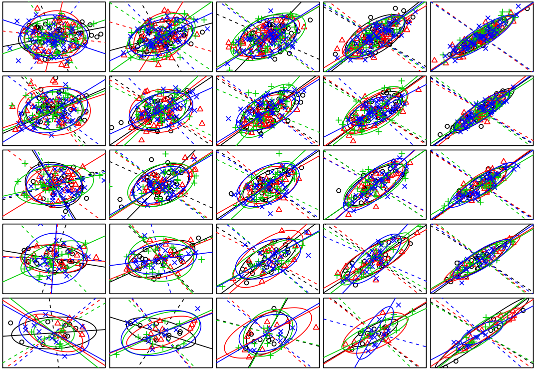
Estatística Multivariada

Slides de apoio às aulas



Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2018/19

Aula 7

Análise em Componentes Principais

- A Análise em Componentes Principais (ACP) é uma técnica exploratória de análise de dados multivariados não estruturados.

- A existência de um número de variáveis demasiadamente elevado coloca, em regra, problemas no processo de análise dos dados (*Curse of dimensionality*)

- *Objetivo geral*: Agrupar as variáveis observadas num número de novas variáveis, designadas por *componentes principais*, inferior ao número de variáveis inicial, perdendo o mínimo de informação estatística (variabilidade) possível.

- *Objetivos específicos*:

  1. redução da dimensionalidade (representação da maior quantidade possível de informação estatística usando um menor número de variáveis);

  2. transformação de variáveis correlacionadas em variáveis independentes e constituição de dimensões mais homogéneas (dentro)/heterogéneas (entre) do que as iniciais.

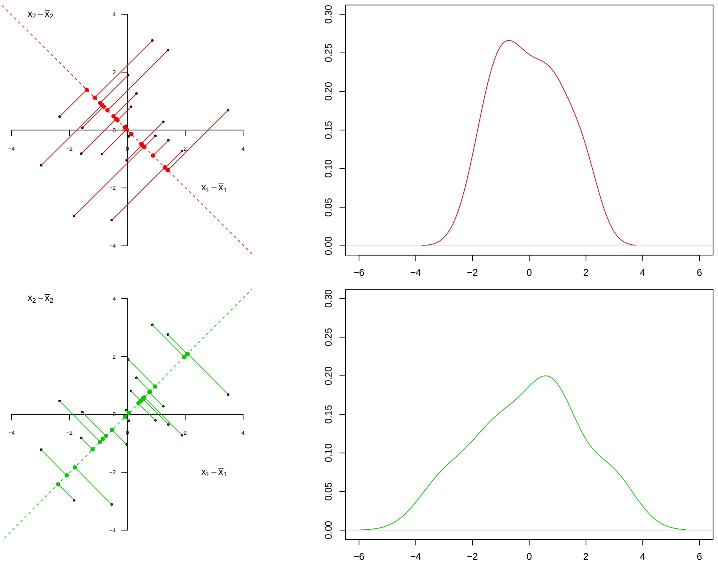
- O objetivo principal de uma ACP é então o de reduzir a dimensão p da matriz inicial $\mathbf{X}_{n\times p}$ à *mais interessante* matriz de todas as que têm menor dimensão do que a inicial ($\mathbf{Y}_{n\times q}$, com $q < p$).

- A solução consiste em encontrar um conjunto de vetores (ortogonais) que definam a direção de projeção de X *mais interessante* num subespaço de dimensão inferior à inicial.

- Do ponto de vista estatístico pode entender-se como *mais interessante* a direção de projeção que permite preservar a maior informação/variabilidade possível e, ao mesmo tempo, permite minimizar as distâncias entre os pontos observados e os pontos projetados.

- A projeção que minimiza o erro (perda de informação) é a definida pela direção de maior variabilidade.

- Ou seja, supondo uma nuvem de pontos elipsóide, o que se pretende é encontrar os eixos "naturais" dessa nuvem com origem em $\bar{\mathbf{x}}$, o que é feito centrando o vetor $\mathbf{x}$ e rodando os eixos.



