Estatística Multivariada
Algumas revisões de conceitos e resultados importantes
2023/2024

Carlos A. Coelho
(cmac@fct.unl.pt)

Departmento de Matemática
NOVA School of Science and Technology

- Propriedade do Traço

$$Tr(ABC) = Tr(BCA) = Tr(CAB)$$

- Característica de uma matriz

$$\begin{aligned} C(A) &= \text{\# colunas ou linhas linearmente independentes} \\ &= \text{\# valores próprios diferentes de 0 (zero)} \end{aligned}$$

- Uma matriz A (quadrada) diz-se idempotente sse

$$A^2 = A$$

  - se A for idempotente os seus valores próprios serão 0 (zero) ou 1 (Demonstrar!)
  - se A for idempotente $C(A) = Tr(A)$ (Demonstrar!)

- Propriedades do Determinante (seja A uma matrix $p \times p$)

$$|A^{-1}| = (|A|)^{-1}$$

$$|nA| = n^p\,|A|$$

$$|AB| = |A|\,|B|$$

$|A| = \prod_{j=1}^{p} a_{jj}$ se A for diagonal ou triangular superior ou inferior

$|A| = \prod_{\alpha=1}^{p} \lambda_\alpha$ (onde λ_α $(\alpha = 1, \ldots, p)$ são os valores próprios de A

- Uma outra propriedade do Traço

$Tr(A) = \sum_{\alpha=1}^{p} \lambda_\alpha$ (onde λ_α $(\alpha = 1, \ldots, p)$ são os valores próprios de A

- Propriedade da transposição: $(AB)' = B'A'$
- Propriedade da inversão: $(AB)^{-1} = B^{-1}A^{-1}$
- Propriedades das matrizes ortogonais: seja U uma matriz ortogonal (de facto ortonormada) $(p \times p)$, então: $UU' = I_p$ e $|U| = \pm 1$

– Sobre a decomposição espetral de uma matriz: $A\underline{u}_\alpha = \lambda_\alpha \underline{u}_\alpha \quad (\alpha = 1, \ldots, p)$ pode ser escrito em notação matricial como

$$AU = U\Lambda \iff A = U\Lambda U'$$

onde U é uma matriz ortonormada, i.e., onde $UU' = U'U = I_p$;
o que mostra que

$$|A| = |U\Lambda U'| = |U|\,|\Lambda|\,|U'| = |\Lambda|\,|U'U| = |\Lambda|\,|I_p| = |\Lambda| = \prod_{\alpha=1}^{p} \lambda_\alpha$$

$$Tr(A) = Tr(U\Lambda U') = Tr(\Lambda U'U) = Tr(\Lambda) = \sum_{\alpha=1}^{p} \lambda_\alpha$$

$$A = U\Lambda U' \iff A^{-1} = (U')^{-1}\Lambda^{-1}U^{-1} = U\Lambda^{-1}U' \iff |A| = \prod_{\alpha=1}^{p} \frac{1}{\lambda_\alpha} = \frac{1}{|A|}$$

$A^k = U\Lambda^k U' \quad$ (para $k \in \mathbb{N}$)

Determinante de matriz 'particionada'

$$\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = |A_{11}|\,|A_{22} - A_{21}A_{11}^{-1}A_{12}|$$
$$= |A_{22}|\,|A_{11} - A_{12}A_{22}^{-1}A_{21}|$$

Um outro resultado útil: seja A $m\times n$ e B $n\times m$, então

$$|I_m + AB| = |I_n + BA|\,.$$

Exemplo de aplicação: seja

$$A = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}.$$

Obter $|I_4 + AA'|$.
Temos

$$|I_4 + AA'| = |I_1 + A'A| = 1 + [1 \;\; 2 \;\; 3 \;\; 4] \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = 1 + 1 + 4 + 9 + 16 = 31$$

– Momento não-centrado de ordem h

$$\begin{aligned} E(X^h) &= \int_S x^h f_X(x)\,dx \quad \text{(v.a.'s contínuas)} \\ &= \sum_{x \in S} x^h f_X(x) \quad \text{(v.a.'s discretas)} \end{aligned}$$

– Momento centrado de ordem h

$$\begin{aligned} E[(X-\mu)^h] &= \int_S (x-\mu)^h f_X(x)\,dx \quad \text{(v.a.'s contínuas)} \\ &= \sum_{x \in S} (x-\mu)^h f_X(x) \quad \text{(v.a.'s discretas)} \end{aligned}$$

$$h = 2 \implies E[(X-\mu)^2] = Var(X) = E(X^2) - (E(X))^2 \quad \text{(Demonstrar!)}$$

– Função geradora de momentos (f.g.m.) de X

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) \qquad \text{(nem sempre existe)} \\ &= \int_S e^{tx} f_X(x)\,dx \quad \text{(v.a.'s contínuas)} \\ &= \sum_{x \in S} e^{tx} f_X(x)\,dx \quad \text{(v.a.'s discretas)} \end{aligned}$$

– Função característica (f.c.) de X: $\Phi_X(t) = E\left(e^{\mathrm{i}tX}\right)$ (existe sempre)

Propriedades elementares do valor esperado:

- $E(aX) = aE(X)$
- $E(a) = a$
- $E(X \pm Y) = E(X) \pm E(Y)$
- $E(a \pm bX) = a \pm bE(X)$

Propriedades elementares da variância:

- $Var(aX) = a^2 Var(X)$
- $Var(a) = 0$
- $Var(X \pm Y) = Var(X) + Vay(Y) \pm 2Cov(X,Y)$
- $Var(a \pm bX) = b^2 Var(X)$

Estimator: r.v. which is a statistic, that is, a function of the random sample (the X_i's), and is used to estimate parameters of distributions or population quantities

Unbiased Estimator: Let $\hat{\theta}$ be an estimator of θ. $\hat{\theta}$ is an unbiased estimator of θ iff $E(\hat{\theta}) = \theta$.

Bias: the bias of $\hat{\theta}$ is $\quad Bias(\hat{\theta}) = b(\hat{\theta}) = E(\hat{\theta}) - \theta$

Consistency: $\hat{\theta}$ is said to be consistent iff $\quad \lim_{n\to\infty} MSE(\hat{\theta}) = 0$

where

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] \qquad \text{(Mean Squared Error)}$$

with

$$\begin{aligned}
MSE(\hat{\theta}) &= E\left[\left(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\right)^2\right] \\
&= E\left[\left(\hat{\theta} - E(\hat{\theta})\right)^2 + \left(E(\hat{\theta}) - \theta\right)^2 + 2\left(\hat{\theta} - E(\hat{\theta})\right)\left(E(\hat{\theta}) - \theta\right)\right] \\
&= E\left[\left(\hat{\theta} - E(\hat{\theta})\right)^2\right] + \left(E(\hat{\theta}) - \theta\right)^2 + 2\left(E(\hat{\theta}) - \theta\right) E\left[\hat{\theta} - E(\hat{\theta})\right] \\
&= Var(\hat{\theta}) + \left(b(\hat{\theta})\right)^2 + 2\left(E(\hat{\theta}) - \theta\right)\left(E(\hat{\theta}) - E(\hat{\theta})\right) \\
&= Var(\hat{\theta}) + \left(b(\hat{\theta})\right)^2
\end{aligned}$$

Define-se

$$E[g(X)] = \int_S g(x)\, f_X(x)\, dx \quad \text{(v.a.'s contínuas)}$$

$$= \sum_{x \in S} g(x)\, f_X(x) \quad \text{(v.a.'s discretas)}$$

As funções $f_X(x)$, $F_X(x)$, $M_X(t)$ e $\Phi_X(y)$ identificam de forma única as v.a.'s.

Exercício: Seja $X \sim Beta(a,b)$. Obtenha $E(X^h)$ e diga para que valores de h é válida a expressão obtida.

Exercício: Seja $X \sim Beta(a,b)$ e seja $Y = -\log X$. Obtenha a f.g.m. e a f.c. de Y.

Considere as v.a.'s $X_1, \ldots, X_n$ e a função $g(X_1, \ldots, X_n)$. Então

$$E[g(X_1,\ldots,X_n)] = \int_{S_{X_1}} \cdots \int_{S_{X_n}} g(x_1,\ldots,x_n)\, f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, dx_n \ldots dx_1 \quad \text{(v.a.'s contínuas)}$$

$$= \sum_{x_1 \in S_{X_1}} \cdots \sum_{x_n \in S_{X_n}} g(x_1,\ldots,x_n)\, f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \quad \text{(v.a.'s discretas)}$$

As v.a.'s $X_1,\ldots,X_n$ são independentes sse

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=\prod_{i=1}^{n} f_{X_i}(x_i)\,.$$

Note-se que então também teremos

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=\prod_{i=1}^{n} F_{X_i}(x_i)\,.$$

Seja

$$g(X_1,\ldots,X_n)=\prod_{i=1}^{n} g_i(X_i)\,.$$

Então se $X_1,\ldots,X_n$ forem n v.a.'s independentes

$$E\left[g(X_1,\ldots,X_n)\right]=\prod_{i=1}^{n} E\left[g_i(X_i)\right]\,.$$

(Demonstração!)

Sejam $X_1,\ldots,X_n$, n v.a.'s independentes, e seja $S=\sum_{i=1}^{n} X_i$. Então

$$M_S(t)=\ldots\quad \text{(em função de } M_{X_i}(t)\text{)}$$

e

$$\Phi_S(t)=\ldots\quad \text{(em função de } \Phi_{X_i}(t)\text{)}\,.$$

Aplicação: Sejam $X_i \sim \Gamma(r_i,\lambda)$, $i=1,\dots,n$, n v.a.'s independentes, e seja $S=\sum_{i=1}^n X_i$. Qual a distribuição de S ?

