

Likelihood Ratio Tests in Multivariate Analysis

whose statistics have quite simple
finite form representations for their distributions

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

Mathematics Department, NOVA School of Science and Technology
Center for Mathematics and Applications (NovaMath)
NOVA University of Lisbon



REPÚBLICA
PORTUGUESA

NOVAMATH
CENTER FOR MATHEMATICS
& APPLICATIONS



NOVA
NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

FCT Fundação
para a Ciência
e a Tecnologia

**NOVA
idFCT**
Associação para a Inovação
e Desenvolvimento da FCT

This work is funded by national funds through the FCT – Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications)

65th Annual Conference
South African Statistical Association
SASA 2024 – Stellenbosch

- Some preliminary results
- Products of independent Beta r.v.'s with finite form representations for their distributions^(*)
-  functions available to read files
-  functions available to handle each LRT
- Several LRTs (Likelihood Ratio Tests)^(*)
 - The LRT for Equality of Mean Vectors (real r.v.'s)
 - The LRT for Equality of Mean Vectors (complex r.v.'s)
 - The LRT for simultaneous Nullity of Mean Vectors (real r.v.'s)
 - The LRT for simultaneous Nullity of Mean Vectors (complex r.v.'s)
 - The LRT for Profile Parallelism (real r.v.'s)
 - The LRT for Profile Parallelism (complex r.v.'s)
 - The LRT for Independence of 2 sets of variables
 - The LRT between a Multivariate Linear model and a submodel

^(*)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, Cham, Switzerland, 515+xviii pp. (ISBN: 978-3-030-28789-4) (<https://doi.org/10.1007/978-3-030-28790-0>)

- The LRT for Independence of several sets of variables
- The LRT for the Outlier test for real r.v.'s
- The LRT for the Outlier test for complex r.v.'s
- The LRT for Complete Symmetrical Equivalence (for real r.v.'s)
- The LRT for Equality of Mean Vectors, with circular cov. matrices
- The LRT for circularity of the covariance matrix
- The LRT for the simultaneous test of circularity of the covariance matrix and equality of means
- The LRT for the simultaneous test of independence of several sets of variables, the circularity of the covariance matrices and the equality of the means in each set

Materials available at <https://github.com/Carlos-Coelho/SASA-2024>

About the validity of the LRTs addressed for non-normal distributions

Although all the LRTs that we will address and the distributions of the associated statistics are derived under Multivariate Normality assumptions,

- the distributions of the LRT statistics for the LRTs addressed remain valid if the distribution of \underline{X} is not multivariate Normal, but some elliptically contoured or left orthogonal-invariant distribution
 - see Theorem 5.3.3 in Fang and Zhang (1990, Chap. V) and Sect. 9.11 in Anderson (2003), and also Jensen and Good (1981), Kariya (1981), Anderson et al. (1986), and Anderson and Fang (1990).

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, New York.

Anderson, T.W., Fang, K.-T. (1990). Inference in multivariate elliptically contoured distributions based on maximum likelihood. In: Fang, K.-T., Anderson, T.W. (eds) *Statistical Inference in Elliptically Contoured and Related Distributions*, pp. 201-216. Allerton Press, Inc., New York.

Anderson, T.W., Fang, K.-T., Hsu, H. (1986). Maximum-likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions. *Can. J. Stat.* 14, 55-59.

Fang, K.T., Zhang, Y.-T. (1990). *Generalized Multivariate Analysis*. Springer, New York.

Jensen, D.R., Good, I.J. (1981). Invariant distributions associated with matrix laws under structural symmetry. *J. R. Stat. Soc., Ser. B* 43, 327-332.

Kariya, T. (1981). Robustness of multivariate tests. *Ann. Stat.* 9, 1267-1275.

Most of the LRT (Likelihood Ratio Test) statistics used in Multivariate Analysis have distributions which are the same as that of a product of independent Beta r.v.'s.

Such distributions are, in general, considered to be non-manageable.

Most often the p.d.f.'s and c.d.f.'s of these statistics are represented using (infinite) series representations or special functions as the Meijer G function.

But these representations present serious problems in terms of computations, not only in terms of precision but also in terms of computation time.

Most of the LRT (Likelihood Ratio Test) statistics used in Multivariate Analysis have distributions which are the same as that of a product of independent Beta r.v.'s.

Such distributions are, in general, considered to be non-manageable.

Most often the p.d.f.'s and c.d.f.'s of these statistics are represented using (infinite) series representations or special functions as the Meijer G function.

But these representations present serious problems in terms of computations, not only in terms of precision but also in terms of computation time.

Most of the LRT (Likelihood Ratio Test) statistics used in Multivariate Analysis have distributions which are the same as that of a product of independent Beta r.v.'s.

Such distributions are, in general, considered to be non-manageable.

Most often the p.d.f.'s and c.d.f.'s of these statistics are represented using (infinite) series representations or special functions as the Meijer G function.

But these representations present serious problems in terms of computations, not only in terms of precision but also in terms of computation time.

Most of the LRT (Likelihood Ratio Test) statistics used in Multivariate Analysis have distributions which are the same as that of a product of independent Beta r.v.'s.

Such distributions are, in general, considered to be non-manageable.

Most often the p.d.f.'s and c.d.f.'s of these statistics are represented using (infinite) series representations or special functions as the Meijer G function.

But these representations present serious problems in terms of computations, not only in terms of precision but also in terms of computation time.

But, a not so well known fact is that in many cases, or quite often in most cases, we may get closed finite and manageable form representations for their p.d.f.'s and c.d.f.'s.

We are going to see how!

But, a not so well known fact is that in many cases, or quite often in most cases, we may get closed finite and manageable form representations for their p.d.f.'s and c.d.f.'s.

We are going to see how!

Preliminary result:

Since for any Beta r.v. with an integer 2nd parameter we can write the distribution of its negative logarithm as that of a sum of independent Exponential r.v.'s

We may write the distribution of the negative logarithm of a product of independent Beta r.v.'s whose 2nd parameters are integer as that of the sum of independent Gamma r.v.'s with integer shape parameters

(given the fact that the sum of independent Exponential r.v.'s all with the same rate parameter is a Gamma distributed r.v. with that same rate parameter and a shape parameter equal to the number of Exponential r.v.'s being added)

Preliminary result:

Since for any Beta r.v. with an integer 2nd parameter we can write the distribution of its negative logarithm as that of a sum of independent Exponential r.v.'s

We may write the distribution of the negative logarithm of a product of independent Beta r.v.'s whose 2nd parameters are integer as that of the sum of independent Gamma r.v.'s with integer shape parameters

(given the fact that the sum of independent Exponential r.v.'s all with the same rate parameter is a Gamma distributed r.v. with that same rate parameter and a shape parameter equal to the number of Exponential r.v.'s being added)

Preliminary result:

Since for any Beta r.v. with an integer 2nd parameter we can write the distribution of its negative logarithm as that of a sum of independent Exponential r.v.'s

We may write the distribution of the negative logarithm of a product of independent Beta r.v.'s whose 2nd parameters are integer as that of the sum of independent Gamma r.v.'s with integer shape parameters

(given the fact that the sum of independent Exponential r.v.'s all with the same rate parameter is a Gamma distributed r.v. with that same rate parameter and a shape parameter equal to the number of Exponential r.v.'s being added)

Background for this preliminary result:

Let us remember that, if $X \sim \text{Exp}(\lambda)$, then

$$\begin{aligned}M_X(t) = E(e^{tX}) &= \int_0^{+\infty} e^{tx} \lambda e^{-\lambda x} dx \\&= \lambda \int_0^{+\infty} e^{-(\lambda-t)x} dx \\&= \lambda (\lambda - t)^{-1} \underbrace{\int_0^{+\infty} \underbrace{(\lambda - t) e^{-(\lambda-t)x}}_{\substack{\text{pdf of } \text{Exp}(\lambda-t) \\ \text{for } \lambda > t}} dx}_{=1} \\&= \lambda (\lambda - t)^{-1} \quad (\text{for } t < \lambda)\end{aligned}$$

or

$$\Phi_X(t) = E(e^{itX}) = \lambda (\lambda - it)^{-1}$$

and that, if $X \sim \Gamma(r, \lambda)$, then

$$\begin{aligned}
 M_X(t) = E(e^{tX}) &= \int_0^{+\infty} e^{tx} \frac{\lambda^r}{\Gamma(r)} e^{-\lambda x} x^{r-1} dx \\
 &= \lambda^r (\lambda - t)^{-r} \int_0^{+\infty} \frac{(\lambda - t)^r}{\Gamma(r)} e^{-(\lambda - t)x} x^{r-1} dx \\
 &= \lambda^r (\lambda - t)^{-r} \underbrace{\int_0^{+\infty} \underbrace{\frac{(\lambda - t)^r}{\Gamma(r)} e^{-(\lambda - t)x} x^{r-1}}_{\text{pdf of } \Gamma(r, \lambda - t)} dx}_{\substack{\text{for } \lambda > t \\ = 1}} \\
 &= \lambda^r (\lambda - t)^{-r} \quad (\text{for } t < \lambda)
 \end{aligned}$$

or

$$\Phi_X(t) = E(e^{itX}) = \lambda^r (\lambda - it)^{-r},$$

which is the c.f. of a Gamma distribution with rate parameter λ and integer shape parameter r .

Moreover, if $X_i \sim \text{Exp}(\lambda)$, $i = 1, \dots, r$, are r independent r.v.'s and $S = \sum_{i=1}^r X_i$, then $S \sim \Gamma(r, \lambda)$ since then

$$\begin{aligned} M_S(t) &= E(e^{tS}) = E\left(e^{t\sum_{i=1}^r X_i}\right) = E\left(\prod_{i=1}^r e^{tX_i}\right) \\ &= \prod_{i=1}^r E(e^{tX_i}) = \prod_{i=1}^r \lambda(\lambda - t)^{-1} = \lambda^r(\lambda - t)^{-r}. \end{aligned}$$

Then, on one hand we also know that if $X \sim \text{Beta}(a, b)$, then

$$\begin{aligned} E(X^h) &= \int_0^1 x^h \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1} dx \\ &= \frac{1}{B(a, b)} \underbrace{\int_0^1 x^{a+h-1} (1-x)^{b-1} dx}_{=B(a+h, b)} = \frac{B(a+h, b)}{B(a, b)} \end{aligned}$$

where

$$B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}$$

so that

$$E(X^h) = \frac{B(a+h, b)}{B(a, b)} = \frac{\Gamma(a+h) \cancel{\Gamma(b)}}{\Gamma(a+b+h)} \frac{\Gamma(a+b)}{\Gamma(a) \cancel{\Gamma(b)}} = \frac{\Gamma(a+b) \Gamma(a+h)}{\Gamma(a+b+h) \Gamma(a)}$$

for any $a+h > 0 \iff h > -a$.

while, on the other hand,

$$\Gamma(r+1) = r\Gamma(r), \quad (\text{for any } r \in \mathbb{C})$$

from where we may write (by applying it repeatedly), for any $r \in \mathbb{C}$ and any $n \in \mathbb{N}$,

$$\frac{\Gamma(r+n)}{\Gamma(r)} = \prod_{\ell=0}^{n-1} (r+\ell).$$

So that we may write the c.f. of $W = -\log X$, for $b \in \mathbb{N}$, as

$$\begin{aligned}\Phi_W(t) &= E(e^{itW}) = E(e^{-it \log X}) = E(X^{-it}) \\ &= \frac{\Gamma(a+b)\Gamma(a-it)}{\Gamma(a+b-it)\Gamma(a)} = \frac{\Gamma(a+b)\Gamma(a-it)}{\Gamma(a)\Gamma(a+b-it)} \\ &= \prod_{\ell=0}^{b-1} (a+\ell)(a+\ell-it)^{-1},\end{aligned}$$

which is the c.f. of a sum of independent Exponential r.v.'s with rate parameters $a+\ell$ ($\ell = 0, \dots, b-1$).

And, as such,

So that we may write the c.f. of $W = -\log X$, for $b \in \mathbb{N}$, as

$$\begin{aligned}\Phi_W(t) &= E(e^{itW}) = E(e^{-it \log X}) = E(X^{-it}) \\ &= \frac{\Gamma(a+b)\Gamma(a-it)}{\Gamma(a+b-it)\Gamma(a)} = \frac{\Gamma(a+b)\Gamma(a-it)}{\Gamma(a)\Gamma(a+b-it)} \\ &= \prod_{\ell=0}^{b-1} (a+\ell)(a+\ell-it)^{-1},\end{aligned}$$

which is the c.f. of a sum of independent Exponential r.v.'s with rate parameters $a+\ell$ ($\ell = 0, \dots, b-1$).

And, as such,

if $X_j \sim \text{Beta}(a_j, b_j)$, ($j = 1, \dots, p$) are p independent r.v.'s, with $b_j \in \mathbb{N}$ ($j = 1, \dots, p$), and $\Lambda = \prod_{j=1}^p X_j$, and we take $W = -\log \Lambda$, then

$$\begin{aligned}\Phi_W(t) &= E(e^{itW}) = E(e^{-it \log \Lambda}) = E(\Lambda^{-it}) \\ &= E\left[\left(\prod_{j=1}^p X_j\right)^{-it}\right] = E\left(\prod_{j=1}^p X_j^{-it}\right) = \prod_{j=1}^p E(X_j^{-it}) \\ &= \prod_{j=1}^p \frac{\Gamma(a_j + b_j) \Gamma(a_j - it)}{\Gamma(a_j) \Gamma(a_j + b_j - it)} = \prod_{j=1}^p \prod_{\ell=0}^{b_j-1} (a_j + \ell) (a_j + \ell - it)^{-1},\end{aligned}$$

which is the c.f. of $\sum_{j=1}^p b_j$ independent Exponential r.v.'s with rate parameters $a_j + \ell$ ($\ell = 0, \dots, b_j - 1$; $j = 1, \dots, p$), which, in case some of the $a_j + \ell$ parameters are equal, is the c.f. of a sum of independent Gamma r.v.'s with integer shape parameters.

And as such, indeed,

We may write the distribution of the negative logarithm of a product of independent Beta r.v.'s whose 2nd parameters are integer as that of the *sum of independent Gamma r.v.'s with integer shape parameters*

and this distribution is a GIG (Generalized Integer Gamma) distribution.

We say that the r.v. X has a Gamma distribution with shape parameter $r (> 0)$ and rate parameter $\lambda (> 0)$, and we will denote this fact by $X \sim \Gamma(r, \lambda)$, if the p.d.f. of X is

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

Let $X_j \sim \Gamma(r_j, \lambda_j)$ ($j = 1, \dots, p$) 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.

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

and c.d.f.

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

for $w > 0$, where

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

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

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

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

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

where

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

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.

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

and c.d.f.

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

But, what happens when the second parameters of the Beta r.v.'s are not integer?

Very much happily, when dealing with the distributions of LRT statistics used in Multivariate Analysis, the second parameters of the Beta r.v.'s involved are in general rational.

And for these cases we have a number of results that show that we can still obtain the p.d.f.'s and c.d.f.'s of their distributions in finite closed forms, through the use of the GIG and EGIG distributions.

But, what happens when the second parameters of the Beta r.v.'s are not integer?

Very much happily, when dealing with the distributions of LRT statistics used in Multivariate Analysis, the second parameters of the Beta r.v.'s involved are in general rational.

And for these cases we have a number of results that show that we can still obtain the p.d.f.'s and c.d.f.'s of their distributions in finite closed forms, through the use of the GIG and EGIG distributions.

But, what happens when the second parameters of the Beta r.v.'s are not integer?

Very much happily, when dealing with the distributions of LRT statistics used in Multivariate Analysis, the second parameters of the Beta r.v.'s involved are in general rational.

And for these cases we have a number of results that show that we can still obtain the p.d.f.'s and c.d.f.'s of their distributions in finite closed forms, through the use of the GIG and EGIG distributions.

Theorem 1 (Theorem 3.1 in Coelho and Arnold (2019))

For positive integers $n_v, k_{v\ell}$, and $m_{v\ell}$ ($v = 1, \dots, m^*; \ell = 1, \dots, n_v$), and for real $a_v > \sum_{\ell=1}^{n_v} k_{v\ell} / \min_{\ell} \{k_{v\ell}\}$ ($v = 1, \dots, m^*$), let

$$Z = \prod_{v=1}^{m^*} \prod_{\ell=1}^{n_v} \prod_{j=1}^{k_{v\ell}} Y_{v\ell j}$$

where

$$Y_{v\ell j} \sim \text{Beta} \left(a_v - \frac{j + \sum_{r=1}^{\ell-1} k_{vr}}{k_{v\ell}}, \frac{m_{v\ell}}{k_{v\ell}} \right), \quad v = 1, \dots, m^*; \ell = 1, \dots, n_v; j = 1, \dots, k_{v\ell},$$

are independent r.v.'s. Then, for $W = -\log Z$,

$$W \stackrel{d}{=} \sum_{v=1}^{m^*} \sum_{\ell=1}^{n_v} \sum_{j=0}^{m_{v\ell}-1} W_{v\ell j}$$

where “ $\stackrel{d}{=}$ ” stands for “is equivalent in distribution to”, or “is stochastically equivalent to”, and

$$W_{v\ell j} \sim \text{Exp} \left(a_v + \frac{j - \sum_{r=1}^{\ell} k_{vr}}{k_{v\ell}} \right)$$

and where the Exponential r.v.'s are all independent.

Corollary 1 (Corollary 4.1 in Coelho and Arnold (2019))

For Z in Theorem 1 we may write, for $0 < z \leq 1$, the p.d.f. of Z as,

$$f_Z(z) = f^{EGIG} \left(z \left| \left\{ \left\{ 1 \right\}_{\substack{v=1:m^* \\ \ell=1:n_v \\ j=0:m_{v\ell}-1}} \right\} \right. \right); \left\{ \left\{ a_v + \frac{j - \sum_{r=1}^{\ell} k_{vr}}{k_{v\ell}} \right\}_{\substack{v=1:m^* \\ \ell=1:n_v \\ j=0:m_{v\ell}-1}} \right\}; g \leq \sum_{v=1}^{m^*} \sum_{\ell=1}^{n_v} m_{v\ell} \right)$$

and the c.d.f. as

$$F_Z(z) = f^{EGIG} \left(z \left| \left\{ \left\{ 1 \right\}_{\substack{v=1:m^* \\ \ell=1:n_v \\ j=0:m_{v\ell}-1}} \right\} \right. \right); \left\{ \left\{ a_v + \frac{j - \sum_{r=1}^{\ell} k_{vr}}{k_{v\ell}} \right\}_{\substack{v=1:m^* \\ \ell=1:n_v \\ j=0:m_{v\ell}-1}} \right\}; g \leq \sum_{v=1}^{m^*} \sum_{\ell=1}^{n_v} m_{v\ell} \right),$$

where we use the notation $\widetilde{\{a_{vj\ell}\}}$ denotes the “contraction” of the set of rate parameters, that is, the set of unique $a_{vj\ell}$ values, and $\widetilde{\{1\}_{vj\ell}}$ denotes the corresponding “contraction” of the corresponding set of shape parameters, that is, the number of times the rate parameter $a_{vj\ell}$ appears.

Theorem 2 (Theorem 3.2 in Coelho and Arnold (2019))

If in Theorem 1 we have $k_{v\ell} = k_v$ and $m_{v\ell} = m_v$, for all $\ell = 1, \dots, n_v$, we will have, for $a_v > n_v$,

$$Z = \prod_{v=1}^{m^*} \prod_{\ell=1}^{n_v} \prod_{j=1}^{k_v} Y_{v\ell j}$$

where

$$Y_{v\ell j} \sim \text{Beta} \left(a_v + 1 - \ell - \frac{j}{k_v}, \frac{m_v}{k_v} \right),$$

are independent r.v.'s, then, for $W = -\log Z$,

$$W \stackrel{d}{=} \sum_{v=1}^{m^*} \sum_{j=1}^{m_v + k_v(n_v - 1)} W_{vj},$$

where,

$$W_{vj} \sim \Gamma \left(r_{vj}, a_v - n_v + \frac{j-1}{k_v} \right),$$

with

$$r_{vj} = \begin{cases} h_{vj} & j = 1, \dots, k_v \\ h_{vj} + r_{v,j-k_v} & j = k_v + 1, \dots, m_v + k_v(n_v - 1) \end{cases}$$

for $v = 1, \dots, m^*$, and where, for $j = 1, \dots, m_v + k_v(n_v - 1)$,

$$h_{vj} = (\# \text{ of elements in } \{p_v, m_v\} \geq j) - 1, \quad v = 1, \dots, m^*,$$

for $p_v = n_v k_v$, and where all Gamma r.v.'s in the double summation are independent.

This shows that in this particular case the exact distribution of Z is an EGIG distribution of depth at most $\sum_{v=1}^{m^*} m_v + k_v (n_v - 1)$, with shape parameters r_{vj} and rate parameters $a_v - n_v + (j - 1)/k_v$, ($v = 1, \dots, m^*$; $j = 1, \dots, m_v + k_v (n_v - 1)$).

Corollary 2 (Corollary 4.2 in Coelho and Arnold (2019))

For the particular case in Theorem 2, we may write the p.d.f. of Z as

$$f_Z(z) = f^{EGIG} \left(z \left| \left\{ \left\{ r_{vj} \right\}_{\substack{v=1:m^* \\ j=1:m_v+k_v(n_v-1)}} \right\}; \left\{ \left\{ a_v - n_v + \frac{j-1}{k_v} \right\}_{\substack{v=1:m^* \\ j=1:m_v+k_v(n_v-1)}} \right\}; \right. \\ \left. g \leq \sum_{v=1}^{m^*} m_v + k_v (n_v - 1) \right)$$

and the c.d.f. as

$$F_Z(z) = F^{EGIG} \left(z \left| \left\{ \left\{ r_{vj} \right\}_{\substack{v=1:m^* \\ j=1:m_v+k_v(n_v-1)}} \right\}; \left\{ \left\{ a_v - n_v + \frac{j-1}{k_v} \right\}_{\substack{v=1:m^* \\ j=1:m_v+k_v(n_v-1)}} \right\}; \right. \\ \left. g \leq \sum_{v=1}^{m^*} m_v + k_v (n_v - 1) \right).$$

Theorem 3 (Theorem 3.3 in Coelho and Arnold (2019))

For positive integers k_v, n_v , nonnegative integers s_v and real $a_v > n_v k_v$ ($v = 1, \dots, m^*$), let

$$Z = \prod_{v=1}^{m^*} \prod_{\ell=1}^{n_v} \prod_{j=1}^{k_v} Y_{v\ell j}$$

where, for $v = 1, \dots, m^*; \ell = 1, \dots, n_v; j = 1, \dots, k_v$,

$$Y_{v\ell j} \sim \text{Beta} \left(a_v - \frac{(\ell-1)k_v + j}{n_v}, \frac{j + (\ell-1)k_v + \ell + s_v - 1}{n_v} \right)$$

are independent r.v.'s. Then, for $W = -\log Z$,

$$W \stackrel{d}{=} \sum_{v=1}^{m^*} \sum_{\ell=1}^{n_v k_v + s_v} W_{v\ell},$$

where

$$W_{v\ell} \sim \Gamma \left(r_{v\ell}, a_v + \frac{s_v - \ell}{n_v} \right),$$

with

$$r_{v\ell} = \begin{cases} k_v & \ell = 1, \dots, s_v \\ k_v + 1 + \left\lfloor \frac{s_v - \ell}{n_v} \right\rfloor & \ell = s_v + 1, \dots, n_v k_v + s_v, \end{cases} \quad (3.18)$$

for $v = 1, \dots, m^*$, and where all the Gamma random variables involved are independent.