$x_2 - \bar{x}_2$
$x_1 - \bar{x}_1$

O que são componentes principais?

- Chamam-se **componentes principais** (amostrais) às p combinações lineares $\mathbf{y} = (y_1, y_2, ..., y_p)$ das p variáveis originais $\mathbf{x} = (x_1, x_2, ..., x_p)$:

$$y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$$
$$y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p$$
$$...$$
$$y_p = a_{p1}x_1 + a_{p2}x_2 + ... + a_{pp}x_p$$

  onde $a_{kj}\,(k, j = 1, ..., p)$ são coeficientes desconhecidos (a estimar).

- A matriz

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1p} \\ a_{21} & \dots & a_{2p} \\ \vdots & \vdots & \vdots \\ a_{q1} & \dots & a_{qp} \end{bmatrix}$$

  é designada por matriz de *saturações*, ou matriz dos *pesos fatoriais* (*PC loadings*).

- Assim,

$$\mathbf{y} = a_{k1}x_1 + a_{k2}x_2 + ... + a_{kp}x_p = \sum_{j=1}^{p} a_{kj}x_j = \mathbf{a}_k'\mathbf{x}$$

  representa as $k\,(k = 1, ..., p)$ componentes principais, com média $\mathbf{a}_k'\bar{\mathbf{x}}$ e variância $\mathbf{a}_k'\mathbf{S}\mathbf{a}_k$

Como se obtêm as componentes principais?

- A obtenção das k CP tem início com a determinação da combinação linear com variância máxima correspondente à primeira componente principal

$$\mathbf{a}_1'\mathbf{x}$$

- Assim, pretende-se estimar o vetor $\mathbf{a}_1$ de dimensão p que maximiza

$$var(\mathbf{a}_1'\mathbf{x}) = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1$$

- O máximo da forma quadrática $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$ restrita a $\mathbf{a}_1'\mathbf{a}_1 = 1$ (normalizado) é dado pelo maior valor próprio da matriz $\mathbf{S}$, ℓ_1, com vetor próprio correspondente $\mathbf{e}_1$, isto é, o máximo de $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$ ocorre para

$$\hat{\mathbf{a}}_1 = \mathbf{e}_1$$

- A sucessivas CP, $\mathbf{a}_k'\mathbf{x}$ com $k \geq 2$, são determinadas por maximização de $\mathbf{a}_k'\mathbf{S}\mathbf{a}_k$ sujeita à ausência de correlação com $\mathbf{a}_{k-1}'\mathbf{x}$, equivalendo à determinação dos pares $(\ell_k, \mathbf{e}_k)\,(k = 1, ..., p)$

- Em resumo, os vetores $\mathbf{a}_k$ são estimados sucessivamente de modo a satisfazer as seguintes condições:

1. As variâncias das componentes têm valores decrescentes para $k = 1, \ldots, p$, i.e., $var(\mathbf{a}_1'\mathbf{x}) \geq var(\mathbf{a}_2'\mathbf{x}) \geq \ldots \geq var(\mathbf{a}_p'\mathbf{x})$ correspondentes aos valores próprios ℓ_k, da matriz de covariâncias $\mathbf{S}$, isto é:
$$\begin{aligned} \ell_1 &= var(\mathbf{a}_1'\mathbf{x}) \\ \ell_2 &= var(\mathbf{a}_2'\mathbf{x}) \\ &\ldots \\ \ell_p &= var(\mathbf{a}_p'\mathbf{x}) \end{aligned}$$

2. A covariância entre quaisquer duas componentes k e k' é nula, i.e., $\mathbf{a}_{k'}'\mathbf{S}\mathbf{a}_k = 0, \; k \neq k'$;

3. Para cada componente, os vetores $\mathbf{a}_k$ possuem norma unitária:
$$a_{k1}^2 + a_{k2}^2 + \ldots + a_{kp}^2 = 1, \; (k = 1, \ldots, p)$$
sendo estimados pelos k vetores próprios de $\mathbf{S}$.