Aplicação: Sejam $X_1 \sim Beta(a,b)$ e $X_2 \sim Beta(a+b,c)$ duas v.a.s' independentes, e seja $Y=X_1X_2$. Qual a distribuição de Y ?

Aplicação: Sejam $X_i \sim Beta(a+i,1)$, $i=1,\dots,n$, n v.a.'s independentes, e seja $Y=\prod_{i=1}^n X_i$. Qual a distribuição de Y ?

Estatística Multivariada

Some plots (2D and 3D) and an outlier test

2023/2024

Carlos A. Coelho
(cmac@fct.unl.pt)

Mathematics Department
NOVA School of Science and Technology

The iris dataset

> iris

> summary(iris)

> head(iris)

> tail(iris)

Plotting (2D)

> **plot(iris[,1:4])**

> **plot(iris[,1:5])**

> **plot(iris)**

> **plot(iris[,1:2])**

> **plot(iris[,1:2],pch=as.numeric(iris$Species),col=as.numeric(iris$Species))**

> **plot(iris[,1:2],pch=as.numeric(iris$Species),col=as.numeric(iris$Species, xlab="...",ylab="..."))**

> help(plot)

> **legend("topleft",levels(iris$Species),pch=1:3,col=1:3,box.col="blue")**

- > **text(iris[,1],iris[,2],1:150)**
- > **text(iris[,1],iris[,2],row.names(iris))**
- > **text(.05+iris[,1],.05+iris[,2],row.names(iris),cex=.5)**
- > help(text)

Plotting (3D) (3D rotatable scatterplot – package 'rgl')

- > **install.packages("rgl")**
- > **library(rgl)**
- > **plot3d(iris[,1],iris[,2],iris[,3],size=1,col=as.numeric(iris$Species),type="s")**
  (what gives different colors for the 3 species)
- > **plot3d(iris[,1],iris[,2],iris[,3],type="n")**
- > **text3d(iris[,1],iris[,2],iris[,3],text=c("∗","+","o")[as.numeric(iris$Species)])**
  (what gives different symbols for the 3 species)
- > **plot3d(iris[,1],iris[,2],iris[,3],size=.5,col=as.numeric(iris$Species),type="s")**
- > **text3d(.1+iris[,1],.1+iris[,2],.1+iris[,3],row.names(iris),**
  font=1+as.numeric(iris$Species))
  (gives points identified with their numbers, with different fonts for each species)

a better idea may be:

```
> plot3d(iris[,1:3],type="s",size=.3,col=as.numeric(iris$Species))
> text3d(.1+iris[,1],.1+iris[,2],.1+iris[,3],row.names(iris),col=as.numeric(iris$Species))
```

Legends in 3D scatter plots

```
> legend3d("topleft",levels(iris$Species),pch=16,col=c(1,2,3),magnify=3)
```

Objective: program in R a simple function that given the dataset, the 3 axes, and the index (for 'species') plots a 3D scatterplot similar to the one we obtained last.

```
plot3dmine <- function(data,axes,index,offset=.1)
{plot3d(data[,axes],type="s",size=.3,col=as.numeric(index))
 text3d(offset+data[,axes],text=row.names(data),col=as.numeric(index))
 legend3d("topleft",levels(index),pch=16,col=as.numeric(levels(factor(as.numeric(index)))),
                                                                              magnify=3)
 aspect3d("iso")}
```

Objective: program an R function, similar to the function **plot3dmine**, which if there is only a 'single set of points' does not plot the legend and does not need the indication of the 'index' argument.

Using only the Iris species *virginica*:

– we can use

$>$ **virg** $<-$ **iris[101:150,]**

or

$>$ **virg** $<-$ **iris[iris$Species=='virginica',]** (more general)

Points 107, 118 and 132 may be candidates to outliers.
How can we test if a given set of points should be or not considered as 'outliers'?
We will now address a test for 'outliers' in a very much empirical way!
Many details are left to be addressed later.

Test for 'outliers' (Wilks, 1963; Coelho and Arnold, 2019, Sec. 5.1.13):
For a sample of size n from $N_p(\underline{\mu},\Sigma)$, the test statistic to test the null hypothesis

$$H_0 : \text{observations } n_1,\dots,n_k \text{ are not outliers}$$

is

$$\Lambda = \frac{|A^*|}{|A|}$$

where A is equal to n times the usual m.l.e. of Σ, based on the whole sample, that is, on the n observations, and A^* is equal to $n-k$ times the m.l.e. of Σ based on the $n-k$ observations that remain after removing observations $n_1,\dots,n_k$.

Wilks, S.S. (1963). Multivariate statistical outliers. Sankhya Ser. A 25, 407–426

Let $X\ (n\times p)\ (n=50, p=4)$ be the sample matrix.
Then, the m.l.e. of Σ is

$$\underset{(p\times p)}{S} = \frac{1}{n}X'\left(I_n - \frac{1}{n}E_{nn}\right)X\,,$$

where E_{nn} is an $n\times n$ matrix of 1's, and then we have

$$\underset{(p\times p)}{A} = nS = X'\left(I_n - \frac{1}{n}E_{nn}\right)X\,.$$

A simple way to compute A in R, is to use the sample variance-covariance matrix

$$\underset{(p\times p)}{S^*} = \frac{1}{n-1}X'\left(I_n - \frac{1}{n}E_{nn}\right)X\,,$$

and compute A as

$$A = (n-1)S^*\,.$$

For us X is the data matrix (data frame) virg, so that we have

```
> X <- virg[,1:4]
> Sstar <- var(X)
> dim(X)
> A <- (dim(X)[1]-1)*Sstar
```

Then we have X^* $((n-k)\times p)$ the matrix of the sample without the k observations we want to test for outliers, with

$$\underset{(p\times p)}{S^{**}} = \frac{1}{n-k-1}(X^*)'\left(I_{n-k} - \frac{1}{n-k}E_{n-k,n-k}\right)X^*$$

and

$$A^* = (n-k-1)S^{**}\,.$$

In R it will computed as

$>$ **Xstar $<-$ X[-c(7,18,32),]**
$>$ **Sstarstar $<-$ var(Xstar)**
$>$ **Astar $<-$ (dim(Xstar)[1]-1)$*$Sstarstar**

e então finalmente

$>$ **Lambda $<-$ det(Astar)/det(A)**

Objective: to program a function in R to obtain the computed value of Λ, given the original data matrix, the variables to be considered, and the indexes of the points to test for outliers.

```
outliers <- function(data,vars,obs)
{mat <- var(data[,vars])*(dim(data)[1]-1)
 mats <- var(data[-obs,vars])*(dim(data)[1]-length(obs)-1)
 det(mats)/det(mat)}
```

How to carry out the test?

Reject H_0 if the computed value of the test is too small!
That is, if for a given value of α (usually $\alpha = 0.5$ or 0.01), the computed value of Λ is smaller than the α quantile of the distribution of Λ, or, equivalently, if the p-value of the test is smaller than α.

p-value=P(the test statistic has a value equal to the computed value obtained, or another value more in favor of H_1)

so that, in our case,

$$\text{p-value} = P(\Lambda \leq \lambda_{calc}) = F_\Lambda(\lambda_{calc}),$$

rejecting H_0 if p-value$< \alpha$.

Since the α quantile of Λ is the value λ_α for which $P(\Lambda \leq \lambda_\alpha) = \alpha$, we have

$$\begin{aligned} P(\Lambda \leq \lambda_{calc}) < \alpha \quad &\Longleftrightarrow \quad P(\Lambda \leq \lambda_{calc}) < P(\Lambda \leq \lambda_\alpha) \\ &\Longleftrightarrow \quad \lambda_{calc} < \lambda_\alpha \end{aligned}$$

given that the function $P(\Lambda \leq \lambda)$ is monotone increasing in λ.

A test for outliers

So, we need to know the distribution of Λ to be able to compute either the α quantile of Λ, or the p-value.

The distribution of Λ (under H_0) is the same as that of a product of independent Beta r.v.'s, with

$$\Lambda \stackrel{d}{\equiv} \prod_{j=1}^{p} Y_j \stackrel{d}{\equiv} \prod_{\ell=1}^{k} Y_\ell^*$$

where, for $j = 1, \dots, p$ and $\ell = 1, \dots, k$,

$$Y_j \sim Beta\left(\frac{n-k-j}{2}, \frac{k}{2}\right) \quad \text{and} \quad Y_\ell^* \sim Beta\left(\frac{n-p-\ell}{2}, \frac{p}{2}\right)$$

are two sets of independent r.v.'s.

(Note that we need to have $n > k+p$ in order to be able to implement the test.)

The statistic Λ may be written as

$$\Lambda = \frac{|A^*|}{|A^*+B|}$$

where $A^* \sim W_p(n-k-1, \Sigma)$ and $B \sim W_p(k, \Sigma)$.

Homework: Obtain the distribution of Λ from this fact.

Later in this course we will see that it is possible to obtain explicit expressions for the exact p.d.f. and c.d.f. of Λ for most cases, while for the other cases we can obtain near-exact p.d.f.'s and c.d.f.'s.

However, by now, we will use the fact that it is possible to compute the p.d.f. and c.d.f. of Λ by inversion of the c.f. or of the m.g.f..

Package CharFun: go to web-page
https://www.rdocumentation.org/packages/CharFun/versions/0.1.0

download file **CharFun.tar.gz** and place it in a folder of your choice.

Then install the package using R-Studio.