Corollary 3 (Corollary 4.3 in Coelho and Arnold (2019))

For Z in Theorem 3 we may write its p.d.f., for $0 < z \leq 1$, as

$$f_Z(z) = f^{EGIG} \left(z \left| \left\{ \left\{ r_{v\ell} \right\}_{\substack{v=1:m^* \\ \ell=1:n_vk_v+s_v}} \right\}; \left\{ \left\{ a_v + \frac{s_v - \ell}{n_v} \right\}_{\substack{v=1:m^* \\ \ell=1:n_vk_v+s_v}} \right\}; \right. \\ \left. g \leq \sum_{v=1}^{m^*} n_vk_v + s_v \right),$$


and its c.d.f. as

$$F_Z(z) = F^{EGIG} \left(z \left| \left\{ \left\{ r_{v\ell} \right\}_{\substack{v=1:m^* \\ \ell=1:n_vk_v+s_v}} \right\}; \left\{ \left\{ a_v + \frac{s_v - \ell}{n_v} \right\}_{\substack{v=1:m^* \\ \ell=1:n_vk_v+s_v}} \right\}; \right. \\ \left. g \leq \sum_{v=1}^{m^*} n_vk_v + s_v \right).$$

All this may seem a little too complicated, but we will see next (and soon)

- how these results may be used and implemented in practice
 - to allow us to carry out many interesting and useful Multivariate Analysis (likelihood ratio) tests.

First of all let us see the type of files that we are interested in handling (and which the R functions provided are able to handle)

We will use two  functions to read multi-sample files (files that have a multi-sample layout) or one-sample files (files that have a one-sample layout) and to generate a multi-sample layout. These are:

ReadFileR – to read real valued files

ReadFileC – to read complex valued files

and

Two functions to read multi-sample or one-sample files and to generate a one-sample layout. These are:

ReadFile1sR – to read real valued files

ReadFile1sC – to read complex valued files

Fig. 1

Examples of multi-sample layouts that work with the **ReadFileR** function

Names of variables				sample 1			
X1	X2	X3	X4	X1	X2	X3	X4
1st sample				23.4	56.7	21.2	36.5
23.4	56.7	21.2	36.5	34.5	23.4	28.9	25.4
34.5	23.4	28.9	25.4	21.2	33.3	56.7	25.6
21.2	33.3	56.7	25.6	sample 2			
2nd sample				X1	X2	X3	X4
56.8	56.4	25.6	43.3	56.8	56.4	25.6	43.3
23.4	56.8	23.4	52.2	23.4	56.8	23.4	52.2
33.5	45.8	23.6	42.2	33.5	45.8	23.6	42.2
33.7	38.9	56.8	53.2	33.7	38.9	56.8	53.2
3rd sample				sample 3			
43.2	45.6	34.5	32.2	X1	X2	X3	X4
56.4	24.7	78.3	34.3	43.2	45.6	34.5	32.2
74.5	83.5	84.3	52.3	56.4	24.7	78.3	34.3
24.5	25.7	37.4	52.3	74.5	83.5	84.3	52.3
35.7	74.5	84.6	83.3	24.5	25.7	37.4	52.3
				35.7	74.5	84.6	83.3
sample 1				23.4	56.7	21.2	36.5
23.4	56.7	21.2	36.5	34.5	23.4	28.9	25.4
34.5	23.4	28.9	25.4	21.2	33.3	56.7	25.6
21.2	33.3	56.7	25.6	sample 2			
sample 2				56.8	56.4	25.6	43.3
56.8	56.4	25.6	43.3	23.4	56.8	23.4	52.2
23.4	56.8	23.4	52.2	33.5	45.8	23.6	42.2
33.5	45.8	23.6	42.2	33.7	38.9	56.8	53.2
33.7	38.9	56.8	53.2	sample 3			
sample 3				43.2	45.6	34.5	32.2
43.2	45.6	34.5	32.2	56.4	24.7	78.3	34.3
56.4	24.7	78.3	34.3	74.5	83.5	84.3	52.3
74.5	83.5	84.3	52.3	24.5	25.7	37.4	52.3
24.5	25.7	37.4	52.3	35.7	74.5	84.6	83.3
35.7	74.5	84.6	83.3				

Fig. 2

Examples of one-sample layouts that work with the **ReadFileR** function

a)					b)			
Names of variables					Names of variables			
X1	X2	X3	X4	ind	X1	X2	X3	X4
23.4	56.7	21.2	36.5	1	23.4	56.7	21.2	36.5
34.5	23.4	28.9	25.4	1	34.5	23.4	28.9	25.4
21.2	33.3	56.7	25.6	1	21.2	33.3	56.7	25.6
56.8	56.4	25.6	43.3	2	56.8	56.4	25.6	43.3
23.4	56.8	23.4	52.2	2	23.4	56.8	23.4	52.2
33.5	45.8	23.6	42.2	2	33.5	45.8	23.6	42.2
33.7	38.9	56.8	53.2	2	33.7	38.9	56.8	53.2
43.2	45.6	34.5	32.2	3	43.2	45.6	34.5	32.2
56.4	24.7	78.3	34.3	3	56.4	24.7	78.3	34.3
74.5	83.5	84.3	52.3	3	74.5	83.5	84.3	52.3
24.5	25.7	37.4	52.3	3	24.5	25.7	37.4	52.3
35.7	74.5	84.6	83.3	3	35.7	74.5	84.6	83.3

c)					d)			
Names of variables					Names of variables			
X1	X2	X3	X4	ind	X1	X2	X3	X4
35.7	74.5	84.6	83.3	3	23.4	56.7	21.2	36.5
23.4	56.7	21.2	36.5	1	34.5	23.4	28.9	25.4
21.2	33.3	56.7	25.6	1	21.2	33.3	56.7	25.6
56.8	56.4	25.6	43.3	2	56.8	56.4	25.6	43.3
34.5	23.4	28.9	25.4	1	23.4	56.8	23.4	52.2
23.4	56.8	23.4	52.2	2	33.5	45.8	23.6	42.2
33.7	38.9	56.8	53.2	2	33.7	38.9	56.8	53.2
43.2	45.6	34.5	32.2	3	43.2	45.6	34.5	32.2
56.4	24.7	78.3	34.3	3	56.4	24.7	78.3	34.3
33.5	45.8	23.6	42.2	2	74.5	83.5	84.3	52.3
74.5	83.5	84.3	52.3	3	24.5	25.7	37.4	52.3
24.5	25.7	37.4	52.3	3	35.7	74.5	84.6	83.3

Files **ex1_1.dat** and **ex1_4.dat** have layouts as the ones in Fig. 1, 1st and 4th displays, while files **ex1_2.dat** and **ex1_3.dat** have layouts as the ones if Fig. 2 c) and d), respectively.

All them will produce exactly the same final output and layout of samples, but the way function **ReadFileR** reads them differs, according to their layouts.

Just try using a command like:

```
> ReadFileR("ex1_1.dat")  
> ReadFileR("ex1_2.dat")  
> ReadFileR("ex1_3.dat")  
or  
> ReadFileR("ex1_4.dat")
```

One other type of file that **ReadFileR** and **ReadFileC** are able to read are files as the **Iris.dat** file, which after all is of the type of the file in Fig. 2 a) or c).

Just try taking a look at the file and use

```
> ReadFileR("Iris.dat")
```

Since the first tests we will be addressing will use either **ReadFileR** or **ReadFileC**, we will address functions **ReadFile1sR** and **ReadFile1sC** only later.

Files **ex1_1.dat** and **ex1_4.dat** have layouts as the ones in Fig. 1, 1st and 4th displays, while files **ex1_2.dat** and **ex1_3.dat** have layouts as the ones if Fig. 2 c) and d), respectively.

All them will produce exactly the same final output and layout of samples, but the way function **ReadFileR** reads them differs, according to their layouts.

Just try using a command like:

```
> ReadFileR("ex1_1.dat")  
> ReadFileR("ex1_2.dat")  
> ReadFileR("ex1_3.dat")  
or  
> ReadFileR("ex1_4.dat")
```

One other type of file that **ReadFileR** and **ReadFileC** are able to read are files as the **Iris.dat** file, which after all is of the type of the file in Fig. 2 a) or c).

Just try taking a look at the file and use

```
> ReadFileR("Iris.dat")
```

Since the first tests we will be addressing will use either **ReadFileR** or **ReadFileC**, we will address functions **ReadFile1sR** and **ReadFile1sC** only later.

Files **ex1_1.dat** and **ex1_4.dat** have layouts as the ones in Fig. 1, 1st and 4th displays, while files **ex1_2.dat** and **ex1_3.dat** have layouts as the ones if Fig. 2 c) and d), respectively.

All them will produce exactly the same final output and layout of samples, but the way function **ReadFileR** reads them differs, according to their layouts.

Just try using a command like:

```
> ReadFileR("ex1_1.dat")  
> ReadFileR("ex1_2.dat")  
> ReadFileR("ex1_3.dat")  
or  
> ReadFileR("ex1_4.dat")
```

One other type of file that **ReadFileR** and **ReadFileC** are able to read are files as the **Iris.dat** file, which after all is of the type of the file in Fig. 2 a) or c).

Just try taking a look at the file and use

```
> ReadFileR("Iris.dat")
```

Since the first tests we will be addressing will use either **ReadFileR** or **ReadFileC**, we will address functions **ReadFile1sR** and **ReadFile1sC** only later.

Files **ex1_1.dat** and **ex1_4.dat** have layouts as the ones in Fig. 1, 1st and 4th displays, while files **ex1_2.dat** and **ex1_3.dat** have layouts as the ones if Fig. 2 c) and d), respectively.

All them will produce exactly the same final output and layout of samples, but the way function **ReadFileR** reads them differs, according to their layouts.

Just try using a command like:

```
> ReadFileR("ex1_1.dat")  
> ReadFileR("ex1_2.dat")  
> ReadFileR("ex1_3.dat")  
or  
> ReadFileR("ex1_4.dat")
```

One other type of file that **ReadFileR** and **ReadFileC** are able to read are files as the **Iris.dat** file, which after all is of the type of the file in Fig. 2 a) or c).

Just try taking a look at the file and use

```
> ReadFileR("Iris.dat")
```

Since the first tests we will be addressing will use either **ReadFileR** or **ReadFileC**, we will address functions **ReadFile1sR** and **ReadFile1sC** only later.

The function **EGIG_help** may be used to obtain information on each of the other functions we will be using.

Try a command like

```
> EGIG_help(ReadFileR)
```

Then try to use function `ReadFileR` with one or both indexes (`index1` and/or `index2`). That is, try for example:

```
> ReadFileR("ex1_1.dat",1)
> ReadFileR("ex1_1.dat",,1)
> ReadFileR("ex1_1.dat",1,1)
```


The function **EGIG_help** may be used to obtain information on each of the other functions we will be using.

Try a command like

```
> EGIG_help(ReadFileR)
```

Then try to use function **ReadFileR** with one or both indexes (**index1** and/or **index2**). That is, try for example:

```
> ReadFileR("ex1_1.dat",1)
> ReadFileR("ex1_1.dat",,1)
> ReadFileR("ex1_1.dat",1,1)
```

Each LRT has an associated acronym, and for each LRT, there are **seven**  functions are available:

PDF<test acronym> – used to compute values of the PDF of the test statistic

CDF<test acronym> – used to compute values of the CDF of the test statistic

PlotPDF<test acronym> – used to plot the PDF of the test statistic


PlotCDF<test acronym> – used to plot the CDF of the test statistic

Pval<test acronym> – used to compute values of the p-value of the test statistic (actually the same as the **CDF<test acronym>** function, since for LRTs we reject the null hypothesis if the computed value of the test statistic is smaller than the α -quantile)

Quant<test acronym> – used to compute quantiles of the test statistic

PvalData<test acronym> – used to obtain the computed value of the test statistic and corresponding p-value, from a data file (probably the most useful function, and the one that will be most used)

(The function **EGIG_help** may be used with any of these functions, to obtain information on their use.)

Each LRT has an associated acronym, and for each LRT, there are **seven**  functions are available:

PDF<test acronym> – used to compute values of the PDF of the test statistic

CDF<test acronym> – used to compute values of the CDF of the test statistic

PlotPDF<test acronym> – used to plot the PDF of the test statistic


PlotCDF<test acronym> – used to plot the CDF of the test statistic

Pval<test acronym> – used to compute values of the p-value of the test statistic (actually the same as the **CDF<test acronym>** function, since for LRTs we reject the null hypothesis if the computed value of the test statistic is smaller than the α -quantile)

Quant<test acronym> – used to compute quantiles of the test statistic

PvalData<test acronym> – used to obtain the computed value of the test statistic and corresponding p-value, from a data file (probably the most useful function, and the one that will be most used)

(The function **EGIG_help** may be used with any of these functions, to obtain information on their use.)

Each LRT has an associated acronym, and for each LRT, there are **seven**  functions are available:

PDF<test acronym> – used to compute values of the PDF of the test statistic

CDF<test acronym> – used to compute values of the CDF of the test statistic

PlotPDF<test acronym> – used to plot the PDF of the test statistic


PlotCDF<test acronym> – used to plot the CDF of the test statistic

Pval<test acronym> – used to compute values of the p-value of the test statistic (actually the same as the **CDF<test acronym>** function, since for LRTs we reject the null hypothesis if the computed value of the test statistic is smaller than the α -quantile)

Quant<test acronym> – used to compute quantiles of the test statistic

PvalData<test acronym> – used to obtain the computed value of the test statistic and corresponding p-value, from a data file (probably the most useful function, and the one that will be most used)

(The function **EGIG_help** may be used with any of these functions, to obtain information on their use.)

Each LRT has an associated acronym, and for each LRT, there are **seven**  functions are available:

PDF<test acronym> – used to compute values of the PDF of the test statistic

CDF<test acronym> – used to compute values of the CDF of the test statistic

PlotPDF<test acronym> – used to plot the PDF of the test statistic


PlotCDF<test acronym> – used to plot the CDF of the test statistic

Pval<test acronym> – used to compute values of the p-value of the test statistic (actually the same as the **CDF<test acronym>** function, since for LRTs we reject the null hypothesis if the computed value of the test statistic is smaller than the α -quantile)

Quant<test acronym> – used to compute quantiles of the test statistic

PvalData<test acronym> – used to obtain the computed value of the test statistic and corresponding p-value, from a data file (probably the most useful function, and the one that will be most used)

(The function **EGIG_help** may be used with any of these functions, to obtain information on their use.)

Each LRT has an associated acronym, and for each LRT, there are **seven**  functions are available:

PDF<test acronym> – used to compute values of the PDF of the test statistic

CDF<test acronym> – used to compute values of the CDF of the test statistic

PlotPDF<test acronym> – used to plot the PDF of the test statistic


PlotCDF<test acronym> – used to plot the CDF of the test statistic

Pval<test acronym> – used to compute values of the p-value of the test statistic (actually the same as the **CDF**<test acronym> function, since for LRTs we reject the null hypothesis if the computed value of the test statistic is smaller than the α -quantile)

Quant<test acronym> – used to compute quantiles of the test statistic

PvalData<test acronym> – used to obtain the computed value of the test statistic and corresponding p-value, from a data file (probably the most useful function, and the one that will be most used)

(The function **EGIG_help** may be used with any of these functions, to obtain information on their use.)

Each LRT has an associated acronym, and for each LRT, there are **seven**  functions are available:

PDF<test acronym> – used to compute values of the PDF of the test statistic

CDF<test acronym> – used to compute values of the CDF of the test statistic

PlotPDF<test acronym> – used to plot the PDF of the test statistic


PlotCDF<test acronym> – used to plot the CDF of the test statistic

Pval<test acronym> – used to compute values of the p-value of the test statistic (actually the same as the **CDF**<test acronym> function, since for LRTs we reject the null hypothesis if the computed value of the test statistic is smaller than the α -quantile)

Quant<test acronym> – used to compute quantiles of the test statistic

PvalData<test acronym> – used to obtain the computed value of the test statistic and corresponding p-value, from a data file (probably the most useful function, and the one that will be most used)

(The function **EGIG_help** may be used with any of these functions, to obtain information on their use.)

Each LRT has an associated acronym, and for each LRT, there are **seven**  functions are available:

PDF<test acronym> – used to compute values of the PDF of the test statistic

CDF<test acronym> – used to compute values of the CDF of the test statistic

PlotPDF<test acronym> – used to plot the PDF of the test statistic


PlotCDF<test acronym> – used to plot the CDF of the test statistic

Pval<test acronym> – used to compute values of the p-value of the test statistic (actually the same as the **CDF**<test acronym> function, since for LRTs we reject the null hypothesis if the computed value of the test statistic is smaller than the α -quantile)

Quant<test acronym> – used to compute quantiles of the test statistic

PvalData<test acronym> – used to obtain the computed value of the test statistic and corresponding p-value, from a data file (probably the most useful function, and the one that will be most used)

(The function **EGIG_help** may be used with any of these functions, to obtain information on their use.)

Each LRT has an associated acronym, and for each LRT, there are **seven**  functions are available:

PDF<test acronym> – used to compute values of the PDF of the test statistic

CDF<test acronym> – used to compute values of the CDF of the test statistic

PlotPDF<test acronym> – used to plot the PDF of the test statistic

PlotCDF<test acronym> – used to plot the CDF of the test statistic

Pval<test acronym> – used to compute values of the p-value of the test statistic (actually the same as the **CDF**<test acronym> function, since for LRTs we reject the null hypothesis if the computed value of the test statistic is smaller than the α -quantile)

Quant<test acronym> – used to compute quantiles of the test statistic

PvalData<test acronym> – used to obtain the computed value of the test statistic and corresponding p-value, from a data file (probably the most useful function, and the one that will be most used)

(The function **EGIG_help** may be used with any of these functions, to obtain information on their use.)

Examples of LRTs whose statistics fall under Theorems 1 or 2

- The LRT for Equality of Mean Vectors (real r.v.'s)
- The LRT for Equality of Mean Vectors (complex r.v.'s)
- The LRT for simultaneous Nullity of Mean Vectors (real r.v.'s)
- The LRT for simultaneous Nullity of Mean Vectors (complex r.v.'s)
- The LRT for Profile Parallelism (real r.v.'s)
- The LRT for Profile Parallelism (complex r.v.'s)
- The LRT for Independence of 2 sets of variables (real r.v.'s)
- The LRT for Independence of 2 sets of variables (complex r.v.'s)
- The LRT for Independence of several sets of variables (real r.v.'s)
- The LRT for Independence of several sets of variables (complex r.v.'s)

(See subsec 5.1.1 in Coelho and Arnold (2019))

Let us suppose that $\underline{X}_k \sim N_p(\underline{\mu}_k, \Sigma)$ ($k = 1, \dots, q$), and that we have q independent samples, one from each \underline{X}_k , with sizes n_k , and that we are interested in testing the null hypothesis

$$H_0 : \underline{\mu}_1 = \dots = \underline{\mu}_k = \dots = \underline{\mu}_q.$$

For $n = \sum_{k=1}^q n_k$, the $(2/n)$ -th power of the LRT statistic for the test to H_0 is

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

where

$$A = \sum_{k=1}^q (n_k - 1) * S_k \quad \text{and} \quad B = \sum_{k=1}^q n_k (\bar{\underline{X}}_k - \bar{\underline{X}}) (\bar{\underline{X}}_k - \bar{\underline{X}})'$$

are respectively the "within" and "between" sum of squares and products of deviations from the sample means, with S_k and $\bar{\underline{X}}_k$ being respectively the sample covariance matrix and mean vector of the k -th sample, and

$$\bar{\underline{X}} = \frac{1}{n} \sum_{k=1}^q n_k \bar{\underline{X}}_k \quad \text{with} \quad n = \sum_{k=1}^q n_k.$$

Then we may show that

$$\Lambda \sim \prod_{j=1}^p Y_j \sim \prod_{k=1}^{q-1} Y_k^*,$$

where, for $n > p + q - 1$,

$$Y_j \sim \text{Beta}\left(\frac{n-q+1-j}{2}, \frac{q-1}{2}\right) \quad \text{and} \quad Y_k^* \sim \text{Beta}\left(\frac{n-p-k}{2}, \frac{p}{2}\right),$$

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

(Note that we need to have $n > p + q - 1$)

As such, for even p , the exact distribution of Λ is given by Theorems 1 or 2, with

$$m^* = 1, \quad a_1 = \frac{n-q+1}{2}, \quad n_1 = \frac{p}{2}, \quad k_1 = 2, \quad \text{and} \quad m_1 = q-1,$$

and for odd q , with

$$m^* = 1, \quad a_1 = \frac{n-p}{2}, \quad n_1 = \frac{q-1}{2}, \quad k_1 = 2, \quad \text{and} \quad m_1 = p,$$

and thus the exact PDF and CDF of Λ are, for even p or odd q , given by Corollaries 1 or 2, through the PDF and CDF of the EGIG distribution, as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-3}; \left\{ \frac{n-2-j}{2} \right\}_{j=1:p+q-3}; p+q-3 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-3}; \left\{ \frac{n-2-j}{2} \right\}_{j=1:p+q-3}; p+q-3 \right).$$


where

$$r_j = \begin{cases} h_j & j = 1, 2 \\ r_{j-2} + h_j & j = 3, \dots, p+q-3, \end{cases}$$

with

$$\begin{aligned} h_j &= \begin{cases} 1, & j = 1, \dots, \min(p, q-1) \\ 0, & j = 1 + \min(p, q-1), \dots, \max(p, q-1) \\ -1, & j = 1 + \max(p, q-1), \dots, p+q-3 \end{cases} \\ &= (\# \text{ of elements in } \{p, q-1\} \geq j) - 1, \quad j = 1, \dots, p+q-3. \end{aligned}$$

Let us take the data in the file **ex1.1.dat** (3 samples, with sizes 3, 4 and 5, and 4 variables), to carry out a test of equality of the 3 population mean vectors.

(Note that in order to obtain the p -values, as well as any other value coming out of computations involving the PDF or CDF of the test statistics, you will need to have installed the package **Rmpfr**, since the computations that involve these  functions use extended precision.)

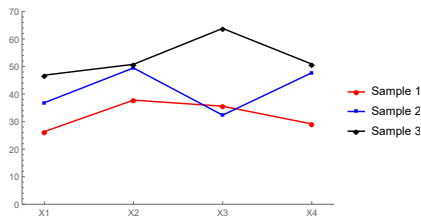
It will be enough to use a command like

```
> PvalDataEqMeanVecR("ex1.1.dat")
```

to obtain a computed value of Λ of

0.209951

and a p -value of 0.178810




Values of the sample means for each sample

So that we should not reject H_0 ! (We should reject H_0 in case the p -value is smaller than our established value of α , or, equivalently, if the computed value of Λ is smaller than the α -quantile)

Note that the samples are very small, so that we have not so much power to reject H_0 .

Let us take the data in the file **ex1.1.dat** (3 samples, with sizes 3, 4 and 5, and 4 variables), to carry out a test of equality of the 3 population mean vectors.

(Note that in order to obtain the p -values, as well as any other value coming out of computations involving the PDF or CDF of the test statistics, you will need to have installed the package **Rmpfr**, since the computations that involve these  functions use extended precision.)

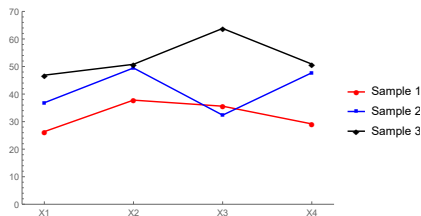
It will be enough to use a command like

```
> PvalDataEqMeanVecR("ex1.1.dat")
```

to obtain a computed value of Λ of

0.209951

and a p -value of 0.178810



Values of the sample means for each sample

So that we should not reject H_0 ! (We should reject H_0 in case the p -value is smaller than our established value of α , or, equivalently, if the computed value of Λ is smaller than the α -quantile)

Note that the samples are very small, so that we have not so much power to reject H_0 .

