados das cargas sobre cada fator tanto quanto for possível. Portanto, espera-se encontrar grupos de variáveis com cargas grandes e grupos com cargas pequenas para cada fator após a rotação.

Algoritmos computacionais para a maximização de $\eta(\check{\boldsymbol{L}}^{\star})$ costumam estar disponíveis. No **R** podemos utilizar a função ou o argumento **varimax**.

Exemplo (continuação)

Rotação das cargas para os dados sobre preferência dos consumidores.

As cargas originais foram obtidas pelo método de componentes principais.

Variável	$\widehat{\ell}_1$	$\widehat{\ell}_2$
Z_1	-0,560	0,816
Z_2	-0,777	-0,524
Z_3	-0,645	0,748
Z_4	-0,939	-0,105
Z_5	-0,798	-0,543

Ver arquivo exemplo_24.r

Após inspeção visual é fácil identificar novos eixos ortogonais que produzem uma estrutura simples.

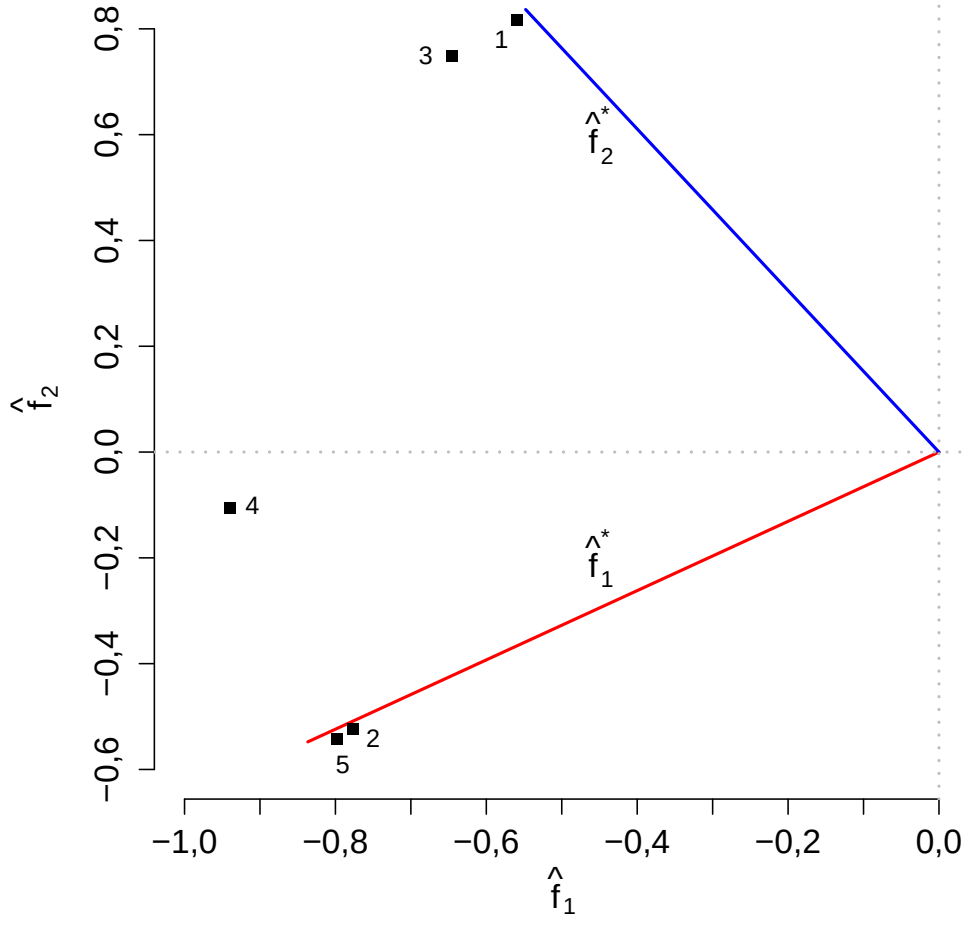

Figura 10: Rotação de fatores via varimax.