The package CharFun in R can handle simple c.f.'s for inversion.
For even k or even p the c.f. can be highly simplified. Let us take $k=2$, that is, make the test for only 2 points. Let us take points 18 and 32.

Then the distribution of Λ is the same as that of (for $p=3$)

$$Beta\left(\frac{50-3}{2},1\right)\,Beta\left(\frac{50-4}{2},1\right)\,Beta\left(\frac{50.5}{2},1\right).$$

So, we can easily obtain the c.f. not of Λ but of $W=-\log\Lambda$, since

$$\begin{aligned}
\Phi_W(t) &= E\left(e^{\mathrm{i}tW}\right)=E\left(e^{-\mathrm{i}t\log\Lambda}\right)=E\left(\Lambda^{-\mathrm{i}t}\right)=\prod_{j=1}^{3}E\left(Y_j^{\mathrm{i}t}\right)\\
&= \frac{\Gamma\left(\frac{47}{2}+1\right)\Gamma\left(\frac{47}{2}-\mathrm{i}t\right)}{\Gamma\left(\frac{47}{2}\right)\Gamma\left(\frac{47}{2}+1-\mathrm{i}t\right)}\,\frac{\Gamma(23+1)\Gamma(23-\mathrm{i}t)}{\Gamma(23)\Gamma(23+1-\mathrm{i}t)}\,\frac{\Gamma\left(\frac{45}{2}+1\right)\Gamma\left(\frac{45}{2}-\mathrm{i}t\right)}{\Gamma\left(\frac{45}{2}\right)\Gamma\left(\frac{45}{2}+1-\mathrm{i}t\right)}\\
&= \frac{47}{2}\left(\frac{47}{2}-\mathrm{i}t\right)^{-1}23(23-\mathrm{i}t)^{-1}\frac{45}{2}\left(\frac{45}{2}-\mathrm{i}t\right)^{-1}
\end{aligned}$$

So, we have to build the R function

```
> cfw <- function(t){47/2*(47/2-t*1i)^(-1)*23*(23-t*1i)^(-1)*45/2*(45/2-t*1i)^(-1)}
```

and then use the function cf2DistGP from package CharFun to obtain the p-value, with:

```
> result <- cf2DistGP(cfw,-log(.6690471))
> 1-result$cdf
```

which gives the result
0.005142764

This result would lead us to reject H_0 for either $\alpha = 0.05$ or $\alpha = 0.01$, assuming that the 2 points (18 and 32) are outliers.

Why do we use $-log(.6690471)$

We should note that we are dealing with the distribution of the r.v. $W = -\log\Lambda$, and we want

$$\begin{aligned}
\text{p-value} &= P(\Lambda \leq 0.6690471) \\
&= P(\log\Lambda \leq \log(0.6690471)) \\
&= P(-\log\Lambda \geq -\log(0.6690471)) \\
&= 1 - P(W < -\log(0.6690471)) \\
&= 1 - P(W \leq -\log(0.6690471)) \\
&= 1 - F_W(-\log(0.6690471))
\end{aligned}$$

However, it happens that the 2 points were not select randomly but rather by inspecting the plot of the overall set of points, so that we indeed, without may be noticing it, we did C_2^{50} tests to choose these 2 points.
As such it is recommended by Wilks himself that we would rather use an α value equal to α / C_k^n, which in our case would be equal to either

$$0.05/C_2^{50} = 4.08\times 10^{-5} \quad \text{or} \quad 0.01/C_2^{50} = 8.16\times 10^{-6},$$

so that, after all, we should not reject the hypothesis that the 2 points are not outliers.

Function **cf2DistGP** is supposed to make plots of the p.d.f. and the c.d.f. with a command as

$>$ **result** $<-$ **cf2DistGP(cfw)**

but there is a problem in the function, so that in most systems it will not work. You can use the function **cf2DistGP2** in the file `cf2DistGP2.txt`, which should work, by typping

$>$ **result** $<-$ **cf2DistGP2(cfw)**

But the c.f. of W

$$\Phi_W(t) = \frac{47}{2}\left(\frac{47}{2} - \mathrm{i}t\right)^{-1} 23(23 - \mathrm{i}t)^{-1}\frac{45}{2}\left(\frac{45}{2} - \mathrm{i}t\right)^{-1}$$

shows that the distribution of W is, in this case, that of a sum of 3 independent r.v.'s with Exponential distributions with rate parameters 47/2, 23 and 45/2, which is a GIG (Generalized Integer Gamma) distribution of depth 3, with shape parameters $r_j = \{1,1,1\}$ and rate parameters $\lambda_j = \{47/2, 23, 45/2\}$.

Let $X \sim \Gamma(r, \lambda)$. Then the p.d.f. of X is

$$f_X(x) = \frac{\lambda^r}{\Gamma(r)} e^{-\lambda x} x^{r-1} \qquad (x > 0).$$

Let $X_j \sim \Gamma(r_j, \lambda_j)$ $(j = 1, \ldots, p)$ be a set of p independent r.v.'s and consider the r.v.

$$W = \sum_{j=1}^{p} X_j .$$

In case all the $r_j \in \mathbb{N}$, the distribution of W is what we call a GIG distribution (Coelho, C. A. (1998). The Generalized Integer Gamma Distribution – a Basis for Distributions in Multivariate Statistics. *Journal of Multivariate Analysis*, **64**, 86-102.).
If all the λ_j are different, W has a GIG distribution of depth p, with shape parameters r_j and rate parameters λ_j, with p.d.f.

$$f_W(w) = f^{GIG}\left(w \,\middle|\, \{r_j\}_{j=1:p}; \{\lambda_j\}_{j=1:p}; p\right) = K \sum_{j=1}^{p} P_j(w)\, e^{-\lambda_j w} ,$$

and c.d.f.

$$F_W(w) = F^{GIG}\left(w \,\middle|\, \{r_j\}_{j=1:p}; \{\lambda_j\}_{j=1:p}; p\right) = 1 - K \sum_{j=1}^{p} P_j^*(w)\, e^{-\lambda_j w} ,$$

for $w > 0$, where

$$K = \prod_{j=1}^{p} \lambda_j^{r_j}, \qquad P_j(w) = \sum_{k=1}^{r_j} c_{j,k}\, w^{k-1}$$

and

$$P_j^*(w) = \sum_{k=1}^{r_j} c_{j,k}(k-1)! \sum_{i=0}^{k-1} \frac{w^i}{i!\, \lambda_j^{k-i}},$$

with

$$c_{j,r_j} = \frac{1}{(r_j-1)!} \prod_{\substack{i=1\\ i\neq j}}^{p} (\lambda_i - \lambda_j)^{-r_i}, \qquad j = 1,\ldots,p,$$

and, for $k = 1,\ldots,r_j - 1$ and $j = 1,\ldots,p$,

$$c_{j,r_j-k} = \frac{1}{k} \sum_{i=1}^{k} \frac{(r_j-k+i-1)!}{(r_j-k-1)!} R(i,j,p)\, c_{j,r_j-(k-i)},$$

where

$$R(i,j,p) = \sum_{\substack{k=1\\ k\neq j}}^{p} r_k(\lambda_j - \lambda_k)^{-i} \qquad (i = 1,\ldots,r_j - 1).$$

The r.v. $Z = e^{-W}$ has then what is called an Exponentiated Generalized Integer Gamma (EGIG) distribution of depth g (Arnold, B. C., Coelho, C. A., Marques, F. J. (2013). The distribution of the product of powers of independent Uniform random variables, *Journal of Multivariate Analysis*, **113**, 19-36),
with p.d.f.

$$\begin{aligned} f_Z(z) &= f^{EGIG}\left(z \,\middle|\, \{r_j\}_{j=1:g}; \{\lambda_j\}_{j=1:g}; g\right) \\ &= f^{GIG}\left(-\log z \,\middle|\, \{r_j\}_{j=1:g}; \{\lambda_j\}_{j=1:g}; g\right) \frac{1}{z} \\ &= K \sum_{j=1}^{g} P_j(-\log z)\, z^{\lambda_j - 1} \qquad (0 < z < 1) \end{aligned}$$

and c.d.f.

$$\begin{aligned} F_Z(z) &= F^{EGIG}\left(z \,\middle|\, \{r_j\}_{j=1:g}; \{\lambda_j\}_{j=1:g}; g\right) \\ &= 1 - F^{GIG}\left(-\log z \,\middle|\, \{r_j\}_{j=1:g}; \{\lambda_j\}_{j=1:g}; g\right) \\ &= K \sum_{j=1}^{g} P_j^*(-\log z)\, z^{\lambda_j - 1} \qquad (0 < z < 1)\,. \end{aligned}$$

We can easily compute the c.d.f. of this distribution with the R function **GIGcdf** (which may be found in the file `GIGcdf.txt`)

This function needs a package called `Rmpfr` for R to work in extended precision.

Then, a command like

$>$ **1-GIGcdf(c(1,1,1),c(47/2,23,45/2),-log(0.6690471))**

will give the desired result:

0.005142764

Let us now consider the test to the 3 points 7, 18, and 32.