We can check with the function **CDFEqMeanVecR** or the function **PvalEqMeanVecR** the p-value, with a command like

```
> CDFEqMeanVecR(12,4,3,0.209951)
```

or

```
> PvalEqMeanVecR(12,4,3,0.209951)
```

which will give a p-value of 0.1788105357.

For more precision, one can use

```
> CDFEqMeanVecR(12,4,3,0.2099507759)
```

or

```
> PvalEqMeanVecR(12,4,3,0.2099507759)
```

to obtain a p-value of 0.1788101398.

The reverse way, one can obtain the 0.1788101398-quantile of Λ with a command as

```
> QuantEqMeanVecR(12,4,3,0.1788101398)
```

which will give 0.20995077590, confirming the results from **PvalDataEqMeanVecR**.

To obtain the 0.05 and 0.01 quantiles of Λ one may use

```
> QuantEqMeanVecR(12,4,3,0.05)
```

and

```
> QuantEqMeanVecR(12,4,3,0.01)
```

We can check with the function **CDFEqMeanVecR** or the function **PvalEqMeanVecR** the p-value, with a command like

```
> CDFEqMeanVecR(12,4,3,0.209951)
```

or

```
> PvalEqMeanVecR(12,4,3,0.209951)
```

which will give a p-value of 0.1788105357.

For more precision, one can use

```
> CDFEqMeanVecR(12,4,3,0.2099507759)
```

or

```
> PvalEqMeanVecR(12,4,3,0.2099507759)
```

to obtain a p-value of 0.1788101398.

The reverse way, one can obtain the 0.1788101398-quantile of Λ with a command as

```
> QuantEqMeanVecR(12,4,3,0.1788101398)
```

which will give 0.20995077590, confirming the results from **PvalDataEqMeanVecR**.

To obtain the 0.05 and 0.01 quantiles of Λ one may use

```
> QuantEqMeanVecR(12,4,3,0.05)
```

and

```
> QuantEqMeanVecR(12,4,3,0.01)
```

We can check with the function **CDFEqMeanVecR** or the function **PvalEqMeanVecR** the p-value, with a command like

```
> CDFEqMeanVecR(12,4,3,0.209951)
```

or

```
> PvalEqMeanVecR(12,4,3,0.209951)
```

which will give a p-value of 0.1788105357.

For more precision, one can use

```
> CDFEqMeanVecR(12,4,3,0.2099507759)
```

or

```
> PvalEqMeanVecR(12,4,3,0.2099507759)
```

to obtain a p-value of 0.1788101398.

The reverse way, one can obtain the 0.1788101398-quantile of Λ with a command as

```
> QuantEqMeanVecR(12,4,3,0.1788101398)
```

which will give 0.20995077590, confirming the results from **PvalDataEqMeanVecR**.

To obtain the 0.05 and 0.01 quantiles of Λ one may use

```
> QuantEqMeanVecR(12,4,3,0.05)
```

and

```
> QuantEqMeanVecR(12,4,3,0.01)
```

We can check with the function **CDFEqMeanVecR** or the function **PvalEqMeanVecR** the p-value, with a command like

```
> CDFEqMeanVecR(12,4,3,0.209951)
```

or

```
> PvalEqMeanVecR(12,4,3,0.209951)
```

which will give a p-value of 0.1788105357.

For more precision, one can use

```
> CDFEqMeanVecR(12,4,3,0.2099507759)
```

or

```
> PvalEqMeanVecR(12,4,3,0.2099507759)
```

to obtain a p-value of 0.1788101398.

The reverse way, one can obtain the 0.1788101398-quantile of Λ with a command as

```
> QuantEqMeanVecR(12,4,3,0.1788101398)
```

which will give 0.20995077590, confirming the results from **PvalDataEqMeanVecR**.

To obtain the 0.05 and 0.01 quantiles of Λ one may use

```
> QuantEqMeanVecR(12,4,3,0.05)
```

and

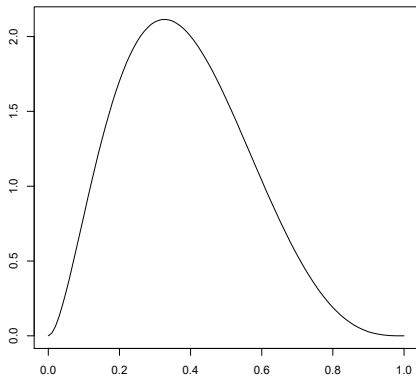
```
> QuantEqMeanVecR(12,4,3,0.01)
```

If one would like to take a look at the PDF and CDF of Λ , one may use

> `PlotPDFEqMeanVecR(12,4,3)` and > `PlotCDFEqMeanVecR(12,4,3)`

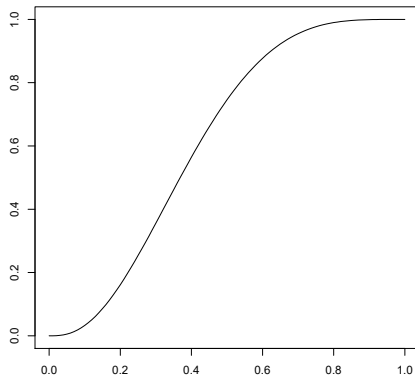
to obtain the plots of the PDF and CDF of Λ

PDFEqMeanVecR



n=12, p=4, q=3

CDFEqMeanVecR



n=12, p=4, q=3

What would it happen if the samples were larger in size (with exactly the same mean values, but also with smaller variances)?

Let us use now the file **ex1.7.dat**, which is similar to the file **ex1.1.dat** with each sample replicated 4 times (thus with exactly the same sample means but with larger sample sizes and smaller sample variances).

Now a command like

```
> PvalDataEqMeanVecR("ex1.7.dat")
```

will give still the same computed value for Λ , but now with a very smaller p-value (1.39252×10^{-11}), which would lead us to reject H_0 .

Note that this makes perfect sense, since larger samples will give us larger power to reject the null hypothesis!

We may wonder if this rejection is due to variable #3, although by looking better at the plot of sample means it seems that its presence affects more the parallelism of the profiles.

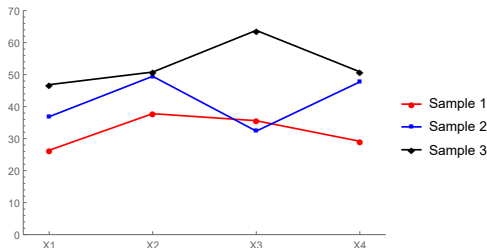
As such we may carry out a test of equality of the population mean vectors considering only variables 1, 2 and 4, by using a command like

```
> PvalDataEqMeanVecR("ex1_7.dat",1)
```

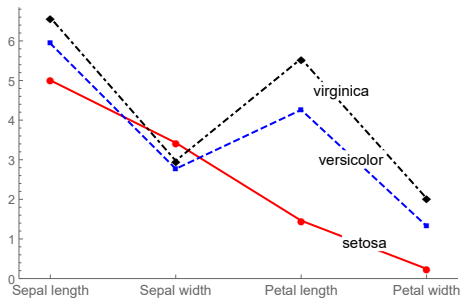
and select variables 1, 2 and 4.

This will still give a very low p-value of 1.279201×10^{-8} , showing that even only considering variables 1, 2 and 4 we should reject the equality of the population mean vectors for the 3 sub-populations.

We will see later which is the effect of removing variable #3 in the test of Profile Parallelism



We may also use the **Iris** dataset in file **Iris.dat**, where we have 3 samples (one for each *Iris* species), each of size 50, measured on 4 variables ('Sepal length', 'Sepal width', 'Petal length' and 'Petal width')



Note that the data file **Iris.dat** has a first column with the number of the observations, so that we have to use a command like

```
> PvalDataEqMeanVecR("Iris.dat",1)
```

and then select "variables" or columns 2, 3, 4 and 5 from the file, for the analysis (it was intentional to leave this column in the file to call the attention to situations like this one).

The extremely low p-value obtained indicates that we should reject the null hypothesis of equality of the population mean vectors for the 3 species of *Iris*, which comes at no surprise given the setup of the sample mean values and the size of the samples, which is rather large.

One can try to see what happens if we only consider the species *virginica* and *versicolor*. One should then use again the command

```
> PvalDataEqMeanVecR("Iris.dat",1)
```

and select samples 2 and 3, and "variables", or columns 2, 3, 4 and 5.

```
> PvalDataEqMeanVecR("iris.dat",1)
Is there a column in the file with the sample assignments ? (1=Yes) 1: 1
Which is the number of the column with the sample assignments ? 1: 6
Do you want to select samples ? (1=Yes) 1: 1
Please insert a list with the numbers of the samples
      you want to keep, separated by spaces: 1: 2 3
Do you want to select variables ? (1=Yes) 1: 1
Please insert a list with the numbers of the variables
      you want to keep, separated by spaces: 1: 2 3 4 5

There are 2 samples, with sizes:  50 50
and 4 variables
Original samples  2 3
Original variables 2 3 4 5

Computed value of Lambda:  0.216110297
p-value:  9.539876265e-31
```

The p-value obtained is still very low, indicating that we should reject the equality of the population mean vectors for these two species.

We might also want to carry out the same test using only the first 10 observations from each species. Then we would use the command

```
> PvalDataEqMeanVecR("iris.dat",1,1)
Is there a column in the file with the sample assignments ? (1=Yes) 1: 1
Which is the number of the column with the sample assignments ? 1: 6
Do you want to select samples ? (1=Yes) 1: 1
Please insert a list with the numbers of the samples
      you want to keep, separated by spaces: 1: 2 3
Do you want to select variables ? (1=Yes) 1: 1
Please insert a list with the numbers of the variables
      you want to keep, separated by spaces: 1: 2 3 4 5

There are 2 samples, with sizes:  50 50
and 4 variables
Original samples  2 3
Original variables 2 3 4 5

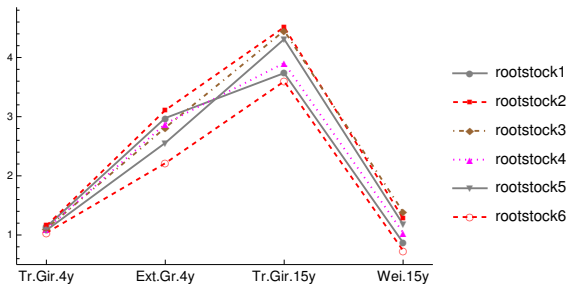
Please insert in each line a list with the observations to keep
in each sample, separated by spaces:
1: 1:10
1: 1:10

There are now  2  samples, with sizes:  10 10
and 4 variables

Computed value of Lambda:  0.08051448615
p-value:  4.914395518e-8
```

One may also consider the **Rootstock** data in Table 6.2 of Rencher (1995, 2002) or in Rencher and Christensen (2012), where are shown measurements for 8 trees from each of 6 different rootstocks on 4 variables ("trunk girth at 4 years (100mm)", "extension growth at 4 years (m)", "trunk girth at 15 years (100mm)", "weight of tree above ground at 15 years (1000lb)").

The data is reproduced in file **Rootstock.dat** and the sample mean values are:



Rencher, A. C. (1995, 2002). *Methods of Multivariate Analysis*, 1st, 2nd ed., Wiley

Rencher, A. C., Christensen, W. F. (2012). *Methods of Multivariate Analysis*, 3rd ed., Wiley

Then if we want to test the equality of the mean vectors for the 6 populations (the 6 different rootstocks), we only have to use the command

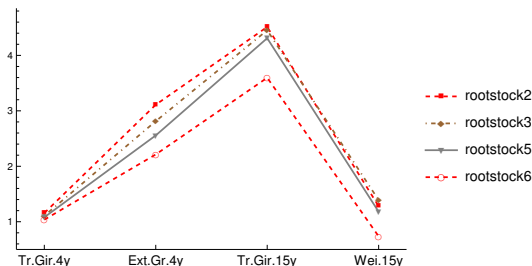
```
> PvalDataEqMeanVecR("Rootstock.dat")
```

from which we obtain a p-value of 7.717446×10^{-9} , which would lead us to reject the null hypothesis of equality of the 6 population mean vectors.

But if we look well at the plot of the sample means, we may think that if we consider only rootstocks 2, 3 and 5, by using the command

```
> PvalDataEqMeanVecR("Rootstock.dat", 1)
```

may be we would not reject the null hypothesis. And this is indeed the case, with a p-value of 0.096172 (which for an α value of 0.05 or smaller, would lead us to not reject the null hypothesis).



Then if we want to test the equality of the mean vectors for the 6 populations (the 6 different rootstocks), we only have to use the command

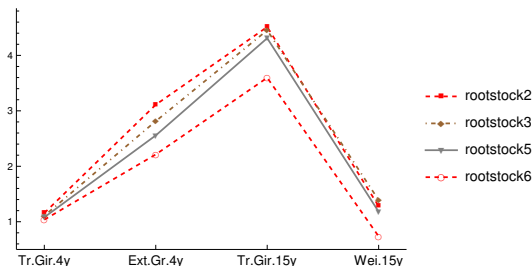
```
> PvalDataEqMeanVecR("Rootstock.dat")
```

from which we obtain a p-value of 7.717446×10^{-9} , which would lead us to reject the null hypothesis of equality of the 6 population mean vectors.

But if we look well at the plot of the sample means, we may think that if we consider only rootstocks 2, 3 and 5, by using the command

```
> PvalDataEqMeanVecR("Rootstock.dat", 1)
```

may be we would not reject the null hypothesis. And this is indeed the case, with a p-value of 0.096172 (which for an α value of 0.05 or smaller, would lead us to not reject the null hypothesis).



In all the above examples we had indeed p even, so that it was in all cases possible to compute the value for the LRT statistic and obtain the corresponding p-value. In case p is odd and q is even, the computed value of the statistic is still obtained, but no p-value is reported. This behavior of the function **PvalDataEqMeanVecR** may be observed by using the command

```
> PvalDataEqMeanVecR("Rootstock.dat",1)
```

and selecting for example only variables 1, 2 and 3.

```
> PvalDataEqMeanVecR("Rootstock.dat",1)
Is there a column in the file with the sample assignments ? (1=Yes) 1: 1
Which is the number of the column with the sample assignments ? 1: 1
Do you want to select samples ? (1=Yes) 1: 0
Do you want to select variables ? (1=Yes) 1: 1
Please insert a list with the numbers of the variables
    you want to keep, separated by spaces: 1: 1 2 3

There are 6 samples, with sizes: 8 8 8 8 8 8
and 3 variables
Original variables 1 2 3

Computed value of Lambda: 0.2459875567
[1] "The number of variables is odd and the number of samples even"
```


What can we do when p is odd and q is even, situation in which the exact distribution of Λ is not possible to be obtained in a closed finite form?

The answer is: one may use the so-called near-exact distributions available for example in

Coelho, C. A., Arnold, B. C., Marques, F. J. (2010) Near-exact distributions for certain likelihood ratio test statistics. *Journal of Statistical Theory and Practice*, 4, 4, 711-725 (invited paper for the special memorial issue in honor of H. C. Gupta)

Marques, F. J., Coelho, C. A., Arnold, B. C. (2011) A general near-exact distribution theory for the most common likelihood ratio test statistics used in Multivariate Analysis. *TEST*, 20, 180-203.

What can we do when p is odd and q is even, situation in which the exact distribution of Λ is not possible to be obtained in a closed finite form?

The answer is: one may use the so-called near-exact distributions available for example in

Coelho, C. A., Arnold, B. C., Marques, F. J. (2010) Near-exact distributions for certain likelihood ratio test statistics. Journal of Statistical Theory and Practice, 4, 4, 711-725 (invited paper for the special memorial issue in honor of H. C. Gupta)

Marques, F. J., Coelho, C. A., Arnold, B. C. (2011) A general near-exact distribution theory for the most common likelihood ratio test statistics used in Multivariate Analysis. TEST, 20, 180-203.

What are near-exact distributions?

These are asymptotic distributions obtained through a different process, where the *major part* of the original distribution is left untouched and only a *minor part* is asymptotically approximated, so that the resulting distribution is manageable (that is, from which it is easy to compute p-values and quantiles).

Typically, and opposite to common asymptotic distributions, near-exact distributions for LRT statistics exhibit very good approximations even for very small samples and are asymptotic for all the parameters in the distribution.

(See subsec 5.1.5 in Coelho and Arnold (2019))

If we consider now $\underline{X}_k \sim CN_p(\underline{\mu}_k, \Sigma)$ ($k = 1, \dots, q$), with $\underline{X}_k = \underline{Z}_{1k} + i\underline{Z}_{2k}$, where \underline{Z}_{1k} and \underline{Z}_{2k} are real normally distributed random vectors, and $\Sigma = 2\Gamma - 2i\Phi$ is Hermitian positive-definite, with

$$\text{Var}(\underline{Z}_{1k}) = \text{Var}(\underline{Z}_{2k}) = \Gamma, \quad (\Gamma \text{ positive-definite})$$

and

$$\text{Cov}(\underline{Z}_{1k}, \underline{Z}_{2k}) = \Phi \quad \text{with} \quad \text{Cov}(\underline{Z}_{2k}, \underline{Z}_{1k}) = \Phi' = -\Phi$$

(that is, Φ is a skew-symmetric matrix).

Then the $(2/n)$ -th power of the LRT statistic will still be given by

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

where

$$A = \sum_{k=1}^q (\underline{X}'_k - \overline{\underline{X}}_k E_{1n_k}) (\underline{X}'_k - \overline{\underline{X}}_k E_{1n_k})^\# \quad \text{and} \quad B = \sum_{k=1}^q n_k (\overline{\underline{X}}_k - \overline{\underline{X}}) (\overline{\underline{X}}_k - \overline{\underline{X}})^\#$$

where \underline{X}_k is the $n_k \times p$ matrix of the sample from \underline{X}_k , $\overline{\underline{X}}_k$ is the corresponding sample mean vector, E_{1n_k} is a $1 \times n_k$ matrix of 1's, $\overline{\underline{X}}$ is the overall sample mean vector, and $(\cdot)^\#$ denotes the transpose of the complex conjugate, but now with

$$\Lambda \sim \prod_{j=1}^p Y_j \sim \prod_{k=1}^{q-1} Y_k^*, \quad \text{where } Y_j \sim \text{Beta}(n - q - j + 1, q - 1) \quad \text{and} \quad Y_k^* \sim \text{Beta}(n - p - k, p),$$

for $n > p + q - 1$, form two sets of independent r.v.'s.

As such, the exact distribution of Λ is given by Theorems 1 or 2, with

$$m^* = 1, \quad a_1 = n - q + 1, \quad n_1 = p, \quad k_1 = 1, \quad \text{and} \quad m_1 = q - 1,$$

or

$$m^* = 1, \quad a_1 = n - p, \quad n_1 = q - 1, \quad k_1 = 1, \quad \text{and} \quad m_1 = p,$$

and thus the exact PDF and CDF of Λ are given by Corollaries 1 or 2, through the PDF and CDF of the EGIG distribution, as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-2}; \{n-1-j\}_{j=1:p+q-2}; p+q-2 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-2}; \{n-1-j\}_{j=1:p+q-2}; p+q-2 \right).$$

where

$$r_j = \begin{cases} h_j & j = 1 \\ r_{j-1} + h_j & j = 2, \dots, p+q-2, \end{cases}$$

with

$$\begin{aligned} h_j &= \begin{cases} 1, & j = 1, \dots, \min(p, q-1) \\ 0, & j = 1 + \min(p, q-1), \dots, \max(p, q-1) \\ -1, & j = 1 + \max(p, q-1), \dots, p+q-2 \end{cases} \\ &= (\# \text{ of elements in } \{p, q-1\} \geq j) - 1, \quad j = 1, \dots, p+q-2. \end{aligned}$$

For this test the data files that can be used have similar structures to those of the files used for the real case. Let us consider the files in Fig. 3,

Fig. 3

Examples of multi-sample layout files that work with the **ReadFileC** and **PvalDataEqMeanVecC** functions

variables					
X1	X2				
1.2+0.5i	2.6-3.1i	1.2+0.5i	2.6-3.1i	1.2+0.5i	2.6-3.1i 1
5.3-0.5i	1.1+5.6i	5.3-0.5i	1.1+5.6i	2.5+3.5i	4.6-3.1i 2
3.5+3.5i	1.6-3.1i	3.5+3.5i	1.6-3.1i	4.3-0.5i	2.1+5.6i 2
		2.5+3.5i	4.6-3.1i	1.5+3.5i	3.6-3.1i 5
		4.3-0.5i	2.1+5.6i	2.3-0.5i	1.1+5.6i 5
		3.5+3.5i	4.6-3.1i	5.3-0.5i	1.1+5.6i 1
		4.3-0.5i	2.1+5.6i	3.5+3.5i	4.6-3.1i 2
		1.5+3.5i	3.6-3.1i	3.5+3.5i	1.6-3.1i 1
		2.3-0.5i	1.1+5.6i	1.5+3.5i	2.6-3.1i 5
		1.5+3.5i	2.6-3.1i	5.3-0.5i	3.1+5.6i 5
		5.3-0.5i	3.1+5.6i	2.3-0.5i	3.1+5.6i 5
		2.3-0.5i	3.1+5.6i	4.3-0.5i	2.1+5.6i 2
		end of data			
ex1_comp.dat		ex2_comp.dat		ex3_comp.dat	

Then to carry out a test of a similar null hypothesis to the one for the real case, that is

$$H_0 : \underline{\mu}_1 = \cdots = \underline{\mu}_k = \cdots = \underline{\mu}_q,$$

we may use a command like

```
> PvalDataEqMeanVecC("ex1_comp.dat")
```

which will give a p-value of 0.648051, leading us to not reject H_0 , that is, to not reject the equality of the 3 population mean vectors.

The sample means are:

1st sample: 3.33+1.17i , 1.77-0.20i

2nd sample: 3.65+1.50i , 3.35+1.25i

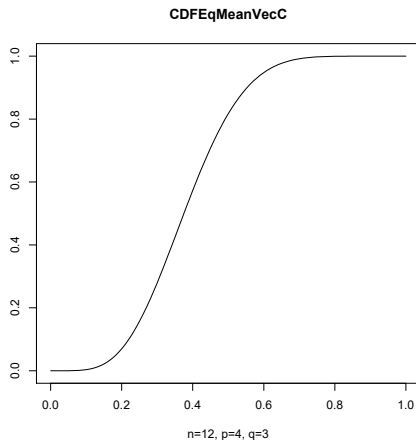
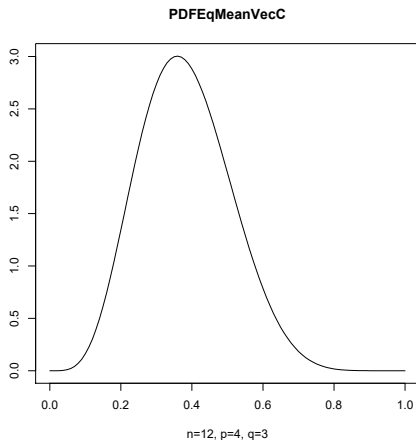
3rd sample: 2.58+1.10i , 2.70+1.22i.

If we use file **ex4_comp.dat**, where each sample is now repeated twice, the p-value would be 0.080897, and if we would use file **ex5_comp.dat**, where each original sample is repeated three times, the p-value would be 0.004794, showing how larger samples give higher power to reject the null hypothesis.