Decomposição da variância total

- Seja $\mathbf{A} = (\mathbf{a}_1, ..., \mathbf{a}_p)$ estimada pelos de vetores próprios $(\mathbf{e}_1, ..., \mathbf{e}_p)$ e $\mathbf{y}$ o vetor de componentes principais $\mathbf{y} = \mathbf{a}'\mathbf{x}$, então a matriz de covariâncias de $\mathbf{y}$ é dada por

$$var(\mathbf{y}) = \mathbf{A}'\mathbf{S}\mathbf{A} = \begin{bmatrix} \ell_1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \ell_p \end{bmatrix}$$

- As variâncias das componentes principais são estimadas pelos valores próprios $\ell_1, ..., \ell_p$ sendo a sua soma o traço de $\mathbf{A}'\mathbf{S}\mathbf{A}$

$$\sum_{k=1}^{p} var(y_k) = \sum_{k=1}^{p} \ell_k = tr(\mathbf{A}'\mathbf{S}\mathbf{A})$$

- Por outro lado, $tr(\mathbf{A}'\mathbf{S}\mathbf{A}) = tr(\mathbf{S}\mathbf{A}\mathbf{A}') = tr(\mathbf{S}) = \sum_{j=1}^{p} var(x_j)$
- Logo

$$\sum_{j=1}^{p} var(x_j) = \sum_{k=1}^{p} var(y_k)$$

- Assim, a proporção da variância total original explicada por cada uma das CP é dada por

$$\frac{\ell_k}{\sum_{k=1}^{p} \ell_k}$$

e as primeiras m CP explicam uma proporção cumulativa de $\frac{\sum_{k=1}^{m} \ell_k}{\sum_{k=1}^{p} \ell_k}$

Seja $\mathbf{X} \in \mathbb{R}^{20\times 2}$ a matriz de 20 valores observados para duas variáveis (x_1, x_2). Considere-se a matriz de covariâncias

$$\mathbf{S} = \begin{bmatrix} 2.459 & 1.019 \\ 1.019 & 2.472 \end{bmatrix}$$

Valores próprios de **S**: $\ell_1 = 3.484 \wedge \ell_2 = 1.447$

Vetores próprios de **S**: $\mathbf{e}_1' = (0.705, 0.709)$ e $\mathbf{e}_2 = (-0.709, 0.705)$.

Código R:

```
> round(S,3)

      [,1]  [,2]
[1,] 2.459 1.019
[2,] 1.019 2.472

> round(eigen(S)$values,3)

[1] 3.484 1.447

> round(eigen(S)$vectors,3)

      [,1]   [,2]
[1,] 0.705 -0.709
[2,] 0.709  0.705
```

Sendo as CP dadas por

$$y_1 = 0.705x_1 + 0.709x_2$$

$$y_2 = -0.709x_1 + 0.705x_2$$

explicando, respetivamente, 71% e 20% da variabilidade original.

Código R:

```
> round(eigen(S)$values[1]/sum(eigen(S)$values),2)
[1] 0.71
> round(eigen(S)$values[2]/sum(eigen(S)$values),2)
[1] 0.29
```

Variáveis centradas

- Frequentemente, consideram-se os vetores centrados $\mathbf{x}_c = \mathbf{x} - \bar{\mathbf{x}}$ na obtenção das componentes principais