Now the distribution of Λ is that of (for $p=3$)

$$Beta\left(\frac{50-4}{2},\frac{3}{2}\right)\,Beta\left(\frac{50-5}{2},\frac{3}{2}\right)\,Beta\left(\frac{50-6}{2},\frac{3}{2}\right).$$

and c.f. of $W=-\log\Lambda$, is

$$\Phi_W(t)=\frac{\Gamma\left(\frac{46}{2}+\frac{3}{2}\right)\Gamma\left(\frac{46}{2}-\mathrm{i}t\right)}{\Gamma\left(\frac{46}{2}\right)\Gamma\left(\frac{46}{2}+\frac{3}{2}-\mathrm{i}t\right)}\,\frac{\Gamma\left(\frac{45}{2}+\frac{3}{2}\right)\Gamma\left(\frac{45}{2}-\mathrm{i}t\right)}{\Gamma\left(\frac{45}{2}\right)\Gamma\left(\frac{45}{2}+\frac{3}{2}-\mathrm{i}t\right)}\,\frac{\Gamma\left(\frac{44}{2}+\frac{3}{2}\right)\Gamma\left(\frac{44}{2}-\mathrm{i}t\right)}{\Gamma\left(\frac{44}{2}\right)\Gamma\left(\frac{44}{2}+\frac{3}{2}-\mathrm{i}t\right)}$$

so that to build the R function **cfw2** with the c.f. of W we will need the function **complex_gamma** from the package **hypergeo**, so that we have to install this package. And, because of precision issues, and issues related with the maximum value allowed in R for the Gamma functions we will have to program this function on the **log** scale, as

```
> cfw2 <- function(t)
{exp(lgamma(49/2)-lgamma(23)+complex_gamma(23-1i*t,log=T)
 -complex_gamma(49/2-1i*t,log=T)+lgamma(24)-lgamma(45/2)
 +complex_gamma(45/2-1i*t,log=T)-complex_gamma(24-1i*t,log=T)
 +lgamma(47/2)-lgamma(22)+complex_gamma(22-1i*t,log=T)
 -complex_gamma(47/2-1i*t,log=T))}
```

A test for outliers

It also happens that with c.f.'s of this type we have to use a different R function to invert the c.f.

We will use the function **FT**, which is an adaptation of the Mathematica® module with the same name provided by Abate and Valkó (2004)[1], which is made available in the file **FT.txt**.

The command
$>$ **outliers(virg,c(1,2,3),c(7,18,32))**
gives
0.5445782

so that we can obtain the desired p-value with

$>$ **1-FT(cfw2,-log(.5445782))**

which gives
0.001100808

[1] Abate, J., Valkó, P.P. (2004). Multi-precision Laplace transform inversion. *Int. J. Numer. Meth. Eng.*, 60, 979–993

Objective: Program an R function (for example called **makecflboutlier**), which given n, p and k will produce the c.f. of W as a function of t, so that it can be used in the following manner:

$>$ **cfw3** $<-$ **function(t) makecflboutlier(50,3,3,t)**
$>$ **FT(cfw3,.5445782)**
0.001100808

Although the p-value obtained is smaller than 0.01, it is still larger than $0.05/C_3^{50} = 2.55\times10^{-6}$, so that we may raise the question:
"Will there be any cases where we will reject the null hypothesis and assume that the points are outliers ?"

Let's use the data from Table 6.10 in Johnson and Wichern (2007). It is in file **trucks.dat**.

How to read this file into R?
$>$ **trucks** $<-$ **read.table("C:/.../trucks.dat",header=T)**

Let's plot in 3D, using our function:

$>$ **plot3dmine(trucks,c(1,2,3),rep(1,36),.5)**

Objective 1: to program a new function **plot3dminen** that does not plot the legend when there is a single 'group' or 'species'
Objective 2: to program a new function **plot3dminenn** that besides doing what **plot3dminen** does, also accepts that in case of a data matrix as **trucks** we can simply give the argument **index** the value 0 (zero), or even better, to not even have to indicate it, that is, that may be used as
$>$ **plot3dminenn(trucks,c(1,2,3),,.5)**

Anyway, let us test if observations 9 and 21 should or not be considered as outliers:

$>$ **outliers(trucks,c(1,2,3),c(9,21))**
0.2778192

$>$ **cflbnew** $<-$ **function(t) makecflboutlier(36,3,2,t)**
$>$ **1-FT(cflbnew,-log(.2778192))**
3.024119e-07

It happens that $0.05/C_2^{36} = 7.94{\times}10^{-5}$ and $0.01/C_2^{36} = 1.59{\times}10^{-5}$, so that in this case we should really reject H_0 and consider these two points as outliers (see Coelho and Arnold (2019, Sec. 5.1.13) for more on similar tests).

Evaluation:
(i) In this case we can simplify the c.f. of W and use the R function **cf2DistGP**.
(ii) Obtain other datasets and perform outlier tests.
(iii) Use situations where you can simplify the c.f.'s
(iv) Transform functions **FT** and **makecflboutlier** so that you can use m.g.f.'s instead.
(v) Program a function 'similar' to the function **makecflboutlier** which uses the product in ℓ instead of the product in j (slide 9)

CANONICAL CORRELATIONS ANALYSIS

Subsection 5.1.9.1 from

Coelho, C. A., Arnold, B. C. (2019). Finite Form Representations for Meijer G and Fox H Functions – Applied to Multivariate likelihood Ratio Tests using Mathematica, Maxima and R.
Lecture Notes in Statistics. Springer.

5.1.9.1 Equivalence Between the Test of Independence of $\underline{X}_1$ and $\underline{X}_2$ and the Test of Nullity of the Canonical Correlations

It is possible to prove that to test H_0 in (5.140) or (5.141) is equivalent to testing the simultaneous nullity of the so-called canonical correlations between the random vectors $\underline{X}_1$ and $\underline{X}_2$. In fact there is a matrix $U = bdiag(U_1', U_2')$ such that

$$\Sigma^* = U\Sigma U' = \begin{bmatrix} I_k & R \\ R & I_k \end{bmatrix}$$

where $k = \min(p_1, p_2)$ and

$$R = diag(\rho_1, \dots, \rho_k)\,, \quad k = \min(p_1, p_2)$$

where $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_k$ are the population canonical correlations between $\underline{X}_1$ and $\underline{X}_2$ (see Muirhead (2005, Sect. 11.3.2)).

Then, to test H_0 in (5.140) or (5.141) is equivalent to testing

$$H_0 : R = 0 \iff H_0 : \rho_1 = \cdots = \rho_k = 0\,, \tag{5.154}$$

that is, to test the independence of $\underline{X}_1$ and $\underline{X}_2$ is equivalent to test the simultaneous nullity of all population canonical correlations.

Matrices U_1 and U_2 are respectively defined as the matrices of eigenvectors of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ and $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ associated with the non-null eigenvalues of these two matrices. More precisely, the matrix

$$U_1 = \left[\, \underline{u}_{11} \,\middle|\, \underline{u}_{12} \,\middle|\, \dots \,\middle|\, \underline{u}_{1k} \,\right]$$

is the matrix with column vectors $\underline{u}_{1\alpha}$ ($\alpha = 1, \dots, k$) which are the eigenvectors of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, associated with the eigenvalues or latent roots ρ_α^2 ($\alpha = 1, \dots, k$), which are the population canonical correlations between $\underline{X}_1$ and $\underline{X}_2$ and are normalized to yield $||\underline{u}_{1\alpha}||^2_{\Sigma_{11}} = \underline{u}_{1\alpha}'\Sigma_{11}\underline{u}_{1\alpha} = 1$ (where $||\underline{u}_{1\alpha}||^2_{\Sigma_{11}}$ represents the square norm of $\underline{u}_{1\alpha}$ with respect to the metric Σ_{11}), and matrix

$$U_2 = \left[\, \underline{u}_{21} \,\middle|\, \underline{u}_{22} \,\middle|\, \dots \,\middle|\, \underline{u}_{2k} \,\right]$$

is the matrix with column vectors $\underline{u}_{2\alpha}$ ($\alpha = 1, \dots, k$) which are the eigenvectors of $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, associated with the non-null eigenvalues ρ_α^2 ($\alpha = 1, \dots, k$) and are normalized to yield $||\underline{u}_{2\alpha}||^2_{\Sigma_{22}} = \underline{u}_{2\alpha}'\Sigma_{22}\underline{u}_{2\alpha} = 1$. This implies that

$$U_1'\Sigma_{11}U_1 = I_k\,, \quad U_2'\Sigma_{22}U_2 = I_k \quad \text{and} \quad U_1'\Sigma_{12}U_2 = R\,.$$

For details see Appendix 9.

We may see that while Σ is the variance-covariance matrix of the vector $\underline{X} = [\underline{X}_1', \underline{X}_2']'$, with

$$\Sigma = Var\left(\left[\frac{\underline{X}_1}{\underline{X}_2}\right]\right) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where

$$Var(\underline{X}_1) = \Sigma_{11}\,, \quad Var(\underline{X}_2) = \Sigma_{22} \quad \text{and} \quad Cov(\underline{X}_1, \underline{X}_2) = \Sigma_{12}\,,$$

Σ^* is the variance-covariance matrix of the vector $\underline{Y} = [\underline{Y}_1', \underline{Y}_2']'$, where $\underline{Y}_1 = U_1'\underline{X}_1$ and $\underline{Y}_2 = U_2'\underline{X}_2$, with

$$\Sigma^* = Var\left(\left[\frac{\underline{Y}_1}{\underline{Y}_2}\right]\right) = \begin{bmatrix} I_k & R \\ R & I_k \end{bmatrix}, \tag{5.155}$$

where

$$\begin{aligned} Var(\underline{Y}_1) &= U_1'\,Var(\underline{X}_1)U_1 = U_1'\Sigma_{11}U_1 = I_k\,, \\ Var(\underline{Y}_2) &= U_2'\,Var(\underline{X}_2)U_2 = U_2'\Sigma_{22}U_2 = I_k \end{aligned}$$

and

$$Cov(\underline{Y}_1, \underline{Y}_2) = U_1'\,Cov(\underline{X}_1, \underline{X}_2)U_2 = U_1'\Sigma_{12}U_2 = R\,. \tag{5.156}$$