If one would like to take a look at the PDF and CDF of Λ , for $n = 12$, $p = 4$ and $q = 3$, to compare it with the one for the real case, one may use

> **PlotPDFEqMeanVecC(12,4,3)** and > **PlotCDFEqMeanVecC(12,4,3)**

to obtain the plots of the PDF and CDF of Λ



(See subsec 5.1.2 in Coelho and Arnold (2019))

Under a similar setup of that of the LRT for equality of mean vectors, the $(2/n)$ -th power of the LRT statistic for the test of simultaneous nullity of the mean vectors, that is, to test the null hypothesis

$$\underline{\mu}_1 = \cdots = \underline{\mu}_k = \cdots = \underline{\mu}_q = \underline{0},$$

is,

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

now with

$$B^* = \sum_{k=1}^q n_k \bar{\underline{X}}_k \bar{\underline{X}}_k',$$

and with

$$\Lambda \sim \prod_{j=1}^p Y_j \sim \prod_{k=1}^q Y_k^*$$

where, for $n > p + q - 1$,

$$Y_j \sim \text{Beta}\left(\frac{n - q + 1 - j}{2}, \frac{q}{2}\right) \quad \text{and} \quad Y_k^* \sim \text{Beta}\left(\frac{n - p + 1 - k}{2}, \frac{p}{2}\right)$$

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

As such, the exact distribution of Λ is given, for even p , by Theorems 1 or 2, with

$$m^* = 1, \quad a_1 = \frac{n-q+1}{2}, \quad n_1 = \frac{p}{2}, \quad k_1 = 2, \quad \text{and} \quad m_1 = q,$$

and for even q , with

$$m^* = 1, \quad a_1 = \frac{n-p+1}{2}, \quad n_1 = \frac{q}{2}, \quad k_1 = 2, \quad \text{and} \quad m_1 = p,$$

and thus the exact PDF and CDF of Λ are, for even p or even q , given by Corollaries 1 or 2, through the PDF and CDF of the EGIG distribution, as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-2}; \left\{ \frac{n-1-j}{2} \right\}_{j=1:p+q-2}; p+q-2 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-2}; \left\{ \frac{n-1-j}{2} \right\}_{j=1:p+q-2}; p+q-2 \right).$$

where

$$r_j = \begin{cases} h_j & j = 1, 2 \\ r_{j-1} + h_j & j = 3, \dots, p+q-2, \end{cases}$$

with

$$\begin{aligned} h_j &= \begin{cases} 1, & j = 1, \dots, \min(p, q) \\ 0, & j = 1 + \min(p, q), \dots, \max(p, q) \\ -1, & j = 1 + \max(p, q), \dots, p+q-2 \end{cases} \\ &= (\# \text{ of elements in } \{p, q\} \geq j) - 1, \quad j = 1, \dots, p+q-2. \end{aligned}$$

Then using the command

```
> PvalDataNullMeanVecR("ex1_1.dat")
```

one will obtain a p-value of 0.001168, which would lead us to reject the null hypothesis of simultaneous nullity of the 3 population mean vectors, and which comes indeed at no surprise.

Also at no surprise comes the rejection of the null hypothesis of simultaneous nullity of the population mean vectors for all 3 species of *Iris*, with the command

```
> PvalDataNullMeanVecR("Iris.dat",1)
```

giving an extremely low p-value of $1.847886 \times 10^{-261}$, and the test to the simultaneous nullity of the population mean vectors for the 6 rootstocks giving a p-value of 1.698012×10^{-56} . Note that there are files for the rootstock data which have the variables on a different scale. Note also that that fact does not affect the results for these tests, neither in terms of the computed value of the statistic, neither then in terms of the p-value. Try to use files **Rootstock.dat** and **Rootstock_2.dat**.

In line with what happens for the test of equality of real mean vectors, in case p and q are both odd, the function **PvalDataNullMeanVecR** will only give the computed value of the LRT statistic and not the p-value.

Similarly to what happened with the LRT for equality of mean vectors of real random vectors, the question is: **what to do when p and q are both odd?**

The answer is: one may easily build near-exact distributions in a similar way of those built for the LRT statistic of the test of equality of real mean vectors (see slide 47), by using the results in the references on that slide, or the results in

Coelho, C. A., Alberto, R. P. and Grilo, L. M. (2006). A mixture of Generalized Integer Gamma distributions as the exact distribution of the product of an odd number of independent Beta random variables: applications. *Journal of Interdisciplinary Mathematics*, 9, 229-248.

Coelho, C. A. (2006). The exact and near-exact distributions of the product of independent Beta random variables whose second parameter is rational. *Journal of Combinatorics, Information & System Sciences*, 31, 21-44.

Grilo, L. M., Coelho, C. A. (2007). Development and study of two near-exact approximations to the distribution of the product of an odd number of independent Beta random variables. *Journal of Statistical Planning and Inference*, 137, 1560-1575.

Similarly to what happened with the LRT for equality of mean vectors of real random vectors, the question is: **what to do when p and q are both odd?**

The answer is: one may easily build near-exact distributions in a similar way of those built for the LRT statistic of the test of equality of real mean vectors (see slide 47), by using the results in the references on that slide, or the results in

Coelho, C. A., Alberto, R. P. and Grilo, L. M. (2006). A mixture of Generalized Integer Gamma distributions as the exact distribution of the product of an odd number of independent Beta random variables: applications. *Journal of Interdisciplinary Mathematics*, 9, 229-248.

Coelho, C. A. (2006). The exact and near-exact distributions of the product of independent Beta random variables whose second parameter is rational. *Journal of Combinatorics, Information & System Sciences*, 31, 21-44.

Grilo, L. M., Coelho, C. A. (2007). Development and study of two near-exact approximations to the distribution of the product of an odd number of independent Beta random variables. *Journal of Statistical Planning and Inference*, 137, 1560-1575.

(See subsec 5.1.6 in Coelho and Arnold (2019))

Similarly to what happens in the real case, we want to test the null hypothesis

$$\underline{\mu}_1 = \cdots = \underline{\mu}_k = \cdots = \underline{\mu}_q = \underline{0},$$

now with

$$\underline{X}_k \sim CN_p(\underline{\mu}_k, \Sigma),$$

where the multivariate complex Normal distribution is defined as in the case of the **EqMeanVecC** test.

Then the $(2/n)$ -th power of the LRT statistic to test H_0 is

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

with A defined as in the case of the **EqMeanVecC** test, and

$$B^* = \sum_{k=1}^q n_k \overline{\underline{X}}_k \overline{\underline{X}}_k^{\#},$$

yielding

$$\Lambda \sim \prod_{j=1}^p Y_j \sim \prod_{k=1}^q Y_k^*,$$

where, for $n > p + q - 1$,

$$Y_j \sim \text{Beta}(n - q + 1 - j, q) \quad \text{and} \quad Y_k^* \sim \text{Beta}(n - p + 1 - k, p)$$

form two sets of independent r.v.'s, and thus with the exact distribution of Λ given by Theorems 1 or 2 with

$$m^* = 1, \quad a_1 = n - q + 1, \quad n_1 = p, \quad k_1 = 1, \quad \text{and} \quad m_1 = q,$$

or

$$m^* = 1, \quad a_1 = n - p + 1, \quad n_1 = q, \quad k_1 = 1, \quad \text{and} \quad m_1 = p,$$

and its PDF and CDF given by Corollaries 1 or 2, through the PDF and cDF of the EGIG distribution as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-1}; \{n-j\}_{j=1:p+q-1}; p+q-1 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-1}; \{n-j\}_{j=1:p+q-1}; p+q-1 \right).$$

where

$$r_j = \begin{cases} h_j & j = 1 \\ r_{j-1} + h_j & j = 2, \dots, p+q-1, \end{cases}$$

with

$$\begin{aligned} h_j &= \begin{cases} 1, & j = 1, \dots, \min(p, q) \\ 0, & j = 1 + \min(p, q), \dots, \max(p, q) \\ -1, & j = 1 + \max(p, q), \dots, p+q-1 \end{cases} \\ &= (\# \text{ of elements in } \{p, q\} \geq j) - 1, \quad j = 1, \dots, p+q-1. \end{aligned}$$

Then, if one wants to test H_0 for the dataset in the file `ex1_comp.dat`, one only has to use a command like

```
> PvalDataNullMeanVecC("ex1_comp.dat")
```

obtaining a p-value of 6.114046×10^{-7} , a very low p-value that actually comes at no surprise, since the sample means were quite different from zero.

If one wants to take a look at the PDF and CDF of Λ , for $n = 12$, $p = 4$ and $q = 3$, to compare it with the one for the test of equality of mean vectors, one may use

```
> PlotPDFNullMeanVecC(12,4,3)    and    > PlotCDFNullMeanVecC(12,4,3)
```

to obtain the plots of the PDF and CDF of Λ

Then, if one wants to test H_0 for the dataset in the file `ex1_comp.dat`, one only has to use a command like

```
> PvalDataNullMeanVecC("ex1_comp.dat")
```

obtaining a p-value of 6.114046×10^{-7} , a very low p-value that actually comes at no surprise, since the sample means were quite different from zero.

If one wants to take a look at the PDF and CDF of Λ , for $n = 12$, $p = 4$ and $q = 3$, to compare it with the one for the test of equality of mean vectors, one may use

```
> PlotPDFNullMeanVecC(12,4,3)    and    > PlotCDFNullMeanVecC(12,4,3)
```

to obtain the plots of the PDF and CDF of Λ

Then, if one wants to test H_0 for the dataset in the file `ex1_comp.dat`, one only has to use a command like

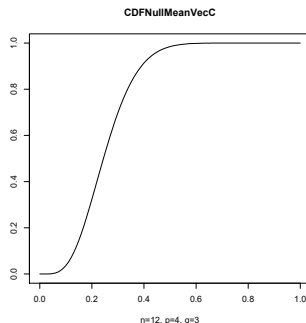
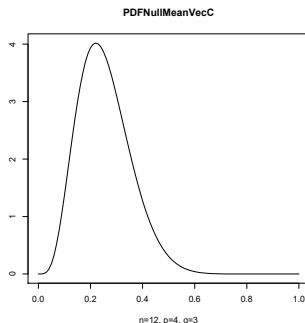
```
> PvalDataNullMeanVecC("ex1_comp.dat")
```

obtaining a p-value of 2.559643×10^{-7} , a very low p-value that actually comes at no surprise, since the sample means were quite different from zero.

If one wants to take a look at the PDF and CDF of Λ , for $n = 12$, $p = 4$ and $q = 3$, to compare it with the one for the test of equality of mean vectors, one may use

```
> PlotPDFNullMeanVecC(12,4,3) and > PlotCDFNullMeanVecC(12,4,3)
```

to obtain the plots of the PDF and CDF of Λ



(See subsec 5.1.3 in Coelho and Arnold (2019))

Let us consider a similar setup to the one used for the test of equality of mean vectors, and let us suppose we are interested in testing the null hypothesis of profile parallelism. This hypothesis may be written as

$$C\mu_1 = \dots = C\mu_k = \dots = C\mu_q,$$

where

$$C_{(p-1) \times p} = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix}.$$

We may note that

$$C\mu_k_{(p-1) \times 1} = \begin{bmatrix} \mu_{k1} - \mu_{k2} \\ \mu_{k2} - \mu_{k3} \\ \vdots \\ \mu_{k,p-1} - \mu_{kp} \end{bmatrix}.$$

(See subsec 5.1.3 in Coelho and Arnold (2019))

Let us consider a similar setup to the one used for the test of equality of mean vectors, and let us suppose we are interested in testing the null hypothesis of profile parallelism. This hypothesis may be written as

$$C\mu_1 = \dots = C\mu_k = \dots = C\mu_q,$$

where

$$C_{(p-1) \times p} = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix}.$$

We may note that

$$C\mu_k_{(p-1) \times 1} = \begin{bmatrix} \mu_{k1} - \mu_{k2} \\ \mu_{k2} - \mu_{k3} \\ \vdots \\ \mu_{k,p-1} - \mu_{kp} \end{bmatrix}.$$

Then the $(2/n)$ -th power of the LRT statistic to test H_0 is

$$\Lambda = \frac{|CAC'|}{|C(A+B)C'|}$$

where A and B are the matrices in the test of equality of mean vectors.

And then it may be shown that

$$\Lambda \sim \prod_{j=1}^{p-1} Y_j \sim \prod_{k=1}^{q-1} Y_k^*$$

where, for $n > p + q - 1$,

$$Y_j \sim \text{Beta}\left(\frac{n-q+1-j}{2}, \frac{q-1}{2}\right) \quad \text{and} \quad Y_k^* \sim \text{Beta}\left(\frac{n-p+1-k}{2}, \frac{p-1}{2}\right)$$

form two sets on independent r.v.'s.

As such, for odd p , the exact distribution of Λ is given by Theorems 1 or 2, with

$$m^* = 1, \quad a_1 = \frac{n-q+1}{2}, \quad n_1 = \frac{p-1}{2}, \quad k_1 = 2, \quad \text{and} \quad m_1 = q-1,$$

and for odd q , with

$$m^* = 1, \quad a_1 = \frac{n-p+1}{2}, \quad n_1 = \frac{q-1}{2}, \quad k_1 = 2, \quad \text{and} \quad m_1 = p-1,$$

Then the $(2/n)$ -th power of the LRT statistic to test H_0 is

$$\Lambda = \frac{|CAC'|}{|C(A+B)C'|}$$

where A and B are the matrices in the test of equality of mean vectors.

And then it may be shown that

$$\Lambda \sim \prod_{j=1}^{p-1} Y_j \sim \prod_{k=1}^{q-1} Y_k^*$$

where, for $n > p + q - 1$,

$$Y_j \sim \text{Beta}\left(\frac{n-q+1-j}{2}, \frac{q-1}{2}\right) \quad \text{and} \quad Y_k^* \sim \text{Beta}\left(\frac{n-p+1-k}{2}, \frac{p-1}{2}\right)$$

form two sets on independent r.v.'s.

As such, for odd p , the exact distribution of Λ is given by Theorems 1 or 2, with

$$m^* = 1, \quad a_1 = \frac{n-q+1}{2}, \quad n_1 = \frac{p-1}{2}, \quad k_1 = 2, \quad \text{and} \quad m_1 = q-1,$$

and for odd q , with

$$m^* = 1, \quad a_1 = \frac{n-p+1}{2}, \quad n_1 = \frac{q-1}{2}, \quad k_1 = 2, \quad \text{and} \quad m_1 = p-1,$$

Then the $(2/n)$ -th power of the LRT statistic to test H_0 is

$$\Lambda = \frac{|CAC'|}{|C(A+B)C'|}$$

where A and B are the matrices in the test of equality of mean vectors.

And then it may be shown that

$$\Lambda \sim \prod_{j=1}^{p-1} Y_j \sim \prod_{k=1}^{q-1} Y_k^*$$

where, for $n > p + q - 1$,

$$Y_j \sim \text{Beta}\left(\frac{n-q+1-j}{2}, \frac{q-1}{2}\right) \quad \text{and} \quad Y_k^* \sim \text{Beta}\left(\frac{n-p+1-k}{2}, \frac{p-1}{2}\right)$$

form two sets on independent r.v.'s.

As such, for odd p , the exact distribution of Λ is given by Theorems 1 or 2, with

$$m^* = 1, \quad a_1 = \frac{n-q+1}{2}, \quad n_1 = \frac{p-1}{2}, \quad k_1 = 2, \quad \text{and} \quad m_1 = q-1,$$

and for odd q , with

$$m^* = 1, \quad a_1 = \frac{n-p+1}{2}, \quad n_1 = \frac{q-1}{2}, \quad k_1 = 2, \quad \text{and} \quad m_1 = p-1,$$

and its PDF and CDF by Corollaries 1 or 2, in terms of the EGIG PDF and CDF as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-4}; \left\{ \frac{n-2-j}{2} \right\}_{j=1:p+q-4}; p+q-4 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-4}; \left\{ \frac{n-2-j}{2} \right\}_{j=1:p+q-4}; p+q-4 \right).$$

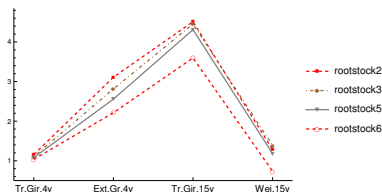
where

$$r_j = \begin{cases} h_j & j = 1, 2 \\ r_{j-2} + h_j & j = 3, \dots, p+q-4, \end{cases}$$

with

$$\begin{aligned} h_j &= \begin{cases} 1, & j = 1, \dots, \min(p-1, q-1) \\ 0, & j = 1 + \min(p-1, q-1), \dots, \max(p-1, q-1) \\ -1, & j = 1 + \max(p-1, q-1), \dots, p+q-4 \end{cases} \\ &= (\# \text{ of elements in } \{p-1, q-1\} \geq j) - 1, \quad j = 1, \dots, p+q-4. \end{aligned}$$

Let us consider the example of the Rootstocks, considering only rootstocks 2, 3, 5 and 6



and use a command like

```
> PvalDataProfParR("Rootstock.dat", 1)
```

and select, for example

- rootstocks 2,3,5 and 6, and all variables (p-value: ?)
- rootstocks 2,3,5 and 6, and variables 2, 3 and 4 (p-value: 0.004339)
- rootstocks 2,3 and 5 and all variables (p-value: 0.032448)
- rootstocks 2,3 and 5 and variables 2, 3 and 4 (p-value: 0.201752)
- rootstocks 2,3 and 6 and all variables (p-value: 5.782751×10^{-5})
- rootstocks 2,3 and 6 and variables 2, 3 and 4 (p-value: 0.003611)

and comment!

Similarly to what happened with the LRTs for equality and nullity of mean vectors of real random vectors, the question is: **what to do when p and q are both even?**

The answer is once again: we should use near-exact distributions.

These may be easily built in a similar way of those for the LRT statistic of the test of equality of real mean vectors (see slide 47 and the results in the references on slide 56).

Similarly to what happened with the LRTs for equality and nullity of mean vectors of real random vectors, the question is: **what to do when p and q are both even?**

The answer is once again: we should use near-exact distributions.

These may be easily built in a similar way of those for the LRT statistic of the test of equality of real mean vectors (see slide 47 and the results in the references on slide 56).

(See subsec 5.1.7 in Coelho and Arnold (2019))

Let us suppose a similar setup to the one used for the tests of equality of mean vectors and of nullity of mean vectors for complex Normal r.v.'s. Then the $(2/n)$ -th power of the LRT statistic to test the null hypothesis of parallelism of the profiles,

$$C\mu_1 = \cdots = C\mu_k = \cdots = C\mu_q,$$

where C is the same matrix used in the case of real r.v.'s, is

$$\Lambda = \frac{|CAC'|}{|C(A+B)C'|}$$

where A and B are the matrices defined for the test of equality of complex mean vectors.

Then,

$$\Lambda \sim \prod_{j=1}^{p-1} Y_j \sim \prod_{k=1}^{q-1} Y_k^*$$

where, for $n > p + q - 1$,

$$Y_j \sim \text{Beta}(n - q - j + 1, q - 1) \quad \text{and} \quad Y_k^* \sim \text{Beta}(n - p - k + 1, p - 1)$$

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

As such, the exact distribution of Λ is given by Theorems 1 or 2, with

$$m^* = 1, \quad a_1 = n - q + 1, \quad n_1 = p - 1, \quad k_1 = 1, \quad \text{and} \quad m_1 = q - 1,$$

or

$$m^* = 1, \quad a_1 = n - p + 1, \quad n_1 = q - 1, \quad k_1 = 1, \quad \text{and} \quad m_1 = p - 1,$$

and thus the exact PDF and CDF of Λ are given by Corollaries 1 or 2, through the PDF and CDF of the EGIG distribution, as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-3}; \{n-1-j\}_{j=1:p+q-3}; p+q-3 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+q-3}; \{n-1-j\}_{j=1:p+q-3}; p+q-3 \right).$$

where

$$r_j = \begin{cases} h_j & j = 1 \\ r_{j-1} + h_j & j = 2, \dots, p+q-3, \end{cases}$$

with

$$\begin{aligned} h_j &= \begin{cases} 1, & j = 1, \dots, \min(p-1, q-1) \\ 0, & j = 1 + \min(p-1, q-1), \dots, \max(p-1, q-1) \\ -1, & j = 1 + \max(p-1, q-1), \dots, p+q-3 \end{cases} \\ &= (\# \text{ of elements in } \{p-1, q-1\} \geq j) - 1, \quad j = 1, \dots, p+q-3. \end{aligned}$$

Then we may carry out a Profile Parallelism test with files `ex1_comp.dat`, `ex4_comp.dat` and `ex5_comp.dat`, with the commands

```
> PvalDataProfParC("ex1_comp.dat")  
> PvalDataProfParC("ex4_comp.dat")  
> PvalDataProfParC("ex5_comp.dat")
```

and analyze the p-values obtained, and discuss their values, namely also comparing them with the p-values obtained for the test of equality of mean vectors.

May be we should take into account the plot of the sample means, where each mean value is represented by a point with the real value as the coordinate on the x-axis and the imaginary value as the coordinate on the y-axis

Then we may carry out a Profile Parallelism test with files `ex1_comp.dat`, `ex4_comp.dat` and `ex5_comp.dat`, with the commands

```
> PvalDataProfParC("ex1_comp.dat")  
> PvalDataProfParC("ex4_comp.dat")  
> PvalDataProfParC("ex5_comp.dat")
```

and analyze the p-values obtained, and discuss their values, namely also comparing them with the p-values obtained for the test of equality of mean vectors.

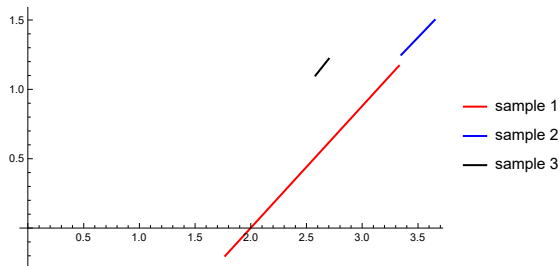
Maybe we should take into account the plot of the sample means, where each mean value is represented by a point with the real value as the coordinate on the x-axis and the imaginary value as the coordinate on the y-axis

Then we may carry out a Profile Parallelism test with files **ex1_comp.dat**, **ex4_comp.dat** and **ex5_comp.dat**, with the commands

```
> PvalDataProfParC("ex1_comp.dat") (p-value: 0.788334)
> PvalDataProfParC("ex4_comp.dat") (p-value: 0.592601)
> PvalDataProfParC("ex5_comp.dat") (p-value: 0.445466)
```

and analyze the p-values obtained, and discuss their values, namely also comparing them with the p-values obtained for the test of equality of mean vectors.

Maybe we should take into account the plot of the sample means, where each mean value is represented by a point with the real value as the coordinate on the x-axis and the imaginary value as the coordinate on the y-axis



The sample means are:

1st sample: $3.33+1.17i$, $1.77-0.20i$
2nd sample: $3.65+1.50i$, $3.35+1.25i$
3rd sample: $2.58+1.10i$, $2.70+1.22i$

Functions `ReadFile1sR` and `ReadFile1sC` used to read files and convert them to the 1 sample format

70

As an example of application let us consider reading the data in file `JW_Tab9_12.dat` which reproduces Table 9.12 of Johnson and Wichern (2014) referring to 3 sales performance indexes ("sales growth", "sales profitability", "new-account sales") and 4 test scores ("creativity test", "mechanical reasoning test", "abstract reasoning test", "mathematics test") measured on 50 sales people. We will just use the command

```
> ReadFile1sR("JW_Tab9_12.dat")
```

We will be using this file in the next test.

One may also read the file `ex1_1.dat` with this function `ReadFile1sR` to see the behavior of the function and possible outputs obtained when reading a multi-sample file with function `ReadFile1sR`. Just use the command

```
> ReadFile1sR("ex1_1.dat")
```

`ReadFile1sR` is the function that will be used by the functions `PvalData<acronym>` that we will be using with the LRTs that we will address next.