Variável	$\widehat{\ell}_1$	$\widehat{\ell}_2$	$\widehat{\ell}_1^{\star}$	$\widehat{\ell}_2^{\star}$
Z_1	-0,560	0,816	-0,021	0,989
Z_2	-0,777	-0,524	-0,937	-0,013
Z_3	-0,645	0,748	-0,130	0,979
Z_4	-0,939	-0,105	-0,843	0,427
Z_5	-0,798	-0,543	-0,965	-0,017

É claro que as variáveis 2, 4 e 5 definem o fator 1 (cargas altas neste fator e desprezíveis no fator 2), enquanto que as variáveis 1 e 3 definem o fator 2. Podemos chamar o fator 1 de fator **nutricional** e o fator 2 de fator de **paladar**.

A matriz de rotação varimax é

$$\boldsymbol{Q} = \begin{bmatrix} 0,837 & -0,548 \\ 0,548 & 0,837 \end{bmatrix}.$$

Exemplo (continuação)

Rotação varimax - mercado de ações - para obter as cargas rotacionadas dos fatores obtidos via método de máxima verossimilhança.

	$\widetilde{\ell}_1$	$\widetilde{\ell}_2$	$\widetilde{\ell}_1^{\star}$	$\widetilde{\ell}_2^{\star}$
JP	0,121	0,754	0,763	0,029
Citi	0,328	0,786	0,819	0,232
Fargo	0,188	0,650	0,668	0,108
Royal	0,997	-0,007	0,113	0,991
Exxon	0,685	0,026	0,108	0,677

As cargas rotacionadas indicam que as ações de bancos apresentam cargas altas no fator 1, enquanto que as ações de petróleo apresentam cargas altas no fator 2.

Os dois fatores rotacionados diferenciam os ramos de atividade. Fica difícil rotular estes fatores de forma inteligente. O fator 1 representa aquelas forças econômicas únicas que acarretam na movimento comum das ações dos bancos. O fator 2 parece representar condições econômicas que afetam ações de petróleo.

Ver arquivo exemplo_25.r

Como já foi observado, um fator geral (isto é, um fator para o qual todas as variáveis apresentam cargas altas) tende a ser “destruído” após a rotação. Por esta razão, em casos nos quais um fator geral é evidente, uma rotação ortogonal é algumas vezes realizada, mantendo o fator geral fixado.

Existem outros métodos de rotação ortogonal.

- Método **quartimax**: minimiza o número de fatores necessários para explicar uma variável.
- Método **equimax**: é um compromisso entre os métodos varimax e quartimax.

O método varimax é o mais popular.

Rotações oblíquas

Rotações ortogonais são apropriadas para um modelo fatorial no qual os fatores comuns são supostos ser independentes. Muitas investigações sociais também consideram rotações oblíquas dos fatores.

Se olhamos os m fatores comuns como eixos coordenados, o ponto com as m coordenadas $(\widetilde{\ell}_{j1}, \widetilde{\ell}_{j2}, \ldots, \widetilde{\ell}_{jm})$ representa a posição da j-ésima variável no espaço dos fatores. Supondo que as variáveis estão agrupadas em conclomerados disjuntos, uma rotação ortogonal para uma estrutura simples corresponde à uma rotação rígida dos eixos coordenados tal que os eixos, após a rotação, passam tão perto quanto possível destes conglomerados.

Uma rotação oblíqua para uma estrutura simples corresponde a uma rotação não rígida do sistema de coordenadas tal que os eixos rotacionados (não mais ortogonais) passam (quase) pelos conglomerados. Uma rotação oblíqua busca expressar cada variável em função de um número mínimo de fatores, preferencialmente um único fator.

Para rotações oblíquas, o método mais popular é o **promax** que tem a vantagem de ser rápido e conceitualmente simples. O método busca ajustar uma matriz alvo que tenha uma estrutura simples.

Escores dos fatores

Na análise fatorial, o interesse está comumente centrado nos parâmetros do modelo fatorial. Porém, os valores estimados dos fatores comuns, chamados **escores dos fatores**, podem também ser necessários para as análises.