The vectors $\underline{Y}_1$ and $\underline{Y}_2$ are the vectors of canonical variables, with

$$\underline{Y}_1 = \left[Y_{11}, \ldots, Y_{1k}\right]' \quad \text{and} \quad \underline{Y}_2 = \left[Y_{21}, \ldots, Y_{2k}\right]',$$

where from (5.155) and (5.156) it is clear that each pair $(Y_{1\alpha}, Y_{2,\alpha})$ has correlation equal to ρ_α, with the first pair (Y_{11}, Y_{21}) maximizing $Corr(Y_{11}, Y_{21}) = Cov(Y_{11}, Y_{21}) = \rho_1$, the second pair (Y_{12}, Y_{22}) maximizing $Corr(Y_{12}, Y_{22}) = Cov(Y_{12}, Y_{22}) = \rho_2$, subject to the conditions

$$Cov(Y_{11}, Y_{12}) = 0 \quad \text{and} \quad Cov(Y_{21}, Y_{22}) = 0\,,$$

the third pair (Y_{13}, Y_{23}) maximizing $Corr(Y_{13}, Y_{23}) = Cov(Y_{13}, Y_{23}) = \rho_3$, subject to the conditions

$$Cov(Y_{11}, Y_{12}) = Cov(Y_{11}, Y_{13}) = Cov(Y_{12}, Y_{13}) = 0$$

and

$$Cov(Y_{21}, Y_{22}) = Cov(Y_{21}, Y_{23}) = Cov(Y_{22}, Y_{23}) = 0\,,$$

and so on, for $\alpha = 1, \ldots, \min(p_1, p_2)$, so that we may indeed write

$$Cov(\underline{Y}_1, \underline{Y}_2) = Corr(\underline{Y}_1, \underline{Y}_2) = R\,,$$

which helps clarify why $\rho_1, \ldots, \rho_k$ $(k = \min(p_1, p_2))$ are called the canonical correlations between $\underline{X}_1$ and $\underline{X}_2$.

As Muirhead (2005, Sect. 11.3.2) remarks, Σ^* is called the canonical form of Σ for the group of transformations $\Sigma \longrightarrow C\Sigma C'$ with $C = bdiag(C_1, C_2)$ where C_1 and C_2 are full-rank matrices.

We may also see that we have

$$|\Sigma^*| = |I_k|\,|I_k - R^{*\prime} I_k^{-1} R| = \prod_{\alpha=1}^{k} (1 - \rho_\alpha^2)\,.$$

There is a parallel relation between the statistic Λ and the sample canonical correlations. These are defined as the sample counterparts of the population canonical correlations ρ_α, that is, as the quantities $\widehat{\rho}_\alpha$, which are the positive square roots of the non-null $k = \min(p_1, p_2)$ eigenvalues of either one of the two matrices

$$A_{11}^{-1/2} A_{12} A_{22}^{-1} A_{21} A_{11}^{-1/2} \quad \text{or} \quad A_{22}^{-1/2} A_{21} A_{11}^{-1} A_{12} A_{22}^{-1/2} \tag{5.157}$$

which are the same eigenvalues as those of the matrices

$$A_{11}^{-1} A_{12} A_{22}^{-1} A_{21} \quad \text{and} \quad A_{22}^{-1} A_{21} A_{11}^{-1} A_{12}\,, \tag{5.158}$$

since if for example $\underline{\hat{u}}_{1\alpha}^*$ are the unitary eigenvectors of the first matrix in (5.157), associated with the eigenvalues $\widehat{\rho}_\alpha^2$, then $\underline{\hat{u}}_{1\alpha} = A_{11}^{-1/2} \underline{\hat{u}}_{1\alpha}^*$ will be the eigenvectors of the first matrix in (5.158), associated with the same eigenvalues, which may be easily verified by left multiplying

$$A_{11}^{-1/2} A_{12} A_{22}^{-1} A_{21} A_{11}^{-1/2} \underline{\hat{u}}_{1\alpha}^* = \widehat{\rho}_\alpha^2 \underline{\hat{u}}_{1\alpha}^*$$

by $A_{11}^{-1/2}$. The existence of the matrices $A_{11}^{-1/2}$ and $A_{22}^{-1/2}$ is assured by the fact that both A_{11} and A_{22} are symmetric positive-definite matrices.

But then from (5.147) we may write

$$\Lambda = |A_{11}|^{-1}\,|A_{11} - A_{12} A_{22}^{-1} A_{21}| = |I_{p_1} - A_{11}^{-1} A_{12} A_{22}^{-1} A_{21}| = \prod_{\alpha=1}^{k} (1 - \widehat{\rho}_\alpha^2)$$

while from (5.148) we may write a similar expression by switching the indexes 1 and 2, or instead, from (5.145) write

$$
\begin{aligned}
\Lambda &= \left| \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{bmatrix} \right| = \begin{vmatrix} I_{p_1} & A_{12}A_{22}^{-1} \\ A_{21}A_{11}^{-1} & I_{p_2} \end{vmatrix} \\
&= |I_{p_1}|\,|I_{p_2} - A_{21}A_{11}^{-1}I_{p_2}^{-1}A_{12}A_{22}^{-1}| = \prod_{\alpha=1}^{k}(1-\widehat{\rho}_\alpha^2)\,.
\end{aligned}
$$

We may note that the set of ordered eigenvalues (from larger to smaller) of the matrix

$$
\left[\begin{array}{c|c} I_{p_1} & A_{12}A_{22}^{-1} \\ \hline A_{21}A_{11}^{-1} & I_{p_2} \end{array}\right]
$$

are the values

$$
\lambda_j = \begin{cases} 1+\widehat{\rho}_j\,, & j=1,\dots,\min(p_1,p_2) \\ 1\,, & j=1+\min(p_1,p_2),\dots,\max(p_1,p_2) \\ 1-\widehat{\rho}_{p_1+p_2-j+1}\,, & j=1+\max(p_1,p_2),\dots,p_1+p_2 \end{cases} \tag{5.159}
$$

yielding once again

$$
\Lambda = \prod_{j=1}^{p_1+p_2}\lambda_j = \prod_{\alpha=1}^{k}(1-\widehat{\rho}_\alpha^2)\,.
$$

We may also notice the interesting parallel between the definition of the eigenvalues λ_j in (5.159) and that of the parameters h_j in (5.153), which may alternatively be written as

$$
h_j = \begin{cases} 1\,, & j=1,\dots,\min(p_1,p_2) \\ 0\,, & j=1+\min(p_1,p_2),\dots,\max(p_1,p_2) \\ -1\,, & j=1+\max(p_1,p_2),\dots,p_1+p_2-2\,. \end{cases}
$$

The equivalence of the test of the null hypotheses in (5.140) or (5.141) and the test of the null hypothesis in (5.154) is the reason why the test of independence of two sets of variables is commonly known as the test of fit of the Canonical Analysis or Canonical Correlation Analysis model, which may be seen as an all-embracing linear model since as Kshirsagar (1972, Sect. 7.8) states “Most of the practical problems arising in statistics can be translated, in some form or the other, as the problems of measurement of association between two vector variates $\underline{x}$ and

$\underline{y}$," and also as Knapp (1978) remarks, "virtually all of the commonly encountered parametric tests of significance can be treated as special cases of canonical-correlation analysis, which is the general procedure for investigating the relationship between two sets of variables," where we would only remark that "parametric tests" should refer to those parametric tests that can be translated in terms of a linear model. This is the reason why we will dedicate detailed attention to the present test.

Tests for Canonical Correlations

Let k be the number of canonical correlations between two groups of variables, the first one with p_1 variables and the second one with p_2 variables. Then, to test the hypothesis

$$H_{0(s)} : \rho_1 \neq 0, \ldots, \rho_s \neq 0, \rho_{s+1} = \ldots = \rho_k = 0\,, \qquad s = 0, \ldots, k-1$$

we will use the statistic

$$\Lambda_{(s)} = \prod_{\alpha=s+1}^{k} (1 - \hat{\rho}_\alpha^2) \tag{1}$$

where, under $H_{0(s)}$,

$$\Lambda_{(s)} \overset{st}{\sim} \prod_{j=1}^{p_1-s} Y_j \overset{st}{\sim} \prod_{\ell=1}^{p_2-s} Y_\ell^*$$

with

$$Y_j \sim Beta\left(\frac{(n-s)-(p_2-s)-j}{2}, \frac{p_2-s}{2}\right), \qquad j = 1, \ldots, p_1 - s$$

and

$$Y_\ell^* \sim Beta\left(\frac{(n-s)-(p_1-s)-j}{2}, \frac{p_1-s}{2}\right), \qquad j = 1, \ldots, p_2 - s$$

form two groups of independent r.v.'s.

It is possible to show that

$$\Lambda_{(s)} = \frac{|A|}{|A+B|}$$

where, under $H_{0(s)}$, A and B are independent, with

$$A \sim W_{p_1-s}\left(n-1-s-(p_2-s), \Sigma_{11}\right) \quad \text{and} \quad B \sim W_{p_1-s}\left(p_2-s, \Sigma_{11}\right)$$

or

$$A \sim W_{p_2-s}\left(n-1-s-(p_1-s), \Sigma_{22}\right) \quad \text{and} \quad B \sim W_{p_2-s}\left(p_1-s, \Sigma_{22}\right).$$

For $s = 0$ the test to $H_{0(s)}$ is indeed the test to the independence of the two sets of variables.