Johnson, R., Wichern, D. (2014). *Applied Multivariate Statistical Analysis*. 6th ed., Pearson New Int. Edition, Pearson

Functions **ReadFile1sR** and **ReadFile1sC** used to read files and convert them to the 1 sample format

70

As an example of application let us consider reading the data in file **JW_Tab9_12.dat** which reproduces Table 9.12 of Johnson and Wichern (2014) referring to 3 sales performance indexes ("sales growth", "sales profitability", "new-account sales") and 4 test scores ("creativity test", "mechanical reasoning test", "abstract reasoning test", "mathematics test") measured on 50 sales people. We will just use the command

```
> ReadFile1sR("JW_Tab9_12.dat")
```

We will be using this file in the next test.

One may also read the file **ex1_1.dat** with this function **ReadFile1sR** to see the behavior of the function and possible outputs obtained when reading a multi-sample file with function **ReadFile1sR**. Just use the command

```
> ReadFile1sR("ex1_1.dat")
```

ReadFile1sR is the function that will be used by the functions **PvalData<acronym>** that we will be using with the LRTs that we will address next.

Johnson, R., Wichern, D. (2014). *Applied Multivariate Statistical Analysis*. 6th ed., Pearson New Int. Edition, Pearson

Functions **ReadFile1sR** and **ReadFile1sC** used to read files and convert them to the 1 sample format

70

As an example of application let us consider reading the data in file **JW_Tab9_12.dat** which reproduces Table 9.12 of Johnson and Wichern (2014) referring to 3 sales performance indexes ("sales growth", "sales profitability", "new-account sales") and 4 test scores ("creativity test", "mechanical reasoning test", "abstract reasoning test", "mathematics test") measured on 50 sales people. We will just use the command

```
> ReadFile1sR("JW_Tab9_12.dat")
```

We will be using this file in the next test.

One may also read the file **ex1_1.dat** with this function **ReadFile1sR** to see the behavior of the function and possible outputs obtained when reading a multi-sample file with function **ReadFile1sR**. Just use the command

```
> ReadFile1sR("ex1_1.dat")
```

ReadFile1sR is the function that will be used by the functions **PvalData**<acronym> that we will be using with the LRTs that we will address next.

Johnson, R., Wichern, D. (2014). *Applied Multivariate Statistical Analysis*. 6th ed., Pearson New Int. Edition, Pearson

One other dataset we may take a look at, and which we will be using later, is the one in file **winequality-red.dat**. This is a dataset obtained by the authors of the paper mentioned at the end of the file. The names of the variables are quite self-explanatory, but a more detailed description of the data may be found at the end of the file.

We may note that functions **ReadFileR**, **ReadFileC**, **ReadFile1sR** and **ReadFile1sC** have no problem reading files that besides the data may have comments on the data inserted at the beginning or the end of the file.

(See subsec 5.1.9 in Coelho and Arnold (2019))

Let us suppose that $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, where $\underline{X} = [\underline{X}'_1, \underline{X}'_2]'$, with \underline{X}_1 of dimension $p_1 \times 1$ and \underline{X}_2 of dimension $p_2 \times 1$, and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}$$

where $\Sigma_{11} = \text{var}(\underline{X}_1)$, $\Sigma_{22} = \text{Var}(\underline{X}_2)$ and $\Sigma_{12} = \text{Cov}(\underline{X}_1, \underline{X}_2) = \Sigma'_{21}$.

Then, the hypothesis of independence of \underline{X}_1 and \underline{X}_2 may be written as

$$H_0 : \Sigma = \text{bdiag}(\Sigma_{11}, \Sigma_{22}) \iff \Sigma_{12} = 0_{p_1 \times p_2}$$

and for a sample of size n , the $(2/n)$ -th power of the LRT statistic to test H_0 is

$$\Lambda = \frac{|A|}{|A_{11}| |A_{22}|},$$

where A is the MLE of Σ , and A_{11} and A_{22} its diagonal blocks, considering a split of A according to the split of Σ .

(See subsec 5.1.9 in Coelho and Arnold (2019))

Let us suppose that $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, where $\underline{X} = [\underline{X}'_1, \underline{X}'_2]'$, with \underline{X}_1 of dimension $p_1 \times 1$ and \underline{X}_2 of dimension $p_2 \times 1$, and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}$$

where $\Sigma_{11} = \text{var}(\underline{X}_1)$, $\Sigma_{22} = \text{Var}(\underline{X}_2)$ and $\Sigma_{12} = \text{Cov}(\underline{X}_1, \underline{X}_2) = \Sigma'_{21}$.

Then, the hypothesis of independence of \underline{X}_1 and \underline{X}_2 may be written as

$$H_0 : \Sigma = \text{bdiag}(\Sigma_{11}, \Sigma_{22}) \iff \Sigma_{12} = 0_{p_1 \times p_2}$$

and for a sample of size n , the $(2/n)$ -th power of the LRT statistic to test H_0 is

$$\Lambda = \frac{|A|}{|A_{11}| |A_{22}|},$$

where A is the MLE of Σ , and A_{11} and A_{22} its diagonal blocks, considering a split of A according to the split of Σ .

Then it is possible to show that

$$\Lambda \sim \prod_{j=1}^{p_1} Y_j \sim \prod_{k=1}^{p_2} Y_k^*$$

where, for $n > p_1 + p_2$,

$$Y_j \sim \text{Beta}\left(\frac{n - p_2 - j}{2}, \frac{p_2}{2}\right) \quad \text{and} \quad Y_k^* \sim \text{Beta}\left(\frac{n - p_1 - k}{2}, \frac{p_1}{2}\right)$$

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

But then, the exact distribution of Λ is given, for even p_1 , by Theorems 1 or 2, with

$$m^* = 1, \quad k_1 = 2, \quad a_1 = \frac{n - p_2}{2}, \quad n_1 = \frac{p_1}{2}, \quad m_1 = p_2,$$

and for even p_2 , by the same Theorems, with

$$m^* = 1, \quad k_1 = 2, \quad a_1 = \frac{n - p_1}{2}, \quad n_1 = \frac{p_2}{2}, \quad m_1 = p_1,$$

and thus its PDF and CDF is for either even p_1 or even p_2 given by Corollaries 1 or 2, in terms of the EGIG PDF and CDF as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p_1+p_2-2}; \left\{ \frac{n-2-j}{2} \right\}_{j=1:p_1+p_2-2}; p_1 + p_2 - 2 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p_1+p_2-2}; \left\{ \frac{n-2-j}{2} \right\}_{j=1:p_1+p_2-2}; p_1 + p_2 - 2 \right).$$

where

$$r_j = \begin{cases} h_j & j = 1, 2 \\ r_{j-2} + h_j & j = 3, \dots, p_1 + p_2 - 2, \end{cases}$$

with

$$\begin{aligned} h_j &= \begin{cases} 1, & j = 1, \dots, \min(p_1, p_2) \\ 0, & j = 1 + \min(p_1, p_2), \dots, \max(p_1, p_2) \\ -1, & j = 1 + \max(p_1, p_2), \dots, p_1 + p_2 - 2 \end{cases} \\ &= (\# \text{ of elements in } \{p_1, p_2\} \geq j) - 1, \quad j = 1, \dots, p_1 + p_2 - 2. \end{aligned}$$

As an example of application let us consider the data in Table 9.12 of Johnson and Wichern (2014). We want to run a test of independence between the 3 sales performance indexes ("sales growth", "sales profitability", "new-account sales") considered as making set \underline{X}_1 and the 4 test scores ("creativity test", "mechanical reasoning test", "abstract reasoning test", "mathematics test"), considered as making part of set \underline{X}_2 . We will consider the data in file `JW_Tab9_12.dat`.

One can use function `EGIG_help` with a command like

```
> EGIG_help(PvalDataInd2R)
```

to see how one can use function `PvalDataInd2R`. Then we can use a command like

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3)
```

to obtain the computed value for the statistic for the test of independence between \underline{X}_1 and \underline{X}_2 . The computed value of the statistic is 0.002148, with a corresponding p-value of 1.254087×10^{-51} , which would lead us to reject the null hypothesis of independence of the two sets of variables, the one of the sales indexes and the one of the scores on the tests, showing a very close relationship between the 2 sets of variables.

As an example of application let us consider the data in Table 9.12 of Johnson and Wichern (2014). We want to run a test of independence between the 3 sales performance indexes ("sales growth", "sales profitability", "new-account sales") considered as making set X_1 and the 4 test scores ("creativity test", "mechanical reasoning test", "abstract reasoning test", "mathematics test"), considered as making part of set X_2 . We will consider the data in file `JW_Tab9_12.dat`.

One can use function `EGIG_help` with a command like

```
> EGIG_help(PvalDataInd2R)
```

to see how one can use function `PvalDataInd2R`. Then we can use a command like

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3)
```

to obtain the computed value for the statistic for the test of independence between X_1 and X_2 . The computed value of the statistic is 0.002148, with a corresponding p-value of 1.254087×10^{-51} , which would lead us to reject the null hypothesis of independence of the two sets of variables, the one of the sales indexes and the one of the scores on the tests, showing a very close relationship between the 2 sets of variables.

As an example of application let us consider the data in Table 9.12 of Johnson and Wichern (2014). We want to run a test of independence between the 3 sales performance indexes ("sales growth", "sales profitability", "new-account sales") considered as making set \underline{X}_1 and the 4 test scores ("creativity test", "mechanical reasoning test", "abstract reasoning test", "mathematics test"), considered as making part of set \underline{X}_2 . We will consider the data in file `JW_Tab9_12.dat`.

One can use function `EGIG_help` with a command like

```
> EGIG_help(PvalDataInd2R)
```

to see how one can use function `PvalDataInd2R`. Then we can use a command like

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3)
```

to obtain the computed value for the statistic for the test of independence between \underline{X}_1 and \underline{X}_2 . The computed value of the statistic is 0.002148, with a corresponding p-value of 1.254087×10^{-51} , which would lead us to reject the null hypothesis of independence of the two sets of variables, the one of the sales indexes and the one of the scores on the tests, showing a very close relationship between the 2 sets of variables.

One could also have used a file with the data for the 1st set, say file `JW_Tab9_12_X1.dat` and another file with the data for the 2nd set, say file `JW_Tab9_12_X2.dat` and then use the command

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat")
```

to obtain exactly the same result.

In order to try to evaluate if the relation between the 3 sales indexes is stronger with the first 2 or the last 2 test scores, and also to see how function `PvalDataInd2R` works, let us do an independence test between the set of the 3 sales indexes and each one of these 2 subsets, first using all the 50 observations and then using only the first 20 observations, for both cases, when using a single data file and when using the 2 data files.

This will be achieved using commands like

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3, 1) (to select variables, using all 50 observations, using a single dataset)
```

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3, 1, 1) (to select variables and observations to be used, using a single dataset)
```

or

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat", , 1) (to select variables, using all 50 observations, using two datasets)
```

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat", , 1, 1) (to select variables and observations to be used, using two datasets)
```

One could also have used a file with the data for the 1st set, say file `JW_Tab9_12_X1.dat` and another file with the data for the 2nd set, say file `JW_Tab9_12_X2.dat` and then use the command

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat")
```

to obtain exactly the same result.

In order to try to evaluate if the relation between the 3 sales indexes is stronger with the first 2 or the last 2 test scores, and also to see how function `PvalDataInd2R` works, let us do an independence test between the set of the 3 sales indexes and each one of these 2 subsets, first using all the 50 observations and then using only the first 20 observations, for both cases, when using a single data file and when using the 2 data files.

This will be achieved using commands like

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3, 1) (to select variables, using all 50 observations, using a single dataset)
```

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3, 1, 1) (to select variables and observations to be used, using a single dataset)
```

or

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat", , 1) (to select variables, using all 50 observations, using two datasets)
```

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat", , 1, 1) (to select variables and observations to be used, using two datasets)
```

One could also have used a file with the data for the 1st set, say file `JW_Tab9_12_X1.dat` and another file with the data for the 2nd set, say file `JW_Tab9_12_X2.dat` and then use the command

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat")
```

to obtain exactly the same result.

In order to try to evaluate if the relation between the 3 sales indexes is stronger with the first 2 or the last 2 test scores, and also to see how function `PvalDataInd2R` works, let us do an independence test between the set of the 3 sales indexes and each one of these 2 subsets, first using all the 50 observations and then using only the first 20 observations, for both cases, when using a single data file and when using the 2 data files.

This will be achieved using commands like

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3, 1) (to select variables, using all 50 observations, using a single dataset)
```

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3, 1, 1) (to select variables and observations to be used, using a single dataset)
```

or

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat", , 1) (to select variables, using all 50 observations, using two datasets)
```

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat", , 1, 1) (to select variables and observations to be used, using two datasets)
```

One could also have used a file with the data for the 1st set, say file `JW_Tab9_12_X1.dat` and another file with the data for the 2nd set, say file `JW_Tab9_12_X2.dat` and then use the command

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat")
```

to obtain exactly the same result.

In order to try to evaluate if the relation between the 3 sales indexes is stronger with the first 2 or the last 2 test scores, and also to see how function `PvalDataInd2R` works, let us do an independence test between the set of the 3 sales indexes and each one of these 2 subsets, first using all the 50 observations and then using only the first 20 observations, for both cases, when using a single data file and when using the 2 data files.

This will be achieved using commands like

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3, 1) (to select variables, using all 50 observations, using a single dataset)
```

```
> PvalDataInd2R("JW_Tab9_12.dat", "", 3, 1, 1) (to select variables and observations to be used, using a single dataset)
```

or

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat", , 1) (to select variables, using all 50 observations, using two datasets)
```

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2.dat", , 1, 1) (to select variables and observations to be used, using two datasets)
```

To see the use of the parameters **transp1** and **transp2** of the function **PvalDataInd2R** we can carry out a test of independence between the same two sets \underline{X}_1 and \underline{X}_2 we have been using, but using now file **JW_Tab9_12t.dat**, which has the data table in a transposed form. Then we should use the command

```
> PvalDataInd2R("JW_Tab9_12t.dat", "", 3, , , , 1)
```

We can also try to use the file for the 1st part of the table, corresponding to the 3 sales indexes, in a non-transposed form and the file for the 2nd part of the table, corresponding to the 4 test scores, in a transposed form in file **JW_Tab9_12_X2t.dat**, using the command

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2t.dat", , , , , 1)
```


To see the use of the parameters **transp1** and **transp2** of the function **PvalDataInd2R** we can carry out a test of independence between the same two sets \underline{X}_1 and \underline{X}_2 we have been using, but using now file **JW_Tab9_12t.dat**, which has the data table in a transposed form. Then we should use the command

```
> PvalDataInd2R("JW_Tab9_12t.dat", "", 3, , , , 1)
```

We can also try to use the file for the 1st part of the table, corresponding to the 3 sales indexes, in a non-transposed form and the file for the 2nd part of the table, corresponding to the 4 test scores, in a transposed form in file **JW_Tab9_12_X2t.dat**, using the command

```
> PvalDataInd2R("JW_Tab9_12_X1.dat", "JW_Tab9_12_X2t.dat", , , , , 1)
```

Now, for this test, the question is: **what to do when p_1 and p_2 are both odd?**

The answer is once again: one may use near-exact distributions.

These may be either the ones expressly developed for this test, as the ones in

Alberto, R. P., Coelho, C. A. (2007). Study of the quality of several asymptotic and near-exact approximations based on moments for the distribution of the Wilks Lambda statistic. *Journal of Statistical Planning and Inference*, 137, 5, 1612-1626.

Grilo, L. M., Coelho, C. A. (2010) The exact and near-exact distributions for the Wilks Lambda statistic used in the test of independence of two sets of variables. *American Journal of Mathematical and Management Sciences*, 30, 111-145.

or the ones developed for the test of independence of several sets of variables, referred on slide 92.

Now, for this test, the question is: **what to do when p_1 and p_2 are both odd?**

The answer is once again: one may use near-exact distributions.

These may be either the ones expressly developed for this test, as the ones in

Alberto, R. P., Coelho, C. A. (2007). Study of the quality of several asymptotic and near-exact approximations based on moments for the distribution of the Wilks Lambda statistic. *Journal of Statistical Planning and Inference*, 137, 5, 1612-1626.

Grilo, L. M., Coelho, C. A. (2010) The exact and near-exact distributions for the Wilks Lambda statistic used in the test of independence of two sets of variables. *American Journal of Mathematical and Management Sciences*, 30, 111-145.

or the ones developed for the test of independence of several sets of variables, referred on slide 92.

(See subsec 5.1.9.5 in Coelho and Arnold (2019))

Another look at the test of independence of 2 sets of variables

An equivalent way to look at the test of independence of two sets of variables is to see it as the test of fit of the Multivariate Linear Model, where one of the sets of variables, say \underline{X}_1 is the set of response variables, and the other set is the set of explanatory variables. Then in case we reject the null hypothesis of independence of the two sets of variables, we will assume that the model

$$\underset{(p_1 \times 1)}{\underline{X}_1} = \underset{(p_1 \times p_2)}{\beta} \underset{(p_2 \times 1)}{\underline{X}_2} + \underset{(p_1 \times 1)}{E}$$

fits.

This way to look at the test brings us to the so-called *partial Wilks Lambda test*, which is a test between the model above and one of its submodels.

We may write the model above as

$$\underset{(p_1 \times 1)}{\underline{X}_1} = \left[\underset{(p_1 \times p_{21})}{\beta_1} \mid \underset{(p_1 \times p_{22})}{\beta_2} \right] \begin{bmatrix} \underset{(p_{21} \times 1)}{\underline{X}_{21}} \\ \underset{(p_{22} \times 1)}{\underline{X}_{22}} \end{bmatrix} + E$$

with $p_2 = p_{21} + p_{22}$, and then consider the submodel

$$\underset{(p_1 \times 1)}{\underline{X}_1} = \underset{(p_1 \times p_{21})}{\beta_1} \underset{(p_{21} \times 1)}{\underline{X}_{21}} + \underset{(p_1 \times 1)}{E^*}.$$

(See subsec 5.1.9.5 in Coelho and Arnold (2019))

Another look at the test of independence of 2 sets of variables

An equivalent way to look at the test of independence of two sets of variables is to see it as the test of fit of the Multivariate Linear Model, where one of the sets of variables, say \underline{X}_1 is the set of response variables, and the other set is the set of explanatory variables. Then in case we reject the null hypothesis of independence of the two sets of variables, we will assume that the model

$$\underset{(p_1 \times 1)}{\underline{X}_1} = \underset{(p_1 \times p_2)}{\beta} \underset{(p_2 \times 1)}{\underline{X}_2} + \underset{(p_1 \times 1)}{E}$$

fits.

This way to look at the test brings us to the so-called *partial Wilks Lambda test*, which is a test between the model above and one of its submodels.

We may write the model above as

$$\underset{(p_1 \times 1)}{\underline{X}_1} = \left[\underset{(p_1 \times p_{21})}{\beta_1} \mid \underset{(p_1 \times p_{22})}{\beta_2} \right] \begin{bmatrix} \underset{(p_{21} \times 1)}{\underline{X}_{21}} \\ \underset{(p_{22} \times 1)}{\underline{X}_{22}} \end{bmatrix} + E$$

with $p_2 = p_{21} + p_{22}$, and then consider the submodel

$$\underset{(p_1 \times 1)}{\underline{X}_1} = \underset{(p_1 \times p_{21})}{\beta_1} \underset{(p_{21} \times 1)}{\underline{X}_{21}} + \underset{(p_1 \times 1)}{E^*}.$$

(See subsec 5.1.9.5 in Coelho and Arnold (2019))

Another look at the test of independence of 2 sets of variables

An equivalent way to look at the test of independence of two sets of variables is to see it as the test of fit of the Multivariate Linear Model, where one of the sets of variables, say \underline{X}_1 is the set of response variables, and the other set is the set of explanatory variables. Then in case we reject the null hypothesis of independence of the two sets of variables, we will assume that the model

$$\underset{(p_1 \times 1)}{\underline{X}_1} = \underset{(p_1 \times p_2)}{\beta} \underset{(p_2 \times 1)}{\underline{X}_2} + \underset{(p_1 \times 1)}{E}$$

fits.

This way to look at the test brings us to the so-called *partial Wilks Lambda test*, which is a test between the model above and one of its submodels.

We may write the model above as

$$\underset{(p_1 \times 1)}{\underline{X}_1} = \left[\underset{(p_1 \times p_{21})}{\beta_1} \mid \underset{(p_1 \times p_{22})}{\beta_2} \right] \left[\begin{array}{c} \underset{(p_{21} \times 1)}{\underline{X}_{21}} \\ \underset{(p_{22} \times 1)}{\underline{X}_{22}} \end{array} \right] + E$$

with $p_2 = p_{21} + p_{22}$, and then consider the submodel

$$\underset{(p_1 \times 1)}{\underline{X}_1} = \underset{(p_1 \times p_{21})}{\beta_1} \underset{(p_{21} \times 1)}{\underline{X}_{21}} + \underset{(p_1 \times 1)}{E^*}.$$

We may then wonder how can we test between the original model and the submodel.

The answer is: by associating the submodel to the null hypothesis of the test, and the original model to the alternative hypothesis, and test

$$H_0 : \underline{X}_1 = \beta_1 \underline{X}_{21} + E^* \quad \text{vs} \quad H_1 : \underline{X}_1 = \beta \underline{X}_2 + E.$$

Then the LRT statistic is the ratio of the LRT statistic to test the fit of the original model, divided by the LRT statistic to test the fit of the submodel

$$\Lambda = \frac{|A|}{|A_{11}| |A_{22}|} \bigg/ \frac{|A^*|}{|A_{11}| |A_{22,11}|}$$

where, A is the MLE of (the original) Σ , with $p = p_1 + p_2$, and

$$A_{(p_1 \times p_2)} = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array} \right]_{\substack{p_1 \\ p_2}} = \left[\begin{array}{c|c|c} A_{11} & A_{12,1} & A_{12,2} \\ \hline A_{21,1} & A_{22,11} & A_{22,12} \\ \hline A_{21,2} & A_{22,21} & A_{22,22} \end{array} \right]_{\substack{p_1 \\ p_{21} \\ p_{22}}}$$

with

$$A^* = \left[\begin{array}{c|c} A_{11} & A_{12,1} \\ \hline A_{21,1} & A_{22,11} \end{array} \right].$$

Then it may be shown that

$$\Lambda \sim \prod_{j=1}^{p_1} Y_j \sim \prod_{k=1}^{p_{22}} Y_k$$

where, for $n > p_1 + p_2$,

$$Y_j \sim \text{Beta}\left(\frac{n - p_2 - j}{2}, \frac{p_{22}}{2}\right) \quad \text{and} \quad Y_k \sim \text{Beta}\left(\frac{n - p_1 - p_{21} - k}{2}, \frac{p_1}{2}\right)$$

form 2 sets on independent r.v.'s.

So that, the exact distribution of Λ is, for even p_1 , given by Theorems 1 or 2, with

$$m^* = 1, \quad k_1 = 2, \quad a_1 = \frac{n - p_2}{2}, \quad n_1 = \frac{p_1}{2}, \quad m_1 = p_{22},$$

and for even p_{22} , with

$$m^* = 1, \quad k_1 = 2, \quad a_1 = \frac{n - p_1 - p_{21}}{2}, \quad n_1 = \frac{p_{22}}{2}, \quad m_1 = p_1,$$

and its PDF and CDF by Corollaries 1 or 2, in terms of the EGIG PDF and CDF as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p_1+p_{22}-2}; \left\{ \frac{n-2-p_{21}-j}{2} \right\}_{j=1:p_1+p_{22}-2}; p_1 + p_{22} - 2 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p_1+p_{22}-2}; \left\{ \frac{n-2-p_{21}-j}{2} \right\}_{j=1:p_1+p_{22}-2}; p_1 + p_{22} - 2 \right).$$

where

$$r_j = \begin{cases} h_j & j = 1, 2 \\ r_{j-2} + h_j & j = 3, \dots, p_1 + p_{22} - 2, \end{cases}$$

with

$$\begin{aligned} h_j &= \begin{cases} 1, & j = 1, \dots, \min(p_1, p_{22}) \\ 0, & j = 1 + \min(p_1, p_2), \dots, \max(p_1, p_{22}) \\ -1, & j = 1 + \max(p_1, p_{22}), \dots, p_1 + p_{22} - 2 \end{cases} \\ &= (\# \text{ of elements in } \{p_1, p_{22}\} \geq j) - 1, \quad j = 1, \dots, p_1 + p_{22} - 2. \end{aligned}$$

In carrying out this test, if one rejects H_0 it means that the we will stay with the original model, meaning that explanatory variables that were taken away from the model were significant in the model, that is, in a model where the other explanatory variables are present, the contribution of these explanatory variables that were taken away from the model was significant in explaining the variability in the response variables.

And, the other way around, if we do not reject H_0 , then we stay with the submodel, meaning that the explanatory variables that were taken away from the model were not contributing significantly to the explanation of the response variables, in a model where the other explanatory variables are present.

As such, if one wants to test if the scores for "mechanical" and "abstract reasoning" are important in a model where the other 2 scores are present, we may use the command