Estas quantidades são em geral usadas para propósitos de diagnósticos, bem como entradas para análises subsequentes. Os escores dos fatores não são estimativas de parâmetros desconhecidos no sentido usual da estimação.

Ao contrário, eles são "estimativas" dos valores para os vetores fatoriais aleatórios não observáveis $\boldsymbol{F}_i$, $i = 1, 2, \ldots, m$. Isto é, os escores $\widetilde{\boldsymbol{F}}_i$ são as estimativas dos valores de $\boldsymbol{f}_i$ atingidos por $\boldsymbol{F}_i$, $i = 1, 2, \ldots, n$.

Este problema de estimação é complicado, pois as quantidades não observáveis $\boldsymbol{F}_i$ e ε_i superam em número as observações $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$.

Para lidar esta dificuldade, algumas abordagens, um tanto heurísticas, mas fundamentadas, para o problema de estimação dos valores dos fatores foram desenvolvidas. Apresentaremos duas delas. Ambas têm dois elementos em comum.

(1) tratam as estimativas das cargas dos fatores $\widetilde{\ell}_{jk}$ e das variâncias específicas $\widetilde{\psi}_j$ como se fossem os valores verdadeiros.

(2) envolvem transformações lineares dos dados originais, centradas ou padronizadas. Comumente, as cargas rotacionadas estimadas são usadas para calcular os escores dos fatores. As fórmulas computacionais são substituídas por cargas não rotacionadas e, portanto, não iremos diferenciar entre elas.

Método dos mínimos quadrados ponderados

$$\text{Modelo:} \quad \boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{LF} + \boldsymbol{\varepsilon}.$$

A soma de quadrados dos erros ponderados pela recíproca de suas variâncias é dada por

$$S^2(\boldsymbol{F}) = \sum_{j=1}^{p} \frac{\varepsilon_j^2}{\psi_j} = \boldsymbol{\varepsilon}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\varepsilon} = (\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{LF}) \boldsymbol{\Psi}^{-1} (\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{LF})$$

Bartlett propôs escolher as estimativas $\widetilde{\boldsymbol{F}}$ de $\boldsymbol{F}$ que minimizam $S^2(\boldsymbol{F})$.

Escores dos fatores obtidos por MQP a partir das estimativas MV:

$$\widetilde{\boldsymbol{F}}_i = (\widetilde{\boldsymbol{L}}^\top \widetilde{\boldsymbol{\Psi}}^{-1} \widetilde{\boldsymbol{L}})^{-1} \widetilde{\boldsymbol{L}}^\top \widetilde{\boldsymbol{\Psi}}^{-1} (\boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}) = \widetilde{\boldsymbol{\Delta}}^{-1} \widetilde{\boldsymbol{L}}^\top \widetilde{\boldsymbol{\Psi}}^{-1} (\boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}),$$

$i = 1, 2, \ldots, n$ ou, se a matriz de correlação é decomposta,

$$\widetilde{\boldsymbol{F}}_i = (\widetilde{\boldsymbol{L}}_{\boldsymbol{Z}}^\top \widetilde{\boldsymbol{\Psi}}_{\boldsymbol{Z}}^{-1} \widetilde{\boldsymbol{L}}_{\boldsymbol{Z}})^{-1} \widetilde{\boldsymbol{L}}_{\boldsymbol{Z}}^\top \widetilde{\boldsymbol{\Psi}}_{\boldsymbol{Z}}^{-1} \boldsymbol{z}_i = \widetilde{\boldsymbol{\Delta}}_{\boldsymbol{Z}}^{-1} \widetilde{\boldsymbol{L}}_{\boldsymbol{Z}}^\top \widetilde{\boldsymbol{\Psi}}_{\boldsymbol{Z}}^{-1} \boldsymbol{z}_i$$

$i = 1, 2, \ldots, n$ com $\boldsymbol{z}_i = \boldsymbol{V}^{-1/2}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$ e $\widetilde{\mathcal{R}} = \widetilde{\boldsymbol{L}}_{\boldsymbol{Z}} \widetilde{\boldsymbol{L}}_{\boldsymbol{Z}}^\top + \widetilde{\boldsymbol{\Psi}}_{\boldsymbol{Z}}$.