We may use for $\Lambda_{(s)}$ Rao's F approximation (Rao(1951), Rao(1973, pp. 556), Mardia, Kent & Bibby (1979 pp. 94,95)), which states that if

$$A \sim W_p(t-q, \Sigma) \quad \text{and} \quad B \sim W_p(q, \Sigma)$$

are two independent matrices, and

$$\Lambda = \frac{|A|}{|A+B|}\,,$$

then

$$\frac{ms-2\lambda}{pq}\,\frac{1-\Lambda^{1/s}}{\Lambda^{1/s}} \overset{a}{\sim} F_{pq,ms-2\lambda}$$

where

$$m = t - \frac{1}{2}(p+q+1)\,, \qquad \lambda = \frac{1}{4}(pq-2)$$

and

$$s^2 = \frac{p^2q^2-4}{p^2+q^2-5}\,.$$

Homework:

– Program an R function to obtain the canonical correlations and the asymptotic p-value from the computed value of a statistic $\Lambda_{(s)}$, using Rao's F approximation. Implement the test on a dataset of your choice.

– Obtain, for the test of nullity of all the canonical correlations, from the exact c.f. of $W = -\log \Lambda_{(0)}$, using an R function similar to tt makecflboutliers and the function `FT`.

– Adapt Rao's F approximation to the other tests where you use statistics with a distribution somewhat similar to $\Lambda_{(s)}$ (that is, the test for outliers, and the LRT test for equality of mean vectors).

Mardia, K.V., Kent, J.T., Bibby, J.M. (1979). Multivariate Analysis. Academic Press, New York.

Rao. C. R. (1951), An asymptotic expansion of the distribution of Wilks' criterion . Bull. Inst. Intern. Statist., 33, 177–180.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications. Wiley, New York.

The test to Canonical Correlations as the test of fit of the Multivariate Regression Model

As noted before, the test to the null hypothesis

$$H_0\colon \rho_1 = \ldots = \rho_k = 0 \tag{2}$$

is the test of independence of the two normal distributed vectors $\underline{X}_1$ (of dimension p_1) and $\underline{X}_2$ (of dimension p_2), where

$$\underset{(p\times 1)}{\underline{X}} = [\underline{X}_1', \underline{X}_2']' \sim N_p(\underline{\mu}, \Sigma)$$

where $p = p_1 + p_2$,

$$\underline{\mu} = \left[\underline{\mu}_1', \underline{\mu}_2'\right]'$$

and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

The test to H_0 in (2) is equivalent to the test to the hypotheses

$$\begin{array}{c} H_0\colon \Sigma_{12} = 0 \\ vs. \\ H_1\colon \Sigma_{12} \neq 0\,, \end{array} \tag{3}$$

and as such it is also the test of fit of the Multivariate Linear Model

$$\underline{X}_1 = \beta\,\underline{X}_2 + \underline{\mathcal{E}} \tag{4}$$

where, $\mathcal{E} \sim N_{p_1}(\underline{0}, \Sigma_{11.2})$, β is the $p_1{\times}p_2$ parameter matrix, and, without any loss of generality, we take $\underline{X}_1$ as the set of response variables and $\underline{X}_2$ as the set of explanatory variables.

If we reject H_0 in (2) or (3), we will say that the model in (4) fits.

The model in (4) covers several common Multivariate Linear Models:

- if $\underline{X}_2$ is a vector of continuous (normal distributed) r.v.'s, the model is a Multivariate Regression model
- if $\underline{X}_2$ is a vector of indicator variables, the model is a MANOVA (Multivariate Analysis of Variance) model, or a Discriminant Analysis model
- if $\underline{X}_2$ is a mixed vector of continuous and indicator variables, we have a MANCOVA (Multivariate Analysis of Covariance) model

The statistic used to test H_0 in (2) or (3) is the statistic $\Lambda_{(0)}$ in (1), which may be alternatively written as

$$\Lambda \equiv \Lambda_{(0)} = \frac{|A|}{|A_{11}|\,|A_{22}|} = \frac{|A_{11.2}|}{|A_{11.2}+A_{12}A_{22}^{-1}A_{21}|} = \frac{|A_{22.1}|}{|A_{22.1}+A_{21}A_{11}^{-1}A_{12}|}\,, \tag{5}$$

where A is the MLE of Σ and $\mathcal{A}_{11}$ and A_{22} its diagonal blocks of dimensions $p_1{\times}p_1$ and $p_2{\times}p_2$, and A_{12} its off-diagonal block of dimensions $p_1{\times}p_2$, with (under H_0 in (2) or (3))

$$A_{11.2} \sim W_{p_1}\left(n-1-p_2, \frac{1}{n}\Sigma_{11}\right) \quad \text{and} \quad A_{12}A_{22}^{-1}A_{21} \sim W_{p_1}\left(p_2, \frac{1}{n}\Sigma_{11}\right),$$

$$A_{22.1} \sim W_{p_2}\left(n-1-p_1, \frac{1}{n}\Sigma_{22}\right) \quad \text{and} \quad A_{21}A_{11}^{-1}A_{12} \sim W_{p_2}\left(p_1, \frac{1}{n}\Sigma_{22}\right),$$

and

$$\Lambda \overset{st}{\sim} \prod_{j=1}^{p_1} Y_j \overset{st}{\sim} \prod_{\ell=1}^{p_2} Y_\ell^*$$

where

$$Y_j \sim Beta\left(\frac{n-p_2-j}{2}, \frac{p_2}{2}\right), \qquad j=1,\dots,p_1$$

and

$$Y_\ell^* \sim Beta\left(\frac{n-p_1-j}{2}, \frac{p_1}{2}\right), \qquad j=1,\dots,p_2$$

form two groups of independent r.v.'s.

We can associate the model in (4) to H_1 in (2) and (3), writing

$$\begin{array}{ll} & H_0\colon \underline{X}_1 = \underline{\mathcal{E}} \\ vs. & \\ & H_1\colon \underline{X}_1 = \beta\,\underline{X}_2 + \underline{\mathcal{E}}\,, \end{array}$$

with a rejection of H_0 implying that we would go with the model in H_1, that is, the model in (4).

We may also implement tests between a given model, taken as the original model, and one of its submodels (that is, a similar model, with the same response variables, but with a set of explanatory variables that is a subset of the set of explanatory variables in the original model), by splitting $\underline{X}_2$ into to subvectors, $\underline{X}_{21}$, with p_{21} variables, and $\underline{X}_{22}$, with p_{22} variables (with $p_2 = p_{21} + p_{22}$), and then taking the submodel as the model

$$\underline{X}_1 = \beta^*\,\underline{X}_{22} + \underline{\mathcal{E}}^*\,,$$

where β^* is now of dimensions $p_1{\times}p_{22}$. If we attach this submodel to a null hypothesis H_0^*, writing

$$H_0^*\colon \underline{X}_1 = \beta^*\,\underline{X}_{22} + \underline{\mathcal{E}}^*\,, \tag{6}$$

then it is possible to show that a test between the original model in H_1 and this submodel in H_0^* may be done by using the statistic

$$\Lambda^* = \frac{\Lambda_{H_0,H_1}}{\Lambda_{H_0^*,H_1}}$$

where Λ_{H_0,H_1} is the LRT statistic used to test between H_0 and H_1, that is, the LRT statistic in (5), and $\Lambda_{H_0^*,H_1}$ is the LRT statistic to test between H_0^* and H_1, that is, the statistic used to test the fit of the submodel in (6).

The statistic Λ^* is indeed the LRT statistic to test between H_0^* and H_1.

It is possible to prove that Λ^* has a distribution similar to that of Λ in (5), with n replaced by $n - p_{21}$ and p_2 replaced by $p_2 - p_{21} = p_{22}$, that is, that

$$\Lambda^* \stackrel{st}{\sim} \prod_{j=1}^{p_1} Y_j^* \stackrel{st}{\sim} \prod_{\ell=1}^{p_{22}} Y_\ell^{**}$$

where

$$Y_j^* \sim Beta\left(\frac{n-p_2-j}{2}, \frac{p_{22}}{2}\right), \qquad j = 1, \ldots, p_1$$

and

$$Y_\ell^{**} \sim Beta\left(\frac{n-p_1-p_{21}-j}{2}, \frac{p_1}{2}\right), \qquad j = 1, \ldots, p_{22}$$

form two groups of independent r.v.'s.

In carrying out this test we will attach the submodel to the null hypothesis of the test and the original model to the alternative hypothesis. A rejection of the null hypothesis will mean that the original model fits significantly better than the submodel, and we would thus, in this case, stay with the original model, and vice-versa.

Homework:

- Program an R function to obtain the computed values of the statistics Λ and Λ^*.

- Program an R function to implement Rao's F approximation for statistics Λ and Λ^*.

- Find a dataset to implement tests of fit for different types of Multivariate Linear models

- Implement tests between models and submodels that would be interesting and meaningful, using your dataset.

Estatística Multivariada
Principal Components Analysis
2023/2024

Carlos A. Coelho
(cmac@fct.unl.pt)

Mathematics Department
NOVA School of Science and Technology

Let

$$\Sigma = Var(\underline{X})\,.$$

(It may be $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, but not necessarily.)

Then let

$$\Sigma \underline{u}_j = \lambda_j \underline{u}_j \quad (j = 1, \ldots, p)\,,$$

that is, let

$$\Sigma U = U\lambda \iff U'\Sigma U = \Lambda = diag(\lambda_1, \ldots, \lambda_p)$$

with $\lambda_1 \geq \cdots \geq \lambda_p$, and

$$U = \left[\underline{u}_1 | \ldots | \underline{u}_p\right]$$

orthogonal, that is, with $U'U = UU' = I_p$ and $|U| = \pm 1$.