```
> PvalDataInd2Rp(c(2,3),"JW_Tab9_12.dat","",3)
```

obtaining a very low p-value, indicating that we should reject H_0 , and infer that these 2 scores have a significant effect in explaining the variability in the response variables (the 3 sales performance indexes), in a model where the other 2 scores are already present.

Use the command

```
> EGIG_help(PvalDataInd2Rp)
```

to see how the function `PvalDataInd2Rp` works, in terms of its arguments.

In carrying out this test, if one rejects H_0 it means that the we will stay with the original model, meaning that explanatory variables that were taken away from the model were significant in the model, that is, in a model where the other explanatory variables are present, the contribution of these explanatory variables that were taken away from the model was significant in explaining the variability in the response variables.

And, the other way around, if we do not reject H_0 , then we stay with the submodel, meaning that the explanatory variables that were taken away from the model were not contributing significantly to the explanation of the response variables, in a model where the other explanatory variables are present.

As such, if one wants to test if the scores for "mechanical" and "abstract reasoning" are important in a model where the other 2 scores are present, we may use the command

```
> PvalDataInd2Rp(c(2,3),"JW_Tab9_12.dat","",3)
```

obtaining a very low p-value, indicating that we should reject H_0 , and infer that these 2 scores have a significant effect in explaining the variability in the response variables (the 3 sales performance indexes), in a model where the other 2 scores are already present.

Use the command

```
> EGIG.help(PvalDataInd2Rp)
```

to see how the function `PvalDataInd2Rp` works, in terms of its arguments.

In carrying out this test, if one rejects H_0 it means that the we will stay with the original model, meaning that explanatory variables that were taken away from the model were significant in the model, that is, in a model where the other explanatory variables are present, the contribution of these explanatory variables that were taken away from the model was significant in explaining the variability in the response variables.

And, the other way around, if we do not reject H_0 , then we stay with the submodel, meaning that the explanatory variables that were taken away from the model were not contributing significantly to the explanation of the response variables, in a model where the other explanatory variables are present.

As such, if one wants to test if the scores for "mechanical" and "abstract reasoning" are important in a model where the other 2 scores are present, we may use the command

```
> PvalDataInd2Rp(c(2,3), "JW-Tab9_12.dat", "", 3)
```

obtaining a very low p-value, indicating that we should reject H_0 , and infer that these 2 scores have a significant effect in explaining the variability in the response variables (the 3 sales performance indexes), in a model where the other 2 scores are already present.

Use the command

```
> EGIG.help(PvalDataInd2Rp)
```

to see how the function `PvalDataInd2Rp` works, in terms of its arguments.

In carrying out this test, if one rejects H_0 it means that the we will stay with the original model, meaning that explanatory variables that were taken away from the model were significant in the model, that is, in a model where the other explanatory variables are present, the contribution of these explanatory variables that were taken away from the model was significant in explaining the variability in the response variables.

And, the other way around, if we do not reject H_0 , then we stay with the submodel, meaning that the explanatory variables that were taken away from the model were not contributing significantly to the explanation of the response variables, in a model where the other explanatory variables are present.

As such, if one wants to test if the scores for "mechanical" and "abstract reasoning" are important in a model where the other 2 scores are present, we may use the command

```
> PvalDataInd2Rp(c(2,3), "JW-Tab9_12.dat", "", 3)
```

obtaining a very low p-value, indicating that we should reject H_0 , and infer that these 2 scores have a significant effect in explaining the variability in the response variables (the 3 sales performance indexes), in a model where the other 2 scores are already present.

Use the command

```
> EGIG.help(PvalDataInd2Rp)
```

to see how the function **PvalDataInd2Rp** works, in terms of its arguments.

What can one do when p_1 and p_{22} are both odd?

The answer is: use a near-exact distribution developed using a similar method to that use to obtain near-exact distributions for the test of independence of 2 sets of variables, that is, for the test with acronym **Ind2R**.

Actually, if one looks well, we may see that the distribution of the *partial Wilks* Λ test is indeed similar to that of the test of independence of 2 sets of variables, replacing p_2 by p_{22} and n by $n - p_{21}$, so that,

we can do the same replacement in the near-exact distributions for the test of independence of 2 sets of variables to obtain near-exact distributions for the *partial Wilks* Λ test.

What can one do when p_1 and p_{22} are both odd?

The answer is: use a near-exact distribution developed using a similar method to that use to obtain near-exact distributions for the test of independence of 2 sets of variables, that is, for the test with acronym **Ind2R**.

Actually, if one looks well, we may see that the distribution of the *partial Wilks* Λ test is indeed similar to that of the test of independence of 2 sets of variables, replacing p_2 by p_{22} and n by $n - p_{21}$, so that,

we can do the same replacement in the near-exact distributions for the test of independence of 2 sets of variables to obtain near-exact distributions for the *partial Wilks* Λ test.

What can one do when p_1 and p_{22} are both odd?

The answer is: use a near-exact distribution developed using a similar method to that use to obtain near-exact distributions for the test of independence of 2 sets of variables, that is, for the test with acronym **Ind2R**.

Actually, if one looks well, we may see that the distribution of the *partial Wilks* Λ test is indeed similar to that of the test of independence of 2 sets of variables, replacing p_2 by p_{22} and n by $n - p_{21}$, so that,

we can do the same replacement in the near-exact distributions for the test of independence of 2 sets of variables to obtain near-exact distributions for the *partial Wilks* Λ test.

(See subsec 5.1.11 in Coelho and Arnold (2019))

Let us suppose that $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ and that the random vector \underline{X} is subdivided into m subsets $\underline{X}_k \sim N_{p_k}(\underline{\mu}_k, \Sigma_{kk})$, with $p = \sum_{k=1}^m p_k$,

$$\underline{X} = [\underline{X}'_1, \dots, \underline{X}'_k, \dots, \underline{X}'_m]', \quad \underline{\mu} = [\underline{\mu}'_1, \dots, \underline{\mu}'_k, \dots, \underline{\mu}'_m]'$$

and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1k} & \cdots & \Sigma_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ \Sigma_{k1} & \cdots & \Sigma_{kk} & \cdots & \Sigma_{km} \\ \vdots & & \vdots & \ddots & \vdots \\ \Sigma_{m1} & \cdots & \Sigma_{mk} & \cdots & \Sigma_{mm} \end{bmatrix}.$$

Then, the hypothesis of independence of the m sets of variables may be written as

$$H_0 : \Sigma = bdiag(\Sigma_{11}, \dots, \Sigma_{kk}, \dots, \Sigma_{mm}),$$

and, for a sample of size n , the $(2/n)$ -th power of the LRT statistic is

$$\Lambda = \frac{|A|}{\prod_{k=1}^m |A_{kk}|}$$

where A is the MLE of Σ (or the sample variance-covariance matrix), assumed to be split according to the split of the matrix Σ .

(See subsec 5.1.11 in Coelho and Arnold (2019))

Let us suppose that $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ and that the random vector \underline{X} is subdivided into m subsets $\underline{X}_k \sim N_{p_k}(\underline{\mu}_k, \Sigma_{kk})$, with $p = \sum_{k=1}^m p_k$,

$$\underline{X} = [\underline{X}'_1, \dots, \underline{X}'_k, \dots, \underline{X}'_m]', \quad \underline{\mu} = [\underline{\mu}'_1, \dots, \underline{\mu}'_k, \dots, \underline{\mu}'_m]'$$

and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1k} & \cdots & \Sigma_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ \Sigma_{k1} & \cdots & \Sigma_{kk} & \cdots & \Sigma_{km} \\ \vdots & & \vdots & \ddots & \vdots \\ \Sigma_{m1} & \cdots & \Sigma_{mk} & \cdots & \Sigma_{mm} \end{bmatrix}.$$

Then, the hypothesis of independence of the m sets of variables may be written as

$$H_0 : \Sigma = b \text{diag}(\Sigma_{11}, \dots, \Sigma_{kk}, \dots, \Sigma_{mm}),$$

and, for a sample of size n , the $(2/n)$ -th power of the LRT statistic is

$$\Lambda = \frac{|A|}{\prod_{k=1}^m |A_{kk}|}$$

where A is the MLE of Σ (or the sample variance-covariance matrix), assumed to be split according to the split of the matrix Σ .

We may then show that

$$\Lambda \sim \prod_{k=1}^{m-1} \prod_{j=1}^{p_k} Y_{jk}$$

where, for $n > p$,

$$Y_{jk} \sim \text{Beta} \left(\frac{n - q_k - j}{2}, \frac{q_k}{2} \right),$$

with $q_k = p_{k+1} + \dots + p_m$, form a set of independent r.v.'s.

Then, if all p_k are even, or at most one of them is odd, the exact distribution of Λ is given by Theorem 2, with

$$m^* = m - 1, \quad k_v = 2, \quad n_v = \frac{p_v}{2}, \quad m_v = q_v, \quad a_v = \frac{n - q_v}{2} \quad (v = 1, \dots, m^*),$$

and thus its PDF and CDF are given, in terms of the PDF and CDF of the EGIG distribution, as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p-2}; \left\{ \frac{n-2-j}{2} \right\}_{j=1:p-2}; p-2 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p-2}; \left\{ \frac{n-2-j}{2} \right\}_{j=1:p-2}; p-2 \right).$$

where

$$r_j = \begin{cases} h_j & j = 1, 2 \\ r_{j-2} + h_j & j = 3, \dots, p-2, \end{cases}$$

with

$$h_j = (\# \text{ of } p_k\text{'s } (k = 1, \dots, m) \geq j) - 1, \quad j = 1, \dots, p-2.$$

If one wants to carry out a test of independence of several sets of variables, one may use the file **JW_Tab9_12.dat** and carry out a test of independence among the 3 sets of variables, the 1st one being that of the 3 sales performance indexes, the 2nd one, the one formed by the scores for "creativity" and "mechanical reasoning" and the 3rd one formed by the scores for "abstract reasoning" and "mathematics". This will be done with the command

```
> PvalDataIndR("JW_Tab9_12.dat", c(3, 2, 2))
```

from which we will obtain a p-value of 1.111458×10^{-53} , suggesting a clear rejection of the null hypothesis of independence of these 3 sets of variables.

One may also use function **PvalDataIndR** to carry out tests of independence between only 2 sets of variables. One may use the command

```
> PvalDataIndR("JW_Tab9_12.dat", c(3,4))
```

to obtain exactly the same values that were obtained with the function **PvalDataInd2R**.

And we may also use function **PvalDataIndR** to test the independence of the group formed by the scores for "creativity" and "mechanical reasoning" and the group formed by the scores for "abstract reasoning" and "mathematics", with the command

```
> PvalDataIndR("JW_Tab9_12.dat", c(2,2), 1)
```

selecting the last 4 variables in the dataset, what will give us a p-value of 0.000300, which would lead us to reject the null hypothesis of independence between these 2 groups of variables, for common values of α .

Also for this function, as for all other functions related with the tests being presented, one can use function **EGIG_help** to obtain help on the use of these functions. One can for example use the command

```
> EGIG_help(PvalDataIndR)
```

to obtain help on the use of function **PvalDataIndR**.

One may also use function **PvalDataIndR** to carry out tests of independence between only 2 sets of variables. One may use the command

```
> PvalDataIndR("JW_Tab9_12.dat", c(3,4))
```

to obtain exactly the same values that were obtained with the function **PvalDataInd2R**.

And we may also use function **PvalDataIndR** to test the independence of the group formed by the scores for "creativity" and "mechanical reasoning" and the group formed by the scores for "abstract reasoning" and "mathematics", with the command

```
> PvalDataIndR("JW_Tab9_12.dat", c(2,2), 1)
```

selecting the last 4 variables in the dataset, what will give us a p-value of 0.000300, which would lead us to reject the null hypothesis of independence between these 2 groups of variables, for common values of α .

Also for this function, as for all other functions related with the tests being presented, one can use function **EGIG_help** to obtain help on the use of these functions. One can for example use the command

```
> EGIG_help(PvalDataIndR)
```

to obtain help on the use of function **PvalDataIndR**.

One may also use function **PvalDataIndR** to carry out tests of independence between only 2 sets of variables. One may use the command

```
> PvalDataIndR("JW_Tab9_12.dat", c(3,4))
```

to obtain exactly the same values that were obtained with the function **PvalDataInd2R**.

And we may also use function **PvalDataIndR** to test the independence of the group formed by the scores for "creativity" and "mechanical reasoning" and the group formed by the scores for "abstract reasoning" and "mathematics", with the command

```
> PvalDataIndR("JW_Tab9_12.dat", c(2,2), 1)
```

selecting the last 4 variables in the dataset, what will give us a p-value of 0.000300, which would lead us to reject the null hypothesis of independence between these 2 groups of variables, for common values of α .

Also for this function, as for all other functions related with the tests being presented, one can use function **EGIG_help** to obtain help on the use of these functions. One can for example use the command

```
> EGIG_help(PvalDataIndR)
```

to obtain help on the use of function **PvalDataIndR**.

One other dataset we can use is the dataset in file **winequality-red.dat**. We will use function **PvalDataIndR** to test the independence of the 3 sets of variables formed by variables 1-4 ("fixed_acidity", "volatile_acidity", "citric_acid", "residual_sugar"), variables 5-7 and 10 ("chlorides", "free_sulfur_dioxide", "total_sulfur_dioxide", "sulphates") and variables 8, 9 and 11 ("density", "pH", "alcohol"), using the command

```
> PvalDataIndR("winequality-red.dat",c(4,4,3),1)
```

paying attention to the fact that we will have to reorder the variables to make each of the sets of variables defined by contiguous variables, by using the option **index1** not to really select variables to be used or not used, but rather to reorder them, which is indeed one other possible use of these parameters in all **PvalData<acronym>** functions.

```
> PvalDataIndR("winequality-red.dat",c(4,4,3),1)
```

```
Please insert a list with the numbers of the  
variables you want to keep, separated by spaces:1: 1 2 3 4 5 6 7 10 8 9 11
```

```
Computed value of Lambda: 0.05798291165
```

```
p-value: 9.883470188e-939
```

We may remark that the p-value obtained is extremely low, leading us to clearly reject the hypothesis of independence of these 3 groups of variables.

One other dataset we can use is the dataset in file **winequality-red.dat**. We will use function **PvalDataIndR** to test the independence of the 3 sets of variables formed by variables 1-4 ("fixed_acidity", "volatile_acidity", "citric_acid", "residual_sugar"), variables 5-7 and 10 ("chlorides", "free_sulfur_dioxide", "total_sulfur_dioxide", "sulphates") and variables 8, 9 and 11 ("density", "pH", "alcohol"), using the command

```
> PvalDataIndR("winequality-red.dat",c(4,4,3),1)
```

paying attention to the fact that we will have to reorder the variables to make each of the sets of variables defined by contiguous variables, by using the option **index1** not to really select variables to be used or not used, but rather to reorder them, which is indeed one other possible use of these parameters in all **PvalData<acronym>** functions.

```
> PvalDataIndR("winequality-red.dat",c(4,4,3),1)
```

```
Please insert a list with the numbers of the  
variables you want to keep, separated by spaces:1: 1 2 3 4 5 6 7 10 8 9 11
```

```
Computed value of Lambda: 0.05798291165
```

```
p-value: 9.883470188e-939
```

We may remark that the p-value obtained is extremely low, leading us to clearly reject the hypothesis of independence of these 3 groups of variables.

We may even think about using function **PvalDataIndR** to run a test of independence between the "alcohol" content and variables 1-4 ("fixed_acidity", "volatile_acidity", "citric_acid", "residual_sugar"), with the command

```
> PvalDataIndR("winequality-red.dat", c(4, 1), 1)
```

and choosing variables 1, 2, 3, 4 and 11 for the analysis.

We will obtain a p-value of 7.443351×10^{-23} , which would lead us to reject the null hypothesis of independence of the two groups of variables, that is, the variable "alcohol" content and the other 4 variables, assuming thus that the "alcohol" content is highly related with the values of the other 4 variables.

We may note that:

- ✓ this is the same result that we would have obtained with the command

```
> PvalDataInd2R("winequality-red.dat", "", 4, 1)
```

selecting exactly the same variables that we selected when using function **PvalDataIndR**
- ✓ the test we carried out is indeed just a test to the fit of a multiple regression model where the variable "alcohol" content acts as the response variable and the other 4 variables as the explanatory variables, and in rejecting the null hypothesis of independence of the two groups of variables, we assume the good fit of this regression model, assuming that the 4 explanatory variables explain 'significantly well' the variations in "alcohol" content.

We may even think about using function **PvalDataIndR** to run a test of independence between the "alcohol" content and variables 1-4 ("fixed_acidity", "volatile_acidity", "citric_acid", "residual_sugar"), with the command

```
> PvalDataIndR("winequality-red.dat", c(4, 1), 1)
```

and choosing variables 1, 2, 3, 4 and 11 for the analysis.

We will obtain a p-value of 7.443351×10^{-23} , which would lead us to reject the null hypothesis of independence of the two groups of variables, that is, the variable "alcohol" content and the other 4 variables, assuming thus that the "alcohol" content is highly related with the values of the other 4 variables.

We may note that:

- ✓ this is the same result that we would have obtained with the command

```
> PvalDataInd2R("winequality-red.dat", "", 4, 1)
```

selecting exactly the same variables that we selected when using function **PvalDataIndR**
- ✓ the test we carried out is indeed just a test to the fit of a multiple regression model where the variable "alcohol" content acts as the response variable and the other 4 variables as the explanatory variables, and in rejecting the null hypothesis of independence of the two groups of variables, we assume the good fit of this regression model, assuming that the 4 explanatory variables explain 'significantly well' the variations in "alcohol" content.

The LRT statistic used for this test of independence of several groups of variables is commonly called the generalized Wilks Λ statistic because it may be written as the product of $m - 1$ Wilks Λ statistics which are used to test the independence between \underline{X}_k and the superset of variables formed by sets \underline{X}_{k+1} through \underline{X}_m . That is, we may write for the LRT statistic in this section

$$\Lambda = \prod_{k=1}^m \Lambda_{k,(k+1,\dots,m)}$$

where $\Lambda_{k,(k+1,\dots,m)}$ is the $(2/n)$ -th power of the LRT statistic used to test the independence between \underline{X}_k and the superset formed by sets \underline{X}_{k+1} through \underline{X}_m . Under the null hypothesis of independence of the m sets of variables \underline{X}_k ($k = 1, \dots, m$), the $m - 1$ statistics $\Lambda_{k,(k+1,\dots,m)}$ are independent.

A question that may then arise is: **what can we do when more than 1 set of variables has an odd number of variables?**

The LRT statistic used for this test of independence of several groups of variables is commonly called the generalized Wilks Λ statistic because it may be written as the product of $m - 1$ Wilks Λ statistics which are used to test the independence between \underline{X}_k and the superset of variables formed by sets \underline{X}_{k+1} through \underline{X}_m . That is, we may write for the LRT statistic in this section

$$\Lambda = \prod_{k=1}^m \Lambda_{k,(k+1,\dots,m)}$$

where $\Lambda_{k,(k+1,\dots,m)}$ is the $(2/n)$ -th power of the LRT statistic used to test the independence between \underline{X}_k and the superset formed by sets \underline{X}_{k+1} through \underline{X}_m . Under the null hypothesis of independence of the m sets of variables \underline{X}_k ($k = 1, \dots, m$), the $m - 1$ statistics $\Lambda_{k,(k+1,\dots,m)}$ are independent.

A question that may then arise is: **what can we do when more than 1 set of variables has an odd number of variables?**

The answer is: one may use the near-exact distributions developed for the generalized Wilks Λ statistic in

Coelho, C. A. (2004). The Generalized Near-Integer Gamma distribution – a basis for 'near-exact' approximations to the distributions of statistics which are the product of an odd number of particular independent Beta random variables. *Journal of Multivariate Analysis*, 89, 2, 191-218.

Coelho, C. A., Arnold, B. C., Marques, F. J. (2010) Near-exact distributions for certain likelihood ratio test statistics. *Journal of Statistical Theory and Practice*, 4, 711-725 (invited paper for the special memorial issue in honor of H. C. Gupta)

Marques, F. J., Coelho, C. A., Arnold, B. C. (2011) A general near-exact distribution theory for the most common likelihood ratio test statistics used in Multivariate Analysis. *TEST*, 20, 180-203.

A test for Outliers (real r.v.'s) [OutR]

(See subsec 5.1.13 in Coelho and Arnold (2019))

Let us suppose we have a random sample of size n from $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ and that we are interested in testing whether the $k (< n)$ observations numbered η_1, \dots, η_k (with $\eta_1, \dots, \eta_k \in \{1, \dots, n\}$) should or should not be considered as outliers. We are then interested in testing, for $\eta_1, \dots, \eta_k \in \{1, \dots, n\}$, with $k < n$, the null hypothesis

$$H_0 : \text{observations } \eta_1, \dots, \eta_k \text{ are not outliers.}$$

Wilks (1963) devised a test for outliers, which, as he himself remarks, although not being an LRT test, has many similitudes with such tests, with a test statistic of the form

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

where A is equal to n times the usual MLE of Σ , based on the whole sample, that is, on the n observations, and A^* is equal to $(n - k)$ times the MLE of Σ , based on the $n - k$ observations that remain after removing observations η_1, \dots, η_k .

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

A test for Outliers (real r.v.'s) [OutR]

It is possible to show that, under H_0 ,

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

where, for $n > p + k$,

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

form 2 sets of independent r.v.'s.

As such, the exact distribution of Λ is given by Theorems 1 or 2, for even p , with

$$m^* = 1, \quad k_1 = 2, \quad a_1 = \frac{n-k}{2}, \quad n_1 = \frac{p}{2}, \quad m_1 = k$$

and for even k with

$$m^* = 1, \quad k_1 = 2, \quad a_1 = \frac{n-p}{2}, \quad n_1 = \frac{k}{2}, \quad m_1 = p$$

and thus its PDF and CDF are given by Corollaries 1 or 2, in terms of the EGIG PDF and CDF, as

$$f_{\Lambda}(z) = f^{EGIG}\left(z \mid \{r_j\}_{j=1:p+k-2}; \left\{\frac{n-2-j}{2}\right\}_{j=1:p+k-2}; p+k-2\right)$$

and

$$F_{\Lambda}(z) = f^{EGIG}\left(z \mid \{r_j\}_{j=1:p+k-2}; \left\{\frac{n-2-j}{2}\right\}_{j=1:p+k-2}; p+k-2\right).$$

A test for Outliers (real r.v.'s) [OutR]

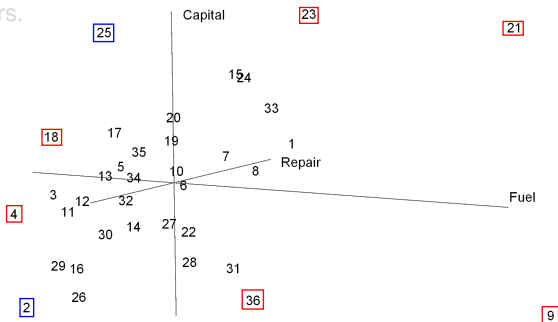
Let us run an outlier test using the data in Table 6.10 of Johnson and Wichern (2014), where data on "Fuel", "Repair" and "Capital" costs were obtained for 36 gasoline trucks used in the distribution of milk. The data is in file **Trucks.dat**

Let us suppose we chose randomly observations (trucks) 4 and 18 to be tested as outliers. Then, we should use the command