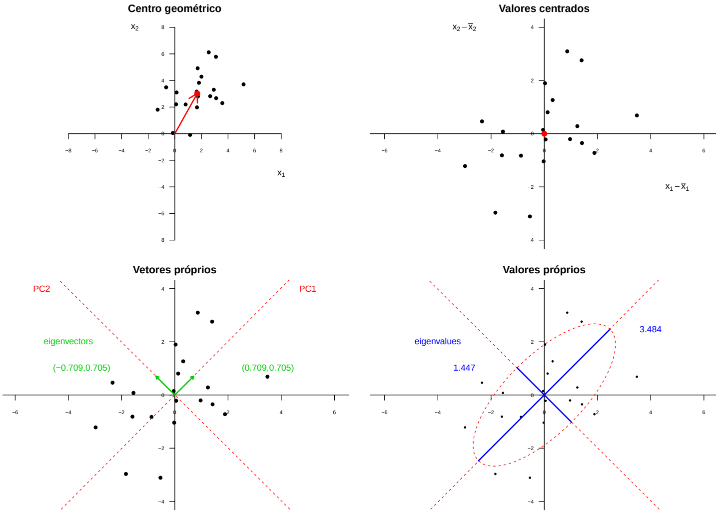
- A transformação de $\mathbf{x}$ em $\mathbf{x}_c$ apenas altera a média de cada variável, que passa a ser zero

- Geometricamente, o centro de gravidade da nuvem de n passa a ser a origem, ou seja, há uma translação do centro de gravidade:

$$(x_1, ..., x_p) \rightarrow (0, ..., 0)$$

- Com esta transformação as componentes principais passam a ter as seguintes características finais:

  1. $\bar{y}_k = 0$
  2. $var(y_k) = \ell_k$
  3. $cov(y_k, y_{k'}) = 0$
  4. $var(y_1) \geq var(y_2) \geq ... \geq var(y_k) \geq 0$
  5. $\sum_{k=1}^{p} var(y_k) = tr(\mathbf{S})$
  6. $\prod_{k=1}^{p} var(y_k) = |\mathbf{S}|$


Centro geométrico
x_2
x_1
Valores centrados
$x_2 - \bar{x}_2$
$x_1 - \bar{x}_1$
Vetores próprios
PC2
PC1
eigenvectors
(−0.709,0.705)
(0.709,0.705)
Valores próprios
eigenvalues
3.484
1.447

- Em muitos casos as variáveis possuem unidades de medida distintas e/ou diferentes graus de heterogeneidade.
- Variáveis com diferentes unidades de medida, não podem ser incluídas conjuntamente na análise porque não são diretamente comparáveis.
- Por outro lado, variáveis mais heterogéneas (com maior variância) têm mais peso na constituição das CP do que as variáveis mais homogéneas e são, por isso, integradas nas primeiras CP (distorcendo a solução final).
- Para ultrapassar este problema é prática comum proceder à padronização das variáveis, i.e., determinar $z_{ij} = \frac{\mathbf{x}_{ij} - \bar{\mathbf{x}}_j}{s_j}$, $(i = 1, ..., n; \, j = 1, ..., p)$, sendo a matriz inicial $\mathbf{X}$ substituída pela matriz $\mathbf{Z}$ com elementos z_{ij}.
- No caso das variáveis estarem padronizadas a matriz $\mathbf{S}$ é igual à matriz das correlações, $\mathbf{R}$.
- Quando as variáveis não estão padronizadas, a solução encontrada partindo de $\mathbf{S}$ ou de $\mathbf{R}$ não é a mesma $\rightarrow$ diferentes vetores próprios
- A maioria dos autores recomenda a padronização prévia das variáveis para que a solução não seja distorcida pela ordem de grandeza e/ou heterogeneidade das variáveis (já que a variância é influenciada pela magnitude dos valores observados).

Considere-se a matriz $\mathbf{X}_{20\times 2}$ e a respetiva matriz de correlações

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.413 \\ 0.413 & 1.000 \end{bmatrix}$$

Valores próprios de **R**: $\ell_1 = 1.413 \wedge \ell_2 = 0.587$

Vetores próprios de **R**: $\mathbf{e}_1' = (0.707, 0.707)$ e $\mathbf{e}_2' = (-0.707, 0.707)$

Código R:

```
> R<-cor(X)
> round(R,3)

      [,1]  [,2]
[1,] 1.000 0.413
[2,] 0.413 1.000

> round(eigen(R)$values,3)

[1] 1.413 0.587

> round(eigen(R)$vectors,3)

      [,1]   [,2]
[1,] 0.707 -0.707
[2,] 0.707  0.707
```

Propriedades das componentes principais

- A soma das variâncias das p CP é igual à soma da variância das p variáveis originais dada por $\sum_{k=1}^{p} \ell_k$. Quando as variáveis são padronizadas $\sum_{k=1}^{p} \ell_k = p$ (porque a variância de cada variável é 1).