Then let

$$\underline{W} = U'\underline{X} = \left[W_1, \ldots, W_p\right]'\,.$$

Then, $W_1, \ldots, W_p$ are the Principal Components of $\underline{X}$, with

$$Var(\underline{W}) = Var(U'\underline{X}) = U'\,Var(\underline{X})\,U = \Lambda\,,$$

namely with

$$Var(W_j) = Var(\underline{u}_j'\underline{X}) = \underline{u}_j'\,Var(\underline{X})\,\underline{u}_j = \underline{u}_j'\Sigma\underline{u}_j = \lambda_j\,,$$

and

$$Cov(W_j, W_{j'}) = Cov(\underline{u}_j'\underline{X}, \underline{u}_{j'}'\underline{X}) = \underline{u}_j'\Sigma\underline{u}_{j'} = 0\,.$$

W_1 – is the 1st Principal Component of $\underline{X}$, it is the (normalized) linear combination of the variables $X_1, \ldots, X_p$, with largest variance $(= \lambda_1)$

W_2 – is the 2nd Principal Component of $\underline{X}$, it is the (normalized) linear combination of the variables $X_1, \ldots, X_p$, orthogonal to W_1, with largest variance $(= \lambda_2)$

... – and so on,

with

$$\lambda_k = \max_{\substack{\underline{v}'\underline{v}=1 \\ \underline{v}'\underline{u}_i=0;i=1,\ldots,k-1}} \underline{v}'\Sigma\underline{v} = \underline{u}_k'\Sigma\underline{u}_k \,.$$

Note that

$$tr(\Sigma) = tr(\Sigma I_p) = tr(\Sigma U U') = tr(U'\Sigma U) = tr(\Lambda)$$

and

$$|\Sigma| = |U'|\,|\Sigma|\,|U| = |U'\Sigma U| = |\Lambda| \,.$$

Let

$$\widehat{\Sigma} = \frac{1}{n}S = \frac{1}{n}X'\left(I_n - \frac{1}{n}E_{nn}\right)X$$

and let

$$\ell_1 \geq \cdots \geq \ell_p$$

be the eigenvalues of $\widehat{\Sigma}$ (or otherwise of the sample variance-covariance matrix), that is, let

$$\widehat{\Sigma}\underline{q}_j = \ell_j\underline{q}_j \quad (j = 1, \ldots, p),$$

or

$$Q'\widehat{\Sigma}Q = L = diag(\ell_1, \ldots, \ell_p)$$

with

$$Q = \left[\underline{q}_1 | \ldots | \underline{q}_p\right].$$

The sample Principal Components are then

$$\widehat{W}_j = \underline{q}_j'\underline{X} \quad (j = 1, \ldots, p)$$

with

$$\underline{\widehat{W}} = Q'\underline{X} = \left[\widehat{W}_1, \ldots, \widehat{W}_p\right].$$

Question: how many (sample) Principal components should we consider?

Homework: Implement a Principal Components Analysis on a dataset of your choice, arguing how many (sample) Principal Components should be considered.

Estatística Multivariada

Tests for eigenvalues of covariance matrices

2023/2024

Carlos A. Coelho
(cmac@fct.unl.pt)

Mathematics Department
NOVA School of Science and Technology

The setup:
Let us consider a sample of size n from a p-variate Normal population $N_p(\mu, \Sigma)$, and let S be the sample covariance matrix. Then $S \sim W_p\left(n-1, \frac{1}{n-1}\Sigma\right)$
Let ℓ_i and λ_i be the i-th eigenvalue, respectively of S and Σ, with $\lambda_1 > \lambda_2 > \cdots > \lambda_p$.

I) Distribution of estimators of eigenvalues (towards CI's and tests for eigenvalues):
Then, from Theorem 13.5.1 in Anderson (2003), we have that

$$\sqrt{n-1}(\ell_i - \lambda_i) \xrightarrow[n\to\infty]{} N(0, 2\lambda_i^2) . \qquad (1)$$

We may see that ℓ_i is a consistent estimator of λ_i, since from (1) we have

$$\ell_i - \lambda_i \overset{a}{\sim} N\left(0, \frac{2\lambda_i^2}{n-1}\right) \Longleftrightarrow \ell_i \overset{a}{\sim} N\left(\lambda_i, \frac{2\lambda_i^2}{n-1}\right),$$

so that

$$\lim_{n\to\infty} MSE(\ell_i) = \lim_{n\to\infty} Var(\ell_i) + \lim_{n\to\infty} (b(\ell_i))^2 = \lim_{n\to\infty} \frac{2\lambda_i^2}{n-1} + \lim_{n\to\infty} 0 = 0 .$$

Also, from (1) we have

$$\frac{\sqrt{n-1}(\ell_i-\lambda_i)}{\sqrt{2}\,\lambda_i} \xrightarrow[n\to\infty]{} N(0,1)\,, \tag{2}$$

and as such

$$\begin{aligned}
1-\alpha &\approx P\left(-z_{1-\alpha/2} \le \frac{\sqrt{n-1}(\ell_i-\lambda_i)}{\sqrt{2}\,\lambda_i} \le z_{1-\alpha/2}\right) \\
&= P\left(-z_{1-\alpha/2}\frac{\sqrt{2}\,\lambda_i}{\sqrt{n-1}} \le \ell_i-\lambda_i \le z_{1-\alpha/2}\frac{\sqrt{2}\,\lambda_i}{\sqrt{n-1}}\right) \\
&= P\left(-\ell_i-z_{1-\alpha/2}\frac{\sqrt{2}\,\lambda_i}{\sqrt{n-1}} \le -\lambda_i \le -\ell_i+z_{1-\alpha/2}\frac{\sqrt{2}\,\lambda_i}{\sqrt{n-1}}\right) \\
&= P\left(\ell_i-z_{1-\alpha/2}\frac{\sqrt{2}\,\lambda_i}{\sqrt{n-1}} \le \lambda_i \le \ell_i+z_{1-\alpha/2}\frac{\sqrt{2}\,\lambda_i}{\sqrt{n-1}}\right) \\
&\approx P\left(\ell_i-z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}} \le \lambda_i \le \ell_i+z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right) \qquad \text{(by consistency of } \ell_i\text{)}
\end{aligned} \tag{3}$$

so that

$$\left[\ell_i-z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\,,\,\ell_i+z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right] \tag{4}$$

is an approximate bilateral asymptotic CI (confidence interval) for λ_i, corresponding to a probability of $1-\alpha$.

Using a similar technique we can also obtain unilateral CI's for λ_i, thus obtaining lower and upper bounds for λ_i. For example, from

$$\begin{aligned}
1-\alpha &= P\left(\frac{\sqrt{n-1}(\ell_i-\lambda_i)}{\sqrt{2}\,\lambda_i} \leq z_{1-\alpha}\right)\\
&= P\left(\ell_i-\lambda_i \leq z_{1-\alpha}\frac{\sqrt{2}\,\lambda_i}{\sqrt{n-1}}\right)\\
&= P\left(-\lambda_i \leq -\ell_i + z_{1-\alpha}\frac{\sqrt{2}\,\lambda_i}{\sqrt{n-1}}\right)\\
&= P\left(\lambda_i \geq \ell_i - z_{1-\alpha}\frac{\sqrt{2}\,\lambda_i}{\sqrt{n-1}}\right)\\
&\approx P\left(\lambda_i \geq \ell_i - z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right) \qquad \text{(by consistency of } \ell_i\text{)}
\end{aligned}$$

so that we have

$$\left[\ell_i - z_{1-\alpha}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}, +\infty\right[\tag{5}$$

as an approximate (left) unilateral asymptotic CI for λ_i, which gives an approximate lower bound for λ_i.

We might had also obtained

$$\left] \cancel{-\infty}\; 0, \ell_i + z_{1-\alpha} \frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}} \right] \qquad (6)$$

as an approximate (right) unilateral asymptotic CI for λ_i, with $\ell_i + z_{1-\alpha} \frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}$ as an approximate upper bound for λ_i.

Note that we could also have worked out from (3) in a slightly different way, as

$$\begin{aligned}
1-\alpha &\approx P\left(-z_{1-\alpha/2}\sqrt{\frac{2}{n-1}} \le \frac{\ell_i-\lambda_i}{\lambda_i} \le z_{1-\alpha/2}\sqrt{\frac{2}{n-1}}\right)\\
&= P\left(1-z_{1-\alpha/2}\sqrt{\frac{2}{n-1}} \le \frac{\ell_i}{\lambda_i} \le 1+z_{1-\alpha/2}\sqrt{\frac{2}{n-1}}\right)\\
&= P\left(\frac{1}{1-z_{1-\alpha/2}\sqrt{\frac{2}{n-1}}} \ge \frac{\lambda_i}{\ell_i} \ge \frac{1}{1+z_{1-\alpha/2}\sqrt{\frac{2}{n-1}}}\right)\\
&= P\left(\frac{\ell_i}{1+z_{1-\alpha/2}\sqrt{\frac{2}{n-1}}} \le \lambda_i \le \frac{\ell_i}{1-z_{1-\alpha/2}\sqrt{\frac{2}{n-1}}}\right)
\end{aligned}$$

to obtain

$$\left[\frac{\ell_i}{1+z_{1-\alpha/2}\sqrt{\frac{2}{n-1}}}, \frac{\ell_i}{1-z_{1-\alpha/2}\sqrt{\frac{2}{n-1}}}\right] \qquad (7)$$

as an asymptotic bilateral CI for λ_i.