```
> PvalDataOutR("Trucks.dat", c(4, 18))
```

obtaining a p-value of 0.395735, which would lead us to not reject the null hypothesis that these 2 points are not outliers.

Now let us, by looking at the cloud of points, choose (not randomly) points (trucks) 9 and 21 to be tested as outliers.



A test for Outliers (real r.v.'s) [OutR]

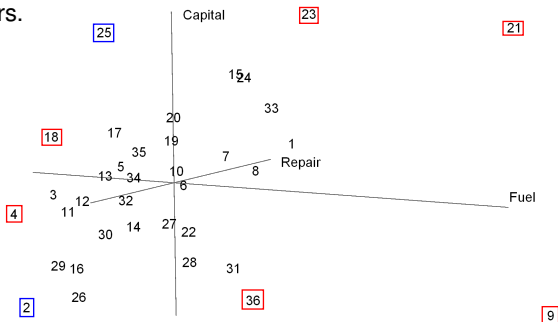
Let us run an outlier test using the data in Table 6.10 of Johnson and Wichern (2014), where data on "Fuel", "Repair" and "Capital" costs were obtained for 36 gasoline trucks used in the distribution of milk. The data is in file **Trucks.dat**

Let us suppose we chose randomly observations (trucks) 4 and 18 to be tested as outliers. Then, we should use the command

```
> PvalDataOutR("Trucks.dat", c(4, 18))
```

obtaining a p-value of 0.395735, which would lead us to not reject the null hypothesis that these 2 points are not outliers.

Now let us, by looking at the cloud of points, choose (not randomly) points (trucks) 9 and 21 to be tested as outliers.



A test for Outliers (real r.v.'s) [OutR]

The command

```
> PvalDataOutR("Trucks.dat",c(9,21))
```

gives us a p-value of 3.024914×10^{-7} , a value that being quite low may at first sight seem to lead us to clearly reject the null hypothesis that these two points are not outliers (thus, considering them as outliers).

However, as Wilks remarks, since we chose these 2 points not randomly but by looking at the cloud of points, if we were planning to use an α value of 0.05, we should really rather use an α value of $\alpha/\binom{n}{k}$, [(why?)] which in this case would be equal to $0.05/\binom{36}{2} = 7.936508 \times 10^{-5}$. Anyway, we should really reject the null hypothesis and consider these 2 points as outliers.

If we run a test for outliers for the 4 points, 9, 21, 36 and 23, with the command

```
> PvalDataOutR("Trucks.dat",c(9,21,36,23))
```

we obtain a p-value of 5.477198×10^{-8} , to be compared with $0.05/\binom{36}{4} = 8.488244 \times 10^{-7}$ for an initial value of $\alpha = 0.05$, while if we test only points 36 and 23 we obtain a p-value of 0.145323. (Try to reason about the values obtained!)

Also test the 2 points 2 and 25 together, and then together with points 36 and 23, and discuss the values obtained.

A test for Outliers (real r.v.'s) [OutR]

The command

```
> PvalDataOutR("Trucks.dat",c(9,21))
```

gives us a p-value of 3.024914×10^{-7} , a value that being quite low may at first sight seem to lead us to clearly reject the null hypothesis that these two points are not outliers (thus, considering them as outliers).

However, as Wilks remarks, since we chose these 2 points not randomly but by looking at the cloud of points, if we were planning to use an α value of 0.05, we should really rather use an α value of $\alpha/\binom{n}{k}$, [(why?)] which in this case would be equal to $0.05/\binom{36}{2} = 7.936508 \times 10^{-5}$. Anyway, we should really reject the null hypothesis and consider these 2 points as outliers.

If we run a test for outliers for the 4 points, 9, 21, 36 and 23, with the command

```
> PvalDataOutR("Trucks.dat",c(9,21,36,23))
```

we obtain a p-value of 5.477198×10^{-8} , to be compared with

$0.05/\binom{36}{4} = 8.488244 \times 10^{-7}$ for an initial value of $\alpha = 0.05$, while if we test only points 36 and 23 we obtain a p-value of 0.145323. (Try to reason about the values obtained!)

Also test the 2 points 2 and 25 together, and then together with points 36 and 23, and discuss the values obtained.

A test for Outliers (real r.v.'s) [OutR]

The command

```
> PvalDataOutR("Trucks.dat", c(9, 21))
```

gives us a p-value of 3.024914×10^{-7} , a value that being quite low may at first sight seem to lead us to clearly reject the null hypothesis that these two points are not outliers (thus, considering them as outliers).

However, as Wilks remarks, since we chose these 2 points not randomly but by looking at the cloud of points, if we were planning to use an α value of 0.05, we should really rather use an α value of $\alpha / \binom{n}{k}$, [(why?)] which in this case would be equal to $0.05 / \binom{36}{2} = 7.936508 \times 10^{-5}$. Anyway, we should really reject the null hypothesis and consider these 2 points as outliers.

If we run a test for outliers for the 4 points, 9, 21, 36 and 23, with the command

```
> PvalDataOutR("Trucks.dat", c(9, 21, 36, 23))
```

we obtain a p-value of 5.477198×10^{-8} , to be compared with $0.05 / \binom{36}{4} = 8.488244 \times 10^{-7}$ for an initial value of $\alpha = 0.05$, while if we test only points 36 and 23 we obtain a p-value of 0.145323. (Try to reason about the values obtained!)

Also test the 2 points 2 and 25 together, and then together with points 36 and 23, and discuss the values obtained.

A test for Outliers (real r.v.'s) [OutR]

The command

```
> PvalDataOutR("Trucks.dat", c(9, 21))
```

gives us a p-value of 3.024914×10^{-7} , a value that being quite low may at first sight seem to lead us to clearly reject the null hypothesis that these two points are not outliers (thus, considering them as outliers).

However, as Wilks remarks, since we chose these 2 points not randomly but by looking at the cloud of points, if we were planning to use an α value of 0.05, we should really rather use an α value of $\alpha / \binom{n}{k}$, [(why?)] which in this case would be equal to $0.05 / \binom{36}{2} = 7.936508 \times 10^{-5}$. Anyway, we should really reject the null hypothesis and consider these 2 points as outliers.

If we run a test for outliers for the 4 points, 9, 21, 36 and 23, with the command

```
> PvalDataOutR("Trucks.dat", c(9, 21, 36, 23))
```

we obtain a p-value of 5.477198×10^{-8} , to be compared with $0.05 / \binom{36}{4} = 8.488244 \times 10^{-7}$ for an initial value of $\alpha = 0.05$, while if we test only points 36 and 23 we obtain a p-value of 0.145323. (Try to reason about the values obtained!)

Also test the 2 points 2 and 25 together, and then together with points 36 and 23, and discuss the values obtained.

A test for Outliers (real r.v.'s) [OutR]

What can one do if p and k are both odd?

I think that by now you already guessed the answer! Yes, we can use near-exact distributions, namely we can use the ones for the test of independence of 2 sets of variables, replacing p_1 by p and p_2 by k (or vice-versa), which is quite interesting.

We did not mention it yet, but one other possible approximation to the distribution of the statistics addressed so far, which works also very well and has a bit of a simpler computation is 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

$$\Lambda = \frac{|A|}{|A+B|}, \quad (1)$$

then

$$\frac{ms-2\lambda}{pq} \frac{1-\Lambda^{1/s}}{\Lambda^{1/s}} \stackrel{a}{\sim} F_{pq, ms-2\lambda} \quad (2)$$

where

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

A test for Outliers (real r.v.'s) [OutR]

What can one do if p and k are both odd?

I think that by now you already guessed the answer! Yes, we can use near-exact distributions, namely we can use the ones for the test of independence of 2 sets of variables, replacing p_1 by p and p_2 by k (or vice-versa), which is quite interesting.

We did not mention it yet, but one other possible approximation to the distribution of the statistics addressed so far, which works also very well and has a bit of a simpler computation is 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

$$\Lambda = \frac{|A|}{|A + B|}, \quad (1)$$

then

$$\frac{ms - 2\lambda}{pq} \frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \stackrel{a}{\sim} F_{pq, ms - 2\lambda} \quad (2)$$

where

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

A test for Outliers (real r.v.'s) [OutR]

What can one do if p and k are both odd?

I think that by now you already guessed the answer! Yes, we can use near-exact distributions, namely we can use the ones for the test of independence of 2 sets of variables, replacing p_1 by p and p_2 by k (or vice-versa), which is quite interesting.

We did not mention it yet, but one other possible approximation to the distribution of the statistics addressed so far, which works also very well and has a bit of a simpler computation is 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

$$\Lambda = \frac{|A|}{|A + B|}, \quad (1)$$

then

$$\frac{ms - 2\lambda}{pq} \frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \stackrel{a}{\sim} F_{pq, ms - 2\lambda} \quad (2)$$

where

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

and

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

And indeed, if you check the sections of Coelho and Arnold (2019), indicated at the beginning of each test, you may see that all the LRT statistics addressed so far, as well as the Wilks statistic for the test of outliers, may be written in the form of (1).

When using this asymptotic distribution one should then reject the corresponding null hypothesis if the transformation of the computed value of Λ through the left hand side of (2) exceeds the $1 - \alpha$ quantile of the F distribution on the right hand side of (2).

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.

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

and

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

And indeed, if you check the sections of Coelho and Arnold (2019), indicated at the beginning of each test, you may see that all the LRT statistics addressed so far, as well as the Wilks statistic for the test of outliers, may be written in the form of (1).

When using this asymptotic distribution one should then reject the corresponding null hypothesis if the transformation of the computed value of Λ through the left hand side of (2) exceeds the $1 - \alpha$ quantile of the F distribution on the right hand side of (2).

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.

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

and

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

And indeed, if you check the sections of Coelho and Arnold (2019), indicated at the beginning of each test, you may see that all the LRT statistics addressed so far, as well as the Wilks statistic for the test of outliers, may be written in the form of (1).

When using this asymptotic distribution one should then reject the corresponding null hypothesis if the transformation of the computed value of Λ through the left hand side of (2) exceeds the $1 - \alpha$ quantile of the F distribution on the right hand side of (2).

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.

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

and

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

And indeed, if you check the sections of Coelho and Arnold (2019), indicated at the beginning of each test, you may see that all the LRT statistics addressed so far, as well as the Wilks statistic for the test of outliers, may be written in the form of (1).

When using this asymptotic distribution one should then reject the corresponding null hypothesis if the transformation of the computed value of Λ through the left hand side of (2) exceeds the $1 - \alpha$ quantile of the F distribution on the right hand side of (2).

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.

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

A test for Outliers (complex r.v.'s) [OutC]

(See subsec 5.1.14 in Coelho and Arnold (2019))

Let us suppose that now we have a random sample of size n from $\underline{X} \sim CN_p(\underline{\mu}, \Sigma)$ and that we are interested in testing a similar hypothesis to that in the real case, that is, to test the null hypothesis

$$H_0 : \text{observations } \eta_1, \dots, \eta_k \text{ are not outliers.}$$

We will still use the statistic developed by Wilks for the real case, still with

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

where, for $n > p + k$,

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

form 2 sets of independent r.v.'s.

As such, the exact distribution of Λ is given by Theorems 1 or 2, for any p and any k , with

$$m^* = 1, \quad k_1 = 1, \quad a_1 = n - k, \quad n_1 = p, \quad m_1 = k$$

or

$$m^* = 1, \quad k_1 = 2, \quad a_1 = n - p, \quad n_1 = k, \quad m_1 = p$$

and thus its PDF and CDF are given by Corollaries 1 or 2, in terms of the EGIG PDF and CDF, as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+k-1}; \left\{ \frac{n-1-j}{2} \right\}_{j=1:p+k-1}; p+k-1 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:p+k-1}; \left\{ \frac{n-1-j}{2} \right\}_{j=1:p+k-1}; p+k-1 \right).$$

where

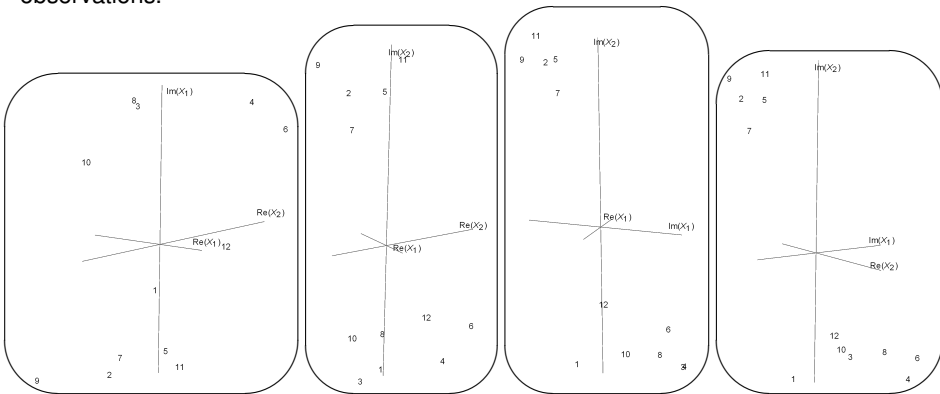
$$r_j = \begin{cases} h_j, & j = 1 \\ r_{j-1} + h_j, & j = 2, \dots, p+k-1 \end{cases}$$

with

$$h_j = (\# \text{ of elements in } \{p, k\} \geq j) - 1, \quad j = 1, \dots, p+k-1.$$

A test for Outliers (complex r.v.'s) [OutC]

Let us take the 12 observations in the file **ex1_comp.dat** for the overall set of the 3 samples and consider the 4 plots of the real and imaginary parts for this 12 observations.



The group of the 5 observations 2, 5, 7, 9 and 11 seems to fall quite apart from the other observations. So, let us test if we should consider this group as a group of outliers.

if we use the command

```
> PvalDataOutC("ex1_comp.dat",c(2,5,7,9,11))
```

we obtain a p-value of 0.001301, which is indeed quite small. However, since we chose these 5 points by looking at the plot of points, and not randomly, we should, if we use for example an α -value of 0.05, actually compare this p-value with the value $0.05 / \binom{12}{5} = 6.3131 \times 10^{-5}$, concluding that indeed we should not consider these 5 points as outliers.

(See subsec 5.1.16 in Coelho and Arnold (2019))

Let us suppose that we have $\underline{X} = [\underline{X}'_1, \underline{X}'_2] \sim N_{2p}(\underline{\mu}, \Sigma)$, where both subvectors \underline{X}_1 and \underline{X}_2 are p -dimensional, with

$$\underline{\mu} = \begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix}, \quad \text{with} \quad \underline{\mu}_1 = E(\underline{X}_1) \quad \text{and} \quad \underline{\mu}_2 = E(\underline{X}_2),$$

and that we want to test the null hypothesis

$$H_0 : \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{bmatrix} \quad \text{and} \quad \underline{\mu}_1 = \underline{\mu}_2,$$

where Σ_1 and Σ_2 are non-specified, but with Σ_1 , $\Sigma_1 + \Sigma_2$ and $\Sigma_1 - \Sigma_2$ positive-definite.

We may note that for

$$\Gamma = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}, \quad \Gamma \otimes I_p = \left[\begin{array}{c|c} \frac{1}{\sqrt{2}} I_p & \frac{1}{\sqrt{2}} I_p \\ \hline \frac{1}{\sqrt{2}} I_p & -\frac{1}{\sqrt{2}} I_p \end{array} \right]$$

under H_0 we have

$$\Sigma^* = (\Gamma \otimes I_p) \Sigma (\Gamma \otimes I_p)' = \begin{bmatrix} \Sigma_1 + \Sigma_2 & 0 \\ 0 & \Sigma_1 - \Sigma_2 \end{bmatrix} = \text{Var}((\Gamma \otimes I_p) \underline{X}).$$

(See subsec 5.1.16 in Coelho and Arnold (2019))

Let us suppose that we have $\underline{X} = [\underline{X}'_1, \underline{X}'_2] \sim N_{2p}(\underline{\mu}, \Sigma)$, where both subvectors \underline{X}_1 and \underline{X}_2 are p -dimensional, with

$$\underline{\mu} = \begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix}, \quad \text{with} \quad \underline{\mu}_1 = E(\underline{X}_1) \quad \text{and} \quad \underline{\mu}_2 = E(\underline{X}_2),$$

and that we want to test the null hypothesis

$$H_0 : \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{bmatrix} \quad \text{and} \quad \underline{\mu}_1 = \underline{\mu}_2,$$

where Σ_1 and Σ_2 are non-specified, but with Σ_1 , $\Sigma_1 + \Sigma_2$ and $\Sigma_1 - \Sigma_2$ positive-definite.

We may note that for

$$\Gamma = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}, \quad \Gamma \otimes I_p = \left[\begin{array}{c|c} \frac{1}{\sqrt{2}} I_p & \frac{1}{\sqrt{2}} I_p \\ \hline \frac{1}{\sqrt{2}} I_p & -\frac{1}{\sqrt{2}} I_p \end{array} \right]$$

under H_0 we have

$$\Sigma^* = (\Gamma \otimes I_p) \Sigma (\Gamma \otimes I_p)' = \begin{bmatrix} \Sigma_1 + \Sigma_2 & 0 \\ 0 & \Sigma_1 - \Sigma_2 \end{bmatrix} = \text{Var}((\Gamma \otimes I_p) \underline{X}).$$

So that, under H_0 ,

$$\underline{X}^* = (\Gamma \otimes I_p) \underline{X} \sim N_{2p}(\underline{\mu}^*, \Sigma^*)$$

where

$$\underline{\mu}^* = \begin{bmatrix} \underline{\mu}_1^* \\ \underline{\mu}_2^* \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} (\underline{\mu}_1 + \underline{\mu}_2) \\ \frac{1}{\sqrt{2}} (\underline{\mu}_2 - \underline{\mu}_1) \end{bmatrix} \quad \text{and} \quad \Sigma^* = \begin{bmatrix} \Sigma_1^* & 0 \\ 0 & \Sigma_2^* \end{bmatrix} = \begin{bmatrix} \Sigma_1 + \Sigma_2 & 0 \\ 0 & \Sigma_1 - \Sigma_2 \end{bmatrix}$$

As such, testing

$$H_0 : \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{bmatrix} \quad \text{and} \quad \underline{\mu}_1 = \underline{\mu}_2,$$

is equivalent to test

$$H_0 : \Sigma^* = b \text{diag}(\Sigma_1^*, \Sigma_2^*), \quad \underline{\mu}_2^* = \underline{0}.$$

Then, for a sample of size n , the $(2/n)$ -th power of the LRT statistic is

$$\Lambda = \frac{|A^*|}{|A_{11}^*| |A_{22}^* + \frac{1}{n} \bar{X}_2 \bar{X}_2^{*T}|},$$

So that, under H_0 ,

$$\underline{X}^* = (\Gamma \otimes I_p) \underline{X} \sim N_{2p}(\underline{\mu}^*, \Sigma^*)$$

where

$$\underline{\mu}^* = \begin{bmatrix} \underline{\mu}_1^* \\ \underline{\mu}_2^* \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} (\underline{\mu}_1 + \underline{\mu}_2) \\ \frac{1}{\sqrt{2}} (\underline{\mu}_2 - \underline{\mu}_1) \end{bmatrix} \quad \text{and} \quad \Sigma^* = \begin{bmatrix} \Sigma_1^* & 0 \\ 0 & \Sigma_2^* \end{bmatrix} = \begin{bmatrix} \Sigma_1 + \Sigma_2 & 0 \\ 0 & \Sigma_1 - \Sigma_2 \end{bmatrix}$$

As such, testing

$$H_0 : \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{bmatrix} \quad \text{and} \quad \underline{\mu}_1 = \underline{\mu}_2,$$

is equivalent to test

$$H_0 : \Sigma^* = b \text{diag}(\Sigma_1^*, \Sigma_2^*), \quad \underline{\mu}_2^* = \underline{0}.$$

Then, for a sample of size n , the $(2/n)$ -th power of the LRT statistic is

$$\Lambda = \frac{|A^*|}{|A_{11}^*| |A_{22}^* + \frac{1}{n} \underline{\bar{X}}_2 \underline{\bar{X}}_2^{*'}|},$$

where

$$A^* = (\Gamma \otimes I_p) A (\Gamma \otimes I_p)' = \begin{bmatrix} A_{11}^* & A_{12}^* \\ A_{21}^* & A_{22}^* \end{bmatrix}$$

for

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

where X is the $n \times (2p)$ data matrix, and where

$$\bar{X}_2^* = \frac{1}{n} X_2^{*'} E_{n1} \quad \text{for} \quad X^* = X (\Gamma \otimes I_p) = [X_1^* | X_2^*],$$

where E_{nm} represents a matrix of 1's, with the given dimensions and X_1^* and X_2^* are both $n \times p$.

We may show that, for $n > p$,

$$\Lambda \sim \prod_{j=1}^{p+1} Y_j \sim \prod_{j=1}^p Y_j^*$$

where

$$Y_j \sim \text{Beta}\left(\frac{n-p-j+1}{2}, \frac{p}{2}\right) \quad \text{and} \quad Y_j^* \sim \text{Beta}\left(\frac{n-p-j}{2}, \frac{p+1}{2}\right)$$

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

As such the exact distribution of Λ is given by Theorems 1 or 2, with

$$m^* = 1, \quad k_1 = 2, \quad a_1 = \frac{n-p+1}{2}, \quad n_1 = \frac{p+1}{2}, \quad \text{and} \quad m_1 = p,$$

or

$$m^* = 1, \quad k_1 = 2, \quad a_1 = \frac{n-p}{2}, \quad n_1 = \frac{p}{2}, \quad \text{and} \quad m_1 = p+1,$$

and thus the exact PDF and CDF of Λ are given by Corollaries 1 or 2, through the PDF and CDF of the EGIG distribution, as

$$f_{\Lambda}(z) = f^{EGIG}\left(z \mid \{r_j\}_{j=1:2p-1}; \left\{\frac{n-1-j}{2}\right\}_{j=1:2p-1}; 2p-1\right)$$

and

$$F_{\Lambda}(z) = f^{EGIG}\left(z \mid \{r_j\}_{j=1:2p-1}; \left\{\frac{n-1-j}{2}\right\}_{j=1:2p-1}; 2p-1\right).$$

We may show that, for $n > p$,

$$\Lambda \sim \prod_{j=1}^{p+1} Y_j \sim \prod_{j=1}^p Y_j^*$$

where

$$Y_j \sim \text{Beta}\left(\frac{n-p-j+1}{2}, \frac{p}{2}\right) \quad \text{and} \quad Y_j^* \sim \text{Beta}\left(\frac{n-p-j}{2}, \frac{p+1}{2}\right)$$

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

As such the exact distribution of Λ is given by Theorems 1 or 2, with

$$m^* = 1, \quad k_1 = 2, \quad a_1 = \frac{n-p+1}{2}, \quad n_1 = \frac{p+1}{2}, \quad \text{and} \quad m_1 = p,$$

or

$$m^* = 1, \quad k_1 = 2, \quad a_1 = \frac{n-p}{2}, \quad n_1 = \frac{p}{2}, \quad \text{and} \quad m_1 = p+1,$$

and thus the exact PDF and CDF of Λ are given by Corollaries 1 or 2, through the PDF and CDF of the EGIG distribution, as

$$f_{\Lambda}(z) = f^{EGIG}\left(z \mid \{r_j\}_{j=1:2p-1}; \left\{\frac{n-1-j}{2}\right\}_{j=1:2p-1}; 2p-1\right)$$

and

$$F_{\Lambda}(z) = f^{EGIG}\left(z \mid \{r_j\}_{j=1:2p-1}; \left\{\frac{n-1-j}{2}\right\}_{j=1:2p-1}; 2p-1\right).$$

To exemplify the application of this test we will use the data in Table 1.8 (the first 24 observations) and Table 6.16 of Johnson and Wichern (2014), placed side by side to form our X matrix, thus formed by measurements of bone mineral content taken on the dominant and nondominant radius, humerus, and ulna of 24 older women, before (Table 1.18) and after (Table 6.16) a treatment.

The data is in file `Tab1.8_6.16_JW.dat`, and the sample variance-covariance matrix is

0.013	0.010	0.023	0.021	0.009	0.008	0.013	0.011	0.026	0.020	0.009	0.009
0.010	0.011	0.018	0.021	0.008	0.008	0.011	0.011	0.022	0.022	0.008	0.009
0.023	0.018	0.082	0.069	0.016	0.012	0.026	0.019	0.088	0.072	0.019	0.015
0.021	0.021	0.069	0.072	0.018	0.016	0.023	0.023	0.077	0.076	0.021	0.020
0.009	0.008	0.016	0.018	0.011	0.007	0.009	0.009	0.021	0.018	0.011	0.007
0.008	0.008	0.012	0.016	0.007	0.010	0.009	0.009	0.016	0.017	0.007	0.010
0.013	0.011	0.026	0.023	0.009	0.009	0.016	0.011	0.029	0.024	0.009	0.009
0.011	0.011	0.019	0.023	0.009	0.009	0.011	0.012	0.023	0.024	0.009	0.010
0.026	0.022	0.088	0.077	0.021	0.016	0.029	0.023	0.105	0.084	0.024	0.018
0.020	0.022	0.072	0.076	0.018	0.017	0.024	0.024	0.084	0.084	0.021	0.020
0.009	0.008	0.019	0.021	0.011	0.007	0.009	0.009	0.024	0.021	0.012	0.008
0.009	0.009	0.015	0.020	0.007	0.010	0.009	0.010	0.018	0.020	0.008	0.013

Further, the two sample mean vectors, for the 1st and 2nd groups of 6 variables, that is, for before and after the treatment are

$$\bar{X}'_1 = [0.84083, 0.81342, 1.78525, 1.72925, 0.69754, 0.68658]$$

and

$$\bar{X}'_2 = [0.84096, 0.81017, 1.77808, 1.71692, 0.71267, 0.68675]$$

To exemplify the application of this test we will use the data in Table 1.8 (the first 24 observations) and Table 6.16 of Johnson and Wichern (2014), placed side by side to form our X matrix, thus formed by measurements of bone mineral content taken on the dominant and nondominant radius, humerus, and ulna of 24 older women, before (Table 1.18) and after (Table 6.16) a treatment.

The data is in file **Tab1.8_6.16_JW.dat**, and the sample variance-covariance matrix is

0.013	0.010	0.023	0.021	0.009	0.008	0.013	0.011	0.026	0.020	0.009	0.009
0.010	0.011	0.018	0.021	0.008	0.008	0.011	0.011	0.022	0.022	0.008	0.009
0.023	0.018	0.082	0.069	0.016	0.012	0.026	0.019	0.088	0.072	0.019	0.015
0.021	0.021	0.069	0.072	0.018	0.016	0.023	0.023	0.077	0.076	0.021	0.020
0.009	0.008	0.016	0.018	0.011	0.007	0.009	0.009	0.021	0.018	0.011	0.007
0.008	0.008	0.012	0.016	0.007	0.010	0.009	0.009	0.016	0.017	0.007	0.010
0.013	0.011	0.026	0.023	0.009	0.009	0.016	0.011	0.029	0.024	0.009	0.009
0.011	0.011	0.019	0.023	0.009	0.009	0.011	0.012	0.023	0.024	0.009	0.010
0.026	0.022	0.088	0.077	0.021	0.016	0.029	0.023	0.105	0.084	0.024	0.018
0.020	0.022	0.072	0.076	0.018	0.017	0.024	0.024	0.084	0.084	0.021	0.020
0.009	0.008	0.019	0.021	0.011	0.007	0.009	0.009	0.024	0.021	0.012	0.008
0.009	0.009	0.015	0.020	0.007	0.010	0.009	0.010	0.018	0.020	0.008	0.013

Further, the two sample mean vectors, for the 1st and 2nd groups of 6 variables, that is, for before and after the treatment are

$$\bar{X}_1' = [0.84083, 0.81342, 1.78525, 1.72925, 0.69754, 0.68658]$$

and

$$\bar{X}_2' = [0.84096, 0.81017, 1.77808, 1.71692, 0.71267, 0.68675]$$

To exemplify the application of this test we will use the data in Table 1.8 (the first 24 observations) and Table 6.16 of Johnson and Wichern (2014), placed side by side to form our X matrix, thus formed by measurements of bone mineral content taken on the dominant and nondominant radius, humerus, and ulna of 24 older women, before (Table 1.18) and after (Table 6.16) a treatment.

The data is in file **Tab1.8_6.16_JW.dat**, and the sample variance-covariance matrix is

0.013	0.010	0.023	0.021	0.009	0.008	0.013	0.011	0.026	0.020	0.009	0.009
0.010	0.011	0.018	0.021	0.008	0.008	0.011	0.011	0.022	0.022	0.008	0.009
0.023	0.018	0.082	0.069	0.016	0.012	0.026	0.019	0.088	0.072	0.019	0.015
0.021	0.021	0.069	0.072	0.018	0.016	0.023	0.023	0.077	0.076	0.021	0.020
0.009	0.008	0.016	0.018	0.011	0.007	0.009	0.009	0.021	0.018	0.011	0.007
0.008	0.008	0.012	0.016	0.007	0.010	0.009	0.009	0.016	0.017	0.007	0.010
0.013	0.011	0.026	0.023	0.009	0.009	0.016	0.011	0.029	0.024	0.009	0.009
0.011	0.011	0.019	0.023	0.009	0.009	0.011	0.012	0.023	0.024	0.009	0.010
0.026	0.022	0.088	0.077	0.021	0.016	0.029	0.023	0.105	0.084	0.024	0.018
0.020	0.022	0.072	0.076	0.018	0.017	0.024	0.024	0.084	0.084	0.021	0.020
0.009	0.008	0.019	0.021	0.011	0.007	0.009	0.009	0.024	0.021	0.012	0.008
0.009	0.009	0.015	0.020	0.007	0.010	0.009	0.010	0.018	0.020	0.008	0.013

Further, the two sample mean vectors, for the 1st and 2nd groups of 6 variables, that is, for before and after the treatment are

$$\bar{X}'_1 = [0.84083, 0.81342, 1.78525, 1.72925, 0.69754, 0.68658]$$

and

$$\bar{X}'_2 = [0.84096, 0.81017, 1.77808, 1.71692, 0.71267, 0.68675]$$

The command