Observações:

1. Se cargas rotacionadas $\widetilde{\boldsymbol{L}}^{\star} = \widetilde{\boldsymbol{L}}\boldsymbol{Q}$ são usadas no lugar das cargas originais, os escores dos fatores correspondentes $\widetilde{\boldsymbol{F}}_i^{\star}$ serão dados por $\widetilde{\boldsymbol{F}}_i^{\star} = \boldsymbol{Q}^{\top}\widetilde{\boldsymbol{F}}_i$, $k = 1, 2, ..., n$.
2. Se as cargas dos fatores são estimadas via método dos componentes principais, costuma-se gerar os escores dos fatores se usando um procedimento de mínimos quadrados ordinários (não ponderados). Implicitamente, isto equivale a assumir que as variâncias específicas são iguais ou aproximadamente iguais. Os escores dos fatores, são, então,

$$\begin{aligned} \widetilde{\boldsymbol{F}}_i &= (\widehat{\boldsymbol{L}}^{\top}\widehat{\boldsymbol{L}})^{-1}\widehat{\boldsymbol{L}}^{\top}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) \quad \text{para} \quad \boldsymbol{\Sigma}, \quad \text{e} \\ \widetilde{\boldsymbol{F}}_i &= (\widehat{\boldsymbol{L}}_{\boldsymbol{Z}}^{\top}\widehat{\boldsymbol{L}}_{\boldsymbol{Z}})^{-1}\widehat{\boldsymbol{L}}_{\boldsymbol{Z}}^{\top}\boldsymbol{z}_i, \qquad \text{para} \quad \boldsymbol{\mathcal{R}}. \end{aligned}$$

Métodos de regressão

Suponha

$$(\boldsymbol{X} - \boldsymbol{\mu}|\boldsymbol{F}) \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}) \quad \text{e} \quad \boldsymbol{F} \sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{I}_m) \quad \text{tal que}$$

$$\begin{pmatrix} \boldsymbol{X} - \boldsymbol{\mu} \\ \boldsymbol{F} \end{pmatrix} \sim \mathcal{N}_{p+m}(\boldsymbol{0}, \boldsymbol{\Sigma}^\star) \quad \text{com}$$

$$\boldsymbol{\Sigma}^\star = \begin{bmatrix} \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi} & \boldsymbol{L} \\ \boldsymbol{L}^\top & \boldsymbol{I}_m \end{bmatrix}.$$

Portanto,

$$\begin{aligned} \mathrm{E}(\boldsymbol{F}|\boldsymbol{x}) &= \boldsymbol{L}^\top(\boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \\ \mathrm{Var}(\boldsymbol{F}|\boldsymbol{x}) &= \boldsymbol{I}_m - \boldsymbol{L}^\top(\boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi})^{-1}\boldsymbol{L} \end{aligned}$$

As quantidades em $\boldsymbol{L}^\top(\boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi})^{-1}$ correspondem aos coeficientes da regressão multivariada dos fatores sobre as variáveis. Estimativas destes coeficientes produzem os escores.

Escores via método de regressão:

$$\begin{aligned} \widetilde{\boldsymbol{F}}_i &= \widetilde{\boldsymbol{L}}^\top \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}), \quad i = 1, 2, \ldots, n \quad \text{ou} \\ \widetilde{\boldsymbol{F}}_i &= \widetilde{\boldsymbol{L}}_{\boldsymbol{Z}}^\top \boldsymbol{R}^{-1}\boldsymbol{z}_i, \quad i = 1, 2, \ldots, n. \end{aligned}$$

Uma medida de concordância dos escores gerados pelos dois métodos é fornecida pelo coeficiente de correlação amostral entre os escores de um mesmo fator.

Não há nenhuma indicação de que um dos métodos seja superior ao outro.

Na função **factanal** do **R** está disponível o argumento `scores=c("none","regression","Bartlett")`.