Then, if one wants to test the hypotheses

$$\begin{array}{c} H_0 : \lambda_i = \lambda_0 \\ \text{vs.} \\ H_1 : \lambda_i \neq \lambda_0 \end{array} \tag{8}$$

one will reject H_0 if λ_0 falls out of the CI's in (4) or (7), and if one wants to test the hypotheses

$$\begin{array}{c} H_0 : \lambda_i \geq \lambda_0 \\ \text{vs.} \\ H_1 : \lambda_i < \lambda_0 \end{array} \tag{9}$$

one will reject H_0 if λ_0 falls out of the CI in (6) or out of a CI like

$$\left]\cancel{-\infty}\; 0, \frac{\ell_i}{1-z_{1-\alpha}\sqrt{\frac{2}{n-1}}}\right] \tag{10}$$

and if one wants to test the hypotheses

$$\begin{array}{c} H_0 : \lambda_i \leq \lambda_0 \\ \text{vs.} \\ H_1 : \lambda_i > \lambda_0 \end{array} \tag{11}$$

one will reject H_0 if λ_0 falls out of the CI in (5) or out of a CI like

$$\left[\frac{\ell_i}{1+z_{1-\alpha}\sqrt{\frac{2}{n-1}}}, +\infty\right[. \tag{12}$$

Tests for (the value of) eigenvalues

While the decision of rejecting the null hypothesis in (8) if λ_0 falls outside the CI's in (4) or (7) is almost intuitive, we should formalize this, because for the other two sets of hypothesis, the decision may be not as intuitive as that, at least at first sight.

In carrying out the test, using the first approach in building the CIs, we will use, from (2), the statistic

$$Z = \frac{\sqrt{n-1}(\ell_i - \lambda_0)}{\sqrt{2}\,\ell_i} \overset{a}{\sim} N(0,1)\,, \tag{13}$$

and would reject H_0 in (8) if $|Z| > z_{1-\alpha/2}$, leading to

$$\begin{aligned}
\alpha &\approx P\left[Z < -z_{1-\alpha/2} \vee Z > z_{1-\alpha/2}\right] \\
&= P\left[\ell_i - \lambda_0 < -z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}} \vee \ell_i - \lambda_0 > z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right] \\
&= P\left[-\lambda_0 < -\ell_i - z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}} \vee -\lambda_0 > -\ell_i + z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right] \\
&= P\left[\lambda_0 > \ell_i + z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}} \vee \lambda_0 < \ell_i - z_{1-\alpha/2}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right]
\end{aligned}$$

and correspondingly for the second approach in building the CIs, where we would use the statistic

$$Z = \frac{\sqrt{n-1}(\ell_i - \lambda_0)}{\sqrt{2}\,\lambda_0} \overset{a}{\sim} N(0,1)\,.$$

Concerning the one-sided tests and CIs

- we would reject the null hypothesis in (9), if $Z < -z_{1-\alpha}$, corresponding to reject H_0 if λ_0 falls out of the CI in (6), since, for the first approach in building the CIs and for the statistic in (13) we would have

$$\begin{aligned}\alpha &\approx P\left[Z < -z_{1-\alpha}\right] = P\left[\ell_i - \lambda_0 < -z_{1-\alpha}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right]\\ &= P\left[-\lambda_0 < -\ell_i - z_{1-\alpha}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right] = P\left[\lambda_0 > \ell_i + z_{1-\alpha}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right]\end{aligned}$$

- and we would reject the null hypothesis in (11), if $Z > z_{1-\alpha}$, corresponding to reject H_0 if λ_0 falls out of the CI in (5), since, for the first approach in building the CIs and for the statistic in (13) we would have

$$\begin{aligned}\alpha &\approx P\left[Z > z_{1-\alpha}\right] = P\left[\ell_i - \lambda_0 > z_{1-\alpha}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right]\\ &= P\left[-\lambda_0 > -\ell_i + z_{1-\alpha}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right] = P\left[\lambda_0 < \ell_i - z_{1-\alpha}\frac{\sqrt{2}\,\ell_i}{\sqrt{n-1}}\right].\end{aligned}$$

Homework:

1 – obtain the CI's in (10) and (12).

2 – i) Program two R functions, such that given ℓ_i, n (the sample size) and the value of α will give the CIs for λ_i built using both approaches shown above (with option of being a bilateral CI or a unilateral CI, with default the bilateral CI).

ii) Obtain at least one dataset onto which you compute Confidence Intervals and carry out tests for the eigenvalues of the population covariance matrix.

II) A test for equality of eigenvalues:
Let L_k be a set of $k \leq p$ indexes.
We are interested in testing the hypotheses

$$\begin{array}{c} H_0 : \lambda_i = \lambda_{i'}\,,\ \text{for all } i,i' \in L_k \\ \text{vs.} \\ H_1 : \lambda_i \neq \lambda_{i'}\,,\ \text{for some } i,i' \in L_k\,. \end{array} \tag{14}$$

The test statistic (the $2/n$-th power of the Likelihood Ratio Test statistic) is (Anderson, 1963, Corollary 1; Mardia, Kent and Bibby, 1979, Sec. 8.4.3)

$$\Lambda = \frac{\prod_{j \in L_k} \ell_j}{\left(\frac{1}{k}\sum_{j \in L_K} \ell_j\right)^k} \tag{15}$$

and we will reject H_0 in (14) for too small values of Λ, that is, we will reject H_0 in (14) if

$$\Lambda_{calc} < \lambda_\alpha\,,$$

where λ_α is the α-quantile of Λ.

For a sample of size n, the distribution of Λ is the same as that of

$$\prod_{j=2}^{k} Y_j \quad \text{with} \quad Y_j \sim Beta\left(\frac{n-j}{2}, \frac{j-1}{k} + \frac{j-1}{2}\right)$$

where the Y_j form a set of $k-1$ independent r.v.'s.

Notes:

(i) Note that we need to have $n > k$ in order to be able to run the test.

(ii) For $k = p$ this is the so-called sphericity test, which is the test to the hypotheses

$$\begin{array}{l} H_0 : \Sigma = \sigma^2 I_p \\ \text{vs.} \\ H_1 : \Sigma \neq \sigma^2 I_p \ \ (\text{although being positive-definite}) . \end{array}$$

Homework:

(i) to program an R function to obtain the computed value of Λ from a data set

(ii) to program an R function to build the c.f. of $W = -\log\Lambda$ to be used with the **FT** inversion function.

(iii) carry out the test of equality of the last 2, and the last 3 eigenvalues of the cov matrix of the iris 3 species, and the virg species.

(iv) obtain at least one dataset onto which you carry out tests of equality for the eigenvalues of the covariance matrix.

In case we do not reject H_0 in (14), let

$$\bar{\ell} = \frac{1}{k} \sum_{j \in L_k} \ell_j$$

and let λ^* be the common unknown (population) value of the k eigenvalues λ_j (for $j \in L_k$). Then (Anderson, 1963, expr. (3.10))

$$\sqrt{n-1}\,(\bar{\ell} - \lambda^*) \xrightarrow[n \to \infty]{} N\left(0, 2(\lambda^*)^2/k\right).$$

Homework:

i) Based on this result obtain CI's and tests for λ^*.

ii) Program an R function to implement the CI's and compute them for a dataset of your choice, implementing also the corresponding tests.

But, if the smallest k eigenvalues of Σ are assumed different (Anderson, 1963, expr. 3.15 and after), then if we take

$$\overline{\ell}_k = \frac{1}{k} \sum_{j=p-k+1}^{p} \ell_j \quad \text{and} \quad \lambda_k^* = \frac{1}{k} \sum_{j=p-k+1}^{p} \lambda_j \,,$$

we have (for a sample of size n),

$$\sqrt{n-1}(\overline{\ell}_k - \lambda_k^*) \xrightarrow[n\to\infty]{} N\left(0, 2\textstyle\sum_{j=p-k+1}^{p} \lambda_j^2/k^2\right).$$

Homework:

i) Based on this result and using the consistent estimator of the variance $2\sum_{j=p-k+1}^{p} \ell_j^2/k^2$, obtain CI's and tests for λ_k^*.

ii) Program an R function to implement the CI's and compute them on a dataset of your choice. Implement also tests for λ_k^*.

Let
$$\Psi = (\lambda_1 + \cdots + \lambda_k)/(\lambda_1 + \cdots + \lambda_p)\,,$$
and let
$$\widehat{\Psi} = (\ell_1 + \cdots + \ell_k)/(\ell_1 + \cdots + \ell_p)\,.$$
Then (Mardia, Kent and Bibby (1979, sec. 8.4.2), for a sample of size n,
$$\widehat{\Psi} \overset{a}{\sim} N(\Psi, \tau^2)$$
where
$$\tau^2 = \frac{2Tr(\Sigma^2)}{(n-1)(Tr(\Sigma))^2}\left(\Psi^2 - 2\alpha\Psi + \alpha\right),$$
with
$$\alpha = (\lambda_1^2 + \cdots + \lambda_k^2)/(\lambda_1^2 + \cdots + \lambda_p^2)\,.$$
In τ^2 we will estimate Ψ by $\widehat{\Psi}$ and Σ by either the sample covariance matrix or the m.l.e. of Σ, and α by
$$\widehat{\alpha} = (\ell_1^2 + \cdots + \ell_k^2)/(\ell_1^2 + \cdots + \ell_p^2)\,.$$

Homework:

(i) using the above approach, obtain CI's and tests for Ψ;

(ii) program R functions to compute such CI's and to carry out such tests;

(iii) implement tests for proportions of eigenvalues on datasets of your choice.