```
> PvalDataCompSymEqR("Tab1_8_6_16_JW.dat")
```

gives a p-value of 0.271746, showing that we should not reject the null hypothesis of complete Symmetrical Equivalence of the two sets of variables (before and after treatment).

This fact will lead us to say that the two sets of bone mineral content variables are "completely symmetrically equivalent", in the sense that using one of them or the other in any statistical analysis will generally lead to similar conclusions.

(See subsec 5.1.22 in Coelho and Arnold (2019))

Let us assume a similar setup to the one of our 1st test, that is, let us suppose that $\underline{X}_k \sim N_p(\underline{\mu}_k, \Sigma)$ ($k = 1, \dots, q$), where now Σ is a circular matrix, and that we have q independent samples, one from each \underline{X}_k , with sizes n_k , and that we are interested in testing the null hypothesis

$$H_0 : \underline{\mu}_1 = \dots = \underline{\mu}_k = \dots = \underline{\mu}_q.$$

A $p \times p$ matrix Σ is said to be circular or circulant if for $i, j = 1, \dots, p$, and $k = 0, \dots, p-1$,

$$\Sigma = [\sigma_{ij}], \quad \text{with } \sigma_{i,i+k} = \sigma_{i+k,i} = \text{Cov}(X_i, X_{i+k}) = \sigma_0^2 \rho_k,$$

where $\rho_0 = \text{Corr}(X_i, X_i) = 1$ and for $i = 1, \dots, p-1$ and $k = 1, \dots, p-i$,
 $\rho_k = \text{Corr}(X_i, X_{i+k}) = \text{Corr}(X_i, X_{i+p-k}) = \rho_{p-k}.$

(See subsec 5.1.22 in Coelho and Arnold (2019))

Let us assume a similar setup to the one of our 1st test, that is, let us suppose that $\underline{X}_k \sim N_p(\underline{\mu}_k, \Sigma)$ ($k = 1, \dots, q$), where now Σ is a circular matrix, and that we have q independent samples, one from each \underline{X}_k , with sizes n_k , and that we are interested in testing the null hypothesis

$$H_0 : \underline{\mu}_1 = \dots = \underline{\mu}_k = \dots = \underline{\mu}_q.$$

A $p \times p$ matrix Σ is said to be circular or circulant if for $i, j = 1, \dots, p$, and $k = 0, \dots, p-1$,

$$\Sigma = [\sigma_{ij}], \quad \text{with } \sigma_{i,i+k} = \sigma_{i+k,i} = \text{Cov}(X_i, X_{i+k}) = \sigma_0^2 \rho_k,$$

where $\rho_0 = \text{Corr}(X_i, X_i) = 1$ and for $i = 1, \dots, p-1$ and $k = 1, \dots, p-i$,
 $\rho_k = \text{Corr}(X_i, X_{i+k}) = \text{Corr}(X_i, X_{i+p-k}) = \rho_{p-k}.$

For example, for $p = 6$ and $p = 7$ we have respectively,

$$\Sigma = \sigma_0^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_2 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_2 & \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}, \quad \Sigma = \sigma_0^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_3 & \rho_2 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_3 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_3 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_3 & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_3 & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_2 & \rho_3 & \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}.$$

For $n = \sum_{k=1}^q n_k$, the $(2/n)$ -th power of the LRT statistic for the test to H_0 is

$$\Lambda = \prod_{j=1}^p \frac{v_j^*}{v_j^{**}}$$

where, for $m = \lfloor p/2 \rfloor$,

$$v_j^* = \begin{cases} a_{jj}^{**}, & j = 1 \text{ and } j = m + 1 \text{ if } p \text{ is even} \\ (a_{jj}^{**} + a_{p-j+2, p-j+2}^{**})/2, & j = 2, \dots, p - m; m + 2, \dots, p \end{cases}$$

and

$$v_j^{**} = \begin{cases} c_{jj}^{**}, & j = 1 \text{ and } j = m + 1 \text{ if } p \text{ is even} \\ (c_{jj}^{**} + c_{p-j+2, p-j+2}^{**})/2, & j = 2, \dots, p - m, m + 2, \dots, p \end{cases}$$

where a_{jj}^{**} and c_{jj}^{**} are the j -th diagonal elements of the matrices

$$A^{**} = UAU' \quad \text{and} \quad C^{**} = U(A+B)U'$$

where A and B are the matrices in our 1st test, and U is a $p \times p$ orthogonal matrix with running element

$$u_{ij} = \frac{1}{\sqrt{p}} \{ \cos(2\pi(i-1)(j-1)/p) + \sin(2\pi(i-1)(j-1)/p) \} \quad (i, j, = 1, \dots, p).$$

It is then possible to show that, for $m = \lfloor p/2 \rfloor$,

$$\Lambda \sim Y_1 (Y^*)^{\text{mod}(p+1,2)} \prod_{j=2}^{p-m} Y_j^2,$$

where

$$Y_1 \stackrel{d}{=} Y^* \sim \text{Beta}\left(\frac{n-q}{2}, \frac{q-1}{2}\right) \quad \text{and} \quad Y_j \sim \text{Beta}(n-q, q-1)$$

form a set of independent r.v.'s.

As such, the exact distribution of Λ is given, for odd q , by Theorem 2 with

$$m^* = \lfloor \frac{p-1}{2} \rfloor + 1 + \text{mod}(p+1,2), \quad n_v = 1, \quad a_v = \frac{n-q}{2} + 1, \quad v = 1, \dots, m^*,$$

and

$$m_v = \begin{cases} q-1, & v = 1, \dots, m^{**} \\ \frac{q-1}{2}, & v = m^{**} + 1, \dots, m^* \end{cases} \quad \text{and} \quad k_v = \begin{cases} 2, & v = 1, \dots, m^{**} \\ 1, & v = m^{**} + 1, \dots, m^*, \end{cases}$$

for $m^{**} = \lfloor (p-1)/2 \rfloor$, so that, in this case, the exact PDF and CDF of Λ are given through the PDF and CDF of the EGIG distribution as

It is then possible to show that, for $m = \lfloor p/2 \rfloor$,

$$\Lambda \sim Y_1 (Y^*)^{\text{mod}(p+1,2)} \prod_{j=2}^{p-m} Y_j^2,$$

where

$$Y_1 \stackrel{d}{=} Y^* \sim \text{Beta}\left(\frac{n-q}{2}, \frac{q-1}{2}\right) \quad \text{and} \quad Y_j \sim \text{Beta}(n-q, q-1)$$

form a set of independent r.v.'s.

As such, the exact distribution of Λ is given, for odd q , by Theorem 2 with

$$m^* = \lfloor \frac{p-1}{2} \rfloor + 1 + \text{mod}(p+1,2), \quad n_v = 1, \quad a_v = \frac{n-q}{2} + 1, \quad v = 1, \dots, m^*,$$

and

$$m_v = \begin{cases} q-1, & v = 1, \dots, m^{**} \\ \frac{q-1}{2}, & v = m^{**} + 1, \dots, m^* \end{cases} \quad \text{and} \quad k_v = \begin{cases} 2, & v = 1, \dots, m^{**} \\ 1, & v = m^{**} + 1, \dots, m^*, \end{cases}$$

for $m^{**} = \lfloor (p-1)/2 \rfloor$, so that, in this case, the exact PDF and CDF of Λ are given through the PDF and CDF of the EGIG distribution as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:q-1}; \left\{ \frac{n-q-1-j}{2} \right\}_{j=1:q-1}; q-1 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=1:q-1}; \left\{ \frac{n-q-1-j}{2} \right\}_{j=1:q-1}; q-1 \right),$$

where

$$r_j = \left\lfloor \frac{p}{2} \right\rfloor + 1 - (1 + \text{mod}(p+1, 2)) \text{mod}(j-1, 2).$$

To exemplify the use and usefulness of the present test we use the data on response times (in seconds) for the words that in a tape-recorded sentence followed a given “token word”, taken from one of five positions in the sentence, reported in Tables 3.9.7 and 3.9.8 of Timm (2002), as forming three independent samples (one corresponding to the sample in Table 3.9.7, acting as the control group, and the other two formed respectively by the low and high short-term memory groups of subjects in Table 3.9.8). We will use these data to test the equality of the corresponding population mean vectors, accounting for the circularity of the covariance matrices, which, according to the results in subsections 5.2.1 and 5.2.3 of Coelho and Arnold (2019), should not be rejected for all three populations. Data were stored in the file **Response_times_all.dat**.

The command

```
> PvalDataEqMeanVecCirc("Response_times_all.dat")
```

gives a p-value of 0.019261, showing that there is some evidence against the null hypothesis of equality of the mean vectors for the 3 populations.

Timm, N.H. (2002). *Applied Multivariate Analysis*. 2nd ed., Springer Texts in Statistics, Springer

To exemplify the use and usefulness of the present test we use the data on response times (in seconds) for the words that in a tape-recorded sentence followed a given “token word”, taken from one of five positions in the sentence, reported in Tables 3.9.7 and 3.9.8 of Timm (2002), as forming three independent samples (one corresponding to the sample in Table 3.9.7, acting as the control group, and the other two formed respectively by the low and high short-term memory groups of subjects in Table 3.9.8). We will use these data to test the equality of the corresponding population mean vectors, accounting for the circularity of the covariance matrices, which, according to the results in subsections 5.2.1 and 5.2.3 of Coelho and Arnold (2019), should not be rejected for all three populations. Data were stored in the file **Response_times_all.dat**.

The command

```
> PvalDataEqMeanVecCirc("Response_times_all.dat")
```

gives a p-value of 0.019261, showing that there is some evidence against the null hypothesis of equality of the mean vectors for the 3 populations.

Timm, N.H. (2002). *Applied Multivariate Analysis*. 2nd ed., Springer Texts in Statistics, Springer

It is interesting to note that if we use the test for equality of mean vectors that does not assume the circularity of the covariance matrices

```
> PvalDataEqMeanVecR("Response_times_all.dat")
```

we obtain a quite larger p-value of 0.048263, indicating that when we assume the circularity of the covariance matrices we gain power in rejecting the null hypothesis (this would really indicate that the assumption of a circular covariance matrix is right – and this is indeed the case, as we will see when we will address the next test).

It is interesting to note that if we use the test for equality of mean vectors that does not assume the circularity of the covariance matrices

```
> PvalDataEqMeanVecR("Response_times_all.dat")
```

we obtain a quite larger p-value of 0.048263, indicating that when we assume the circularity of the covariance matrices we gain power in rejecting the null hypothesis (this would really indicate that the assumption of a circular covariance matrix is right – and this is indeed the case, as we will see when we will address the next test).

(See subsec 5.2.1 in Coelho and Arnold (2019))

Let us suppose we have a sample of size n from a population $\underline{X} \sim N_P(\underline{\mu}, \Sigma)$ and that we want to test the null hypothesis

$$H_0 : \Sigma \text{ is circular.}$$

The LRT was derived by Olkin and Press (1969). The $(2/n)$ -th power of the LRT statistic is

$$\Lambda = \frac{|V|}{\prod_{j=1}^p v_j^*}$$

where $V = nUAU'$, for U the orthogonal matrix defined in the previous test and A the (unstructured) MLE of Σ , and where, for $m = \lfloor p/2 \rfloor$,

$$v_j^* = \begin{cases} v_{jj} & j = 1, \text{ and also } j = m + 1 \text{ if } p \text{ is even} \\ (v_{jj} + v_{p-j+2, p-j+2})/2, & j = 2, \dots, p-m; m+2, \dots, p, \end{cases}$$

with $v_j^* = v_{p-j+2}^*$, where v_{jj} is the j -th diagonal element of V .

Olkin, I., Press, S.J. (1969). Testing and estimation for a circular stationary model. *Ann. Math. Stat.* 40, 1358-1373.

It may then be shown that

$$\Lambda \sim \prod_{j=1}^{p-1} Y_j$$

where

$$Y_j \sim \begin{cases} \text{Beta}\left(\frac{n-1-j}{2}, \frac{j}{2}\right) & j = 1, \dots, m \\ \text{Beta}\left(\frac{n-1-j}{2}, \frac{j+1}{2}\right) & j = m+1, \dots, p-1. \end{cases}$$

As such, the exact distribution of Λ is given, for odd p , by Theorem 3 with

$$m^* = 1, \quad n_1 = 2, \quad k_1 = \frac{p-1}{2}, \quad a_1 = \frac{n-1}{2}, \quad \text{and} \quad s_1 = 0,$$

so that, in this case, the exact PDF and CDF of Λ are given by Corollary 3, through the PDF and CDF of the EGIG distribution as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=2:p}; \left\{ \frac{n-j}{2} \right\}_{j=2:p}; p-1 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=2:p}; \left\{ \frac{n-j}{2} \right\}_{j=2:p}; p-1 \right),$$

It may then be shown that

$$\Lambda \sim \prod_{j=1}^{p-1} Y_j$$

where

$$Y_j \sim \begin{cases} \text{Beta}\left(\frac{n-1-j}{2}, \frac{j}{2}\right) & j = 1, \dots, m \\ \text{Beta}\left(\frac{n-1-j}{2}, \frac{j+1}{2}\right) & j = m+1, \dots, p-1. \end{cases}$$

As such, the exact distribution of Λ is given, for odd p , by Theorem 3 with

$$m^* = 1, \quad n_1 = 2, \quad k_1 = \frac{p-1}{2}, \quad a_1 = \frac{n-1}{2}, \quad \text{and} \quad s_1 = 0,$$

so that, in this case, the exact PDF and CDF of Λ are given by Corollary 3, through the PDF and CDF of the EGIG distribution as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=2:p}; \left\{ \frac{n-j}{2} \right\}_{j=2:p}; p-1 \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j\}_{j=2:p}; \left\{ \frac{n-j}{2} \right\}_{j=2:p}; p-1 \right),$$

where

$$r_j = 1 + \left\lfloor \frac{p-j}{2} \right\rfloor, \quad j = 2, \dots, p.$$

An application of this test will be done by testing the circularity for the population of “controls”, which sample is in Table 3.9.7 of Timm (2002) and also for the populations of the low and high short-term memory groups, which samples are in Table 3.9.8 of the same reference.

We can use the same file `Response_times_all.dat` that we used for the previous test and just run a command like

```
> PvalDataCircOddp("Response_times_all.dat")
```

three times, choosing each time one of the samples in the file.

For population 1 we get a p-value of 0.982209 and for populations 2 and 3 p-values of 0.443470 and 0.389817 respectively. And as such, we should not reject the null hypothesis that the covariance matrices for all 3 populations are circular.

where

$$r_j = 1 + \left\lfloor \frac{p-j}{2} \right\rfloor, \quad j = 2, \dots, p.$$

An application of this test will be done by testing the circularity for the population of “controls”, which sample is in Table 3.9.7 of Timm (2002) and also for the populations of the low and high short-term memory groups, which samples are in Table 3.9.8 of the same reference.

We can use the same file `Response_times_all.dat` that we used for the previous test and just run a command like

```
> PvalDataCircOddp("Response_times_all.dat")
```

three times, choosing each time one of the samples in the file.

For population 1 we get a p-value of 0.982209 and for populations 2 and 3 p-values of 0.443470 and 0.389817 respectively. And as such, we should not reject the null hypothesis that the covariance matrices for all 3 populations are circular.

where

$$r_j = 1 + \left\lfloor \frac{p-j}{2} \right\rfloor, \quad j = 2, \dots, p.$$

An application of this test will be done by testing the circularity for the population of “controls”, which sample is in Table 3.9.7 of Timm (2002) and also for the populations of the low and high short-term memory groups, which samples are in Table 3.9.8 of the same reference.

We can use the same file **Response_times_all.dat** that we used for the previous test and just run a command like

```
> PvalDataCircOddp("Response_times_all.dat")
```

three times, choosing each time one of the samples in the file.

For population 1 we get a p-value of 0.982209 and for populations 2 and 3 p-values of 0.443470 and 0.389817 respectively. And as such, we should not reject the null