Exemplo (continuação)

Estimar os escores dos dois fatores - mercado de ações - pelos métodos de Bartlett e da regressão, usando o modelo ajustado por (máxima verossimilhança) para os dados sobre ações e calcule as correlações entre os escores de um mesmo fator.

A correlação estimada entre os fatores pelos dois métodos é aproximadamente

Regressão	Bartlet	
	$\widetilde{F}_1$	$\widetilde{F}_2$
$\widetilde{F}_1$	1	
$\widetilde{F}_2$	0	1

Ver arquivo exemplo_26.r

Perspectivas e estratégias para a análise fatorial

Há muitas decisões que devem ser tomadas em qualquer análise fatorial.

A mais importante é a escolha de m, o número de fatores comuns.

Apesar de um teste para a adequação do modelo para um dado m estar disponível (sob grandes amostras), ele é adequado somente para dados aproximadamente normais. Além disso, o teste provavelmente irá rejeitar o modelo para m pequeno, se o número de variáveis e observações for grande. No entanto, esta é a situação em análise fatorial que fornecerá uma aproximação útil.

Na maioria das vezes, a escolha de m é baseada em uma combinação de:

1. proporção da variância amostral explicada;
2. conhecimento da situação específica em questão; e
3. “razoabilidade” dos resultados (por exemplo, interpretação dos fatores).

A escolha do método de solução e tipo de rotação é uma decisão menos crucial. De fato, as análises fatoriais mais satisfatórias são aquelas nas quais as rotações são experimentadas com mais do que um método e todos os resultados substancialmente confirmam a mesma estrutura de fatores.

Não existe uma estratégia única para a análise fatorial.

Sugere-se a seguinte rotina.

1. Realize uma ACP. Este método é particularmente apropriado para um primeiro olhar sobre os dados. (Não é necessário que $\boldsymbol{S}$ ou $\boldsymbol{R}$ sejam não singulares.)
   - 1.1 Procure por observações suspeitas construindo os gráficos de dispersão dos escores dos fatores. Também, calcule escores padronizados para cada observação e distâncias quadradas como foi descrito na seção de verificação de normalidade.
   - 1.2 Tente uma rotação varimax.
2. Realize uma análise fatorial via máxima verossimilhança, incluindo uma rotação varimax.

3. Compare as soluções obtidas das duas análises fatoriais.
   - 3.1 Os grupos de cargas são parecidos?
   - 3.2 Construa o gráfico de dispersão dos escores do fatores obtidos por componentes principais versus os obtidos por máxima verossimilhança.
4. Repita os três primeiros passos para outros números m de fatores comuns. Fatores extras necessariamente contribuem na compreensão e interpretação dos dados?
5. Para grandes conjuntos de dados, divida-os ao meio e realize uma análise fatorial em cada metade. Compare os dois resultados entre si e com o obtido a partir do conjunto completo de dados para verificar a estabilidade da solução. (Os dados podem ser divididos se colocando a primeira metade dos casos em um grupo e, a segunda, em outro. Isto poderá revelar mudanças ao longo do tempo, se os dados forem temporais.)

A análise fatorial tem um apelo intuitivo forte para as ciências sociais e comportamentais. Nestas áreas, é natural olhar dados multivariados sobre processos animais, sociais, como manifestações de “traços” não observáveis.

A análise fatorial fornece um modo de explicar a variabilidade observada no comportamento em função destes traços.

Ainda, quando tudo é dito e feito, a análise fatorial permanece muito subjetiva. Os exemplos trabalhados, em comum com a maior parte das publicações, consistem de situações nas quais o **modelo fatorial ortogonal** fornece explicações razoáveis em função de uns poucos fatores interpretáveis.

Na prática, a grande maioria das análises fatoriais não produzem tais resultados claros.

O critério para julgar a qualidade de qualquer análise fatorial não foi bem quantificado.

Exemplo (continuação)

Dividir o conjunto de dados - mercado de ações - em dois subconjuntos, o primeiro contendo as 52 primeiras linhas e, o segundo, contendo as úlimas 51 linhas. Rodar o modelo fatorial a dois fatores usando máxima verossimilhança e rotação varimax. Verificar a estabilidade da solução encontrada.