- A *proporção da variância total* explicada pela componente k é dada por

$$\frac{\ell_k}{\sum_{k=1}^{p} \ell_k} = \frac{\ell_k}{tr(\mathbf{S})}$$

- Quando as variáveis são padronizadas esta proporção é dada por $\frac{\ell_k}{p}$ $(k = 1, ..., p)$.

- A correlação entre a j-ésima variável e a k-ésima CP $y_k = a_k' \mathbf{x}$ é dada por

$$r_{(x_j, y_k)} = \sqrt{\frac{\ell_k}{s_{jj}}}\, \mathbf{e}_{kj}$$

- Donde, em variáveis padronizadas,

$$r_{(z_j, y_k)} = \sqrt{\ell_k}\, \mathbf{e}_{kj}$$

sendo os valores destas correlações particularmente úteis na interpretação das CP.

Número de componentes a reter

- O principal objetivo em muitas aplicações da ACP é substituir as p variáveis originais por um número de novas variáveis $q << p$ retendo a maior quantidade de informação estatística possível. É crucial saber quão pequeno pode ser q sem que a perda de informação seja "relevante".

- Existem vários critérios de decisão quanto ao número de componentes a reter:

1. **Percentagem de variabilidade explicada**: Reter um número q tal que a percentagem cumulativa de variabilidade explicada pelas q CP não seja inferior a um valor de corte fixado *a priori*. Tipicamente este valor situa-se entre 70 e 90%.

2. **Critério de Kaiser**: Segundo este critério retêm-se as componentes cujo valor próprio é superior à média dos valores próprios $\sum_{k=1}^{p} \ell_i / p$.
   Em variáveis padronizadas, uma componente com valor próprio inferior a 1 indica que não retem a informação (variabilidade) equivalente a uma das variáveis originais (com variância 1, por estarem padronizadas). Assim, de acordo com este critério, em variáveis padronizadas, são retidas as CP com valor próprio superior a 1.

3. **Scree plot**: Um *scree plot* relaciona os valores próprios com as CP ordenadas por ordem de grandeza dos respetivos valores próprios. O número de CP é ditado pelo ponto onde a linha desenhada diminui acentuadamente de inclinação, tornando-se tendencialmente horizontal.

4. **Teste aos valores próprios**: Reter as primeiras q componentes se

$$H_{0q} : \lambda_{q+1} = \ldots = \lambda_p$$

   não for rejeitada (abordado na secção de inferência sobre valores próprios)

Exemplo 2

Os dados do ficheiro "data8.xlsx" foram recolhidos como parte de um estudo preliminar de uma possível ligação entre design de capacete em jogadores de futebol americano e lesões no pescoço. Foram amostrados 60 sujeitos de três grupos tendo sido realizadas seis medições:

- WDIM = largura da cabeça (maior dimensão);
- CIRCUM = circunferência da cabeça;
- FBEYE = medição da frente para trás ao nível dos olhos;
- EYEHD = medição do olho ao topo da cabeça;
- EARHD = medida da orelha ao topo da cabeça;
- JAW = largura da mandíbula.

Faça uma análise em componentes principais dos dados do ficheiro.

Código R:

```
> dados8<-as.data.frame(readxl::read_xlsx("./Datasets/data8.xlsx", col_names = TRUE))
> dados<-scale(dados8,center = TRUE,scale = TRUE)
> S<-var(dados)
> round(S,3)

        WDIM CIRCUM  FBEYE  EYEHD  EARHD    JAW
WDIM   1.000  0.608  0.361  0.060  0.252  0.605
CIRCUM 0.608  1.000  0.733  0.336  0.090  0.410
FBEYE  0.361  0.733  1.000  0.014 -0.028  0.311
EYEHD  0.060  0.336  0.014  1.000  0.297 -0.079
EARHD  0.252  0.090 -0.028  0.297  1.000 -0.090
JAW    0.605  0.410  0.311 -0.079 -0.090  1.000
```

Adequabilidade da ACP

- Em regra nos procedimentos multivariados $n > p$. Contudo, hoje em dia, fruto da crescente capacidade computacional de registo e armazenamento de informação são frequentes as situações onde $n < p$ (e.g. expressão de genes).

- Nestes casos nada impede que se realize uma ACP - *Sparse ACP*. Contudo, o número de CP fica limitado pelo $min\{n, p\}$, ou seja, se $n < p$ poderão ser definidas apenas n CP para explicar toda a variabilidade original.

- A ACP baseia-se na existência de padrões de associação estatística entre as variáveis originais, isto é na existência de correlações (pelo menos uma parte delas) elevadas.

- Antes de efetuar uma ACP deve por isso analisar-se a **matriz das correlações**

- Porém, correlações bivariadas elevadas não são condição suficiente para garantir a existência de variáveis latentes. A existência de um padrão latente que simultaneamente afeta várias variáveis não é detectada por este tipo de correlações. Para detectar este tipo de estrutura é aconselhável a análise da **matriz de correlações parciais**

- As correlações parciais são medidas de associação linear bivariada, retirando o efeito de todas as restantes variáveis em análise. Para $p = 3$

$$r_{x,y|z} = \frac{r_{xy} - r_{yz} r_{xz}}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}}$$

- Os seus valores quantificam a contribuição unitária, não comum, de cada variável para a variância total. Quanto maior for a variância comum, menor o valor das correlações parciais. Assim, quanto menores forem as correlações parciais, mais adequada será a ACP.

- **Teste de Esfericidade (Mauchly)**: Este procedimento permite testar se a matriz de correlações é igual à matriz identidade:

$$H_0 : \boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

isto é, se as variáveis são independentes e têm a mesma variância, equivalente a

$$H_0 : \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$$

Qual a importância deste procedimento?

- A aplicação deste teste supõe a existência de normalidade multivariada, i.e., $\mathbf{x} \frown N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Sob H_0, a estatística do teste (U^*)

$$U^* = -\left(n - 1 - \frac{2p^2 + p + 2}{6p}\right) \ln U$$

com $U = \Lambda^{2/n}$ e $\Lambda = \frac{|\mathbf{S}|^{n/2}}{(tr(\mathbf{S})/p)^{np/2}}$, ou seja,

$$U = \frac{p^p |\mathbf{S}|}{(tr(\mathbf{S}))^p}$$

tem distribuição aproximada $\chi^2_{\frac{1}{2}p(p+1)-1}$

Interpretação das componentes principais

- A interpretação das CP é frequentemente feita atendendo aos *loadings* (elementos dos vetores próprios)

- Tal como vimos anteriormente a correlação ente a k-ésima CP e a j-ésima variável original (dados standartizados) é dada por: $r_{kj} = \sqrt{\ell_k} e_{kj}$

- Assim, valores absolutos elevados indicam uma "identificação"das CP com essas variáveis, devendo a interpretação das CP basear-se nessas variáveis

- Regras usuais para considerar as correlações "elevadas":

  1. $r_{k\cdot}^2 \geq \ell_k / \sum_{k=1}^{p} \ell_k$ ou $r_{k\cdot}^2 \geq \ell_k / p$ (var. standartizadas)
  2. $|r_{kj}^2 \geq 0.5|$ $r_{kj}^2 \geq 0.25$

- Como interpretar r_{kj}^2? Proporção de variabilidade j-ésima variável original explicada pela k-ésima CP

- Advertência: É frequente ignorar variáveis com correlações (ou loadings) próximas de zero. Isto pode induzir em erro, e convém utilizar informação complementar para validar as interpretações baseadas nos coeficientes!