	Completo		Parte 1		Parte 2	
	$\widetilde{\ell}_1$	$\widetilde{\ell}_2$	$\widetilde{\ell}_1$	$\widetilde{\ell}_2$	$\widetilde{\ell}_1$	$\widetilde{\ell}_2$
JP	0,763	0,029	0,799	-0,006	0,741	0,061
Citi	0,819	0,232	0,813	0,258	0,805	0,231
Fargo	0,668	0,108	0,469	0,183	0,918	0,047
Royal	0,113	0,991	0,119	0,990	0,110	0,991
Exxon	0,108	0,677	0,180	0,687	0,081	0,684

Ver arquivo exemplo_27.r

Discriminação e classificação

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 como:

Discriminação: Descrever gráfica 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

Suponha que se dispõe 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 $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\top$. 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 função (de densidade) de probabilidade conjunta de $\boldsymbol{X}$ é dada por $f_1(\boldsymbol{x})$, caso contrário, função (de densidade) de probabilidade 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 em uma 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:

$$\Pr(\text{classificar incorretamente em } \pi_1) = \Pr(\boldsymbol{X} \in R_1|\pi_2)\Pr(\pi_2) = \Pr(1|2)p_2.$$

$$\Pr(\text{classificar incorretamente em } \pi_2) = \Pr(\boldsymbol{X} \in R_2|\pi_1)\Pr(\pi_1) = \Pr(2|1)p_1.$$

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

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:

(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 se reduz 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)**.

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

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

A regra do CECI mínimo é dada por

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

R_2: Caso contrário.

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

$$R_1\colon \left\{ \boldsymbol{x} \middle| (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1} \boldsymbol{x} - \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_c^{-1}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \geqslant \ln\left(\frac{C(1|2)p_2}{C(2|1)p_1}\right) \right\};$$

R_2: Caso contrário.

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 se comportar 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 se comparando os valores da estatística com o valor de

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

Exemplo

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. Além das três medidas, o biólogo identificou o gênero de cada lagarto *m*-macho, *f*-fêmea. O ojetivo é construir uma regra de classificação de gênero a partir das três medidas usando os dados disponíveis.

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

Tabela 17: Probabilidade a priori 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 18: Grupo de médias.

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

Ver arquivo exemplo_28.r

Tabela 19: Coeficientes de discriminação linear.

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

Tabela 20: 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

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

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

usando um argumento completamente diferente. A ideia de Fisher foi transformar as observações multivariadas $\boldsymbol{x}$ em observações univariadas y tal que os y 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, 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

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 se considerando 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.

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right)^{1/2} \exp\left\{-\frac{1}{2}\text{SQ}\right\},$$

em que

$$\text{SQ} = (\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2).$$

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

$$R_1: \; -\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} - k \geqslant \ln\left(\frac{C(1|2)}{C(2|1)}\frac{p_2}{p_1}\right);$$

R_2: caso contrário.

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

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 se substituindo 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

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

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 ou 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

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,

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

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

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

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

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

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 se calculandoe 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 se considerando 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 proporção 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, $k = 1, 2, \ldots, g$.
- $p_1, p_2, \ldots, p_g$ - probabilidades 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

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

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

- 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,

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

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

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

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

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

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

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

- Aloque $\boldsymbol{x}_0$ a π_k se

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

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

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

O escore quadrático $d_k^Q(\underline{x})$ é composto pelas contribuições da variância generalizada $|\boldsymbol{\Sigma}_k|$, 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:

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

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 $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_s$, $s \leq \min\{g-1, p\}$ 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 se criando 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 se comportar tanto pobremente como satisfatoriamente, dependendo das correlações entre as variáveis contínuas e qualitativas. “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. 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 se comportam tão bem quanto à regressão logística e as funções discriminantes discutidas aqui.

Seleção de variáveis

Em algumas aplicações da análise discriminante, os dados estão disponíveis para um grande número de variáveis. Alguns autores 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. 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. 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.