- Até aqui a ACP foi abordada numa perspectiva estritamente exploratória, descritiva. A utilização desta técnica com um objectivo inferencial obriga a considerar as **distribuições amostrais dos valores e vetores próprios** da matriz de covariâncias.

- Notação:

1. **População**:

   $\boldsymbol{\lambda}' = (\lambda_1, \ldots, \lambda_p) \rightarrow$ vetor de valores próprios de $\boldsymbol{\Sigma}$, com elementos λ_k $(k = 1, \ldots, p)$

   $\boldsymbol{\alpha}_k$ $(k = 1, \ldots, p) \rightarrow$ vetores próprios de $\boldsymbol{\Sigma}$, com elementos α_{kj}

2. **Amostra**:

   $\boldsymbol{\ell}' = (\ell_1, \ldots, \ell_p) \rightarrow$ vetor de valores próprios de $\mathbf{S}$, com elementos ℓ_k $(k = 1, \ldots, p)$ $\mathbf{a}_k$ $(k = 1, \ldots, p) \rightarrow$ vetores próprios de $\mathbf{S}$, com elementos a_{kj}

- Numa população normal multivariada de dimensão p, i.e., $\mathbf{x} \frown N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ os EMV dos valores próprios (todos positivos e distintos) e vetores próprios de $\boldsymbol{\Sigma}$ são os valores e vetores próprios de $\mathbf{S}$, i.e,

$$\hat{\lambda}_k = \ell_k$$

$$\hat{\boldsymbol{\alpha}}_k = \mathbf{a}_k$$

1. **Valores próprios**: ℓ_k tem distribuição de probabilidade assintótica normal multivariada (dimensão p) com

    - Valor médio: $E(\ell_k) = \lambda_k$
    - Variância: $Var(\ell_k) = \frac{2\lambda_k^2}{n-1}$

    ou seja,

$$\ell_k \overset{a}{\sim} N\left(\lambda_k, \frac{2\lambda_k^2}{n-1}\right)$$

2. **Vetores próprios**: $\mathbf{a}_k$ tem distribuição de probabilidade assintótica normal multivariada (dimensão p) com

    - Valor médio: $E(\mathbf{a}_k) = \boldsymbol{\alpha}_k$
    - Variância:

$$Var(\mathbf{a}_k) = \frac{\lambda_k}{n-1} \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j - \lambda_k)^2} \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j' \; (j \neq k)$$

    ou seja,

$$\mathbf{a}_k \overset{a}{\sim} N_p\left(\boldsymbol{\alpha}_k, \frac{\lambda_k}{n-1} \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j - \lambda_k)^2} \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j'\right)$$

- Para n suficientemente grande:

1. ℓ_k e $\mathbf{a}_k$ são variáveis aleatórias independentes;

2. Distribuição assintótica centrada e reduzida de ℓ_k:

$$\sqrt{n-1}\,(\ell_k - \lambda_k) \xrightarrow[n\to\infty]{d} N(0, 2\lambda_k^2)$$

Dado que, pelo Teorema de Slutsky, $\lambda_k/\ell_k \xrightarrow[n\to\infty]{p} 1$,

$$\frac{\sqrt{n-1}\,(\ell_k - \lambda_k)}{\ell_k\sqrt{2}} \xrightarrow[n\to\infty]{d} N(0, 1)$$

3. Distribuição assintótica centrada e reduzida de $\mathbf{a}_k$:

$$\sqrt{n-1}\,(\mathbf{a}_k - \boldsymbol{\alpha}_k) \xrightarrow[n\to\infty]{d} N(0, \mathbf{T}_k)$$

com $\mathbf{T}_k = \lambda_k \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j - \lambda_k)^2}\boldsymbol{\alpha}_j\boldsymbol{\alpha}_j'\ (j \neq k)$. Dado que, pelo Teorema de Slutsky, $\boldsymbol{\alpha}_j/\mathbf{a}_j \xrightarrow[n\to\infty]{p} 1$,

$$\frac{\sqrt{n-1}(\mathbf{a}_k - \boldsymbol{\alpha}_k)}{\left(\ell_k \sum_{j=1}^{p} \frac{\ell_j}{(\ell_j - \ell_k)^2}\mathbf{a}_j\mathbf{a}_j'\right)^{1/2}} \xrightarrow[n\to\infty]{d} N(0, 1)$$

- Considerando que $\sqrt{n-1}\,(\ell_k - \lambda_k) \overset{a}{\sim} N(0, 2\lambda_k^2)$, o intervalo de confiança assintótico para λ_k a $(1-\delta)\times 100$ é dado por

$$\left(\frac{\ell_k}{1 + z_{1-\delta/2}\sqrt{2/(n-1)}}, \frac{\ell_k}{1 - z_{1-\delta/2}\sqrt{2/(n-1}}\right)$$

- Tendo em conta o Teorema de Slutsky, o intervalo de confiança assintótico para λ_k a $(1-\delta)\times 100$ pode aproximar-se por

$$\left(\ell_k - z_{1-\delta/2}\,\ell_k\sqrt{\frac{2}{n-1}}, \ell_k + z_{1-\delta/2}\,\ell_k\sqrt{\frac{2}{n-1}}\right)$$

$z_{1-\delta/2}$ representa o quantil de probabilidade $1-\delta/2$ da distribuição normal padrão.

Estimação intervalar

- No caso dos vetores próprios, mostra-se que (Mardia et al., 1979, p. 233 e Joliffe, 2002, p. 52) que (baseado no comportamento assintótico das distâncias de Mahalanobis)

$$(n-1)(\mathbf{a}_k - \boldsymbol{\alpha}_k)'\mathbf{W}(\mathbf{a}_k - \boldsymbol{\alpha}_k) \overset{a}{\sim} \chi^2_{p-1}$$

  com

$$\mathbf{W} = \ell_k \mathbf{S}^{-1} + \ell_k^{-1}\mathbf{S} - 2\mathbf{I}_p$$

- Sendo a região de confiança para $\boldsymbol{\alpha}_k$ a $(1-\delta)\times 100$ é dada por

$$(n-1)\boldsymbol{\alpha}_k'(\ell_k \mathbf{S}^{-1} + \ell_k^{-1}\mathbf{S} - 2\mathbf{I}_p)\boldsymbol{\alpha}_k \leq \chi^2_{p-1}(1-\delta)$$

- Para testar $H_0 : \lambda_k = \lambda_{k0}$ tem-se a estatística adequada

$$\sqrt{\frac{n-1}{2}}\frac{\ell_k - \lambda_{k0}}{\lambda_{k0}}$$

com distribuição assintótica $N(0,1)$, sob H_0

- Para testar a existência de esfericidade parcial, i.e., $H_0 : \lambda_{q+1} = ... = \lambda_p$ tem-se a estatística adequada

$$Q = \left[\frac{\prod_{k=q+1}^{p} \ell_k}{\left(\sum_{k=q+1}^{p} \ell_k/(p-q)\right)^{p-q}}\right]^{n/2}$$

com $-2\ln Q \frown \chi^2_\nu$ e $\nu = \frac{1}{2}(p-q+2)(p-q-1)$, sob H_0

Testes de hipóteses

- Para testar $H_0 : \boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k0}$ tem-se a estatística adequada

$$(n-1)\boldsymbol{\alpha}'_{k0}\left(\ell_k \mathbf{S}^{-1} + \ell_k^{-1}\mathbf{S} - 2\mathbf{I}_p\right)\boldsymbol{\alpha}_{k0}$$

e distribuição assintótica χ^2_{p-1}, sob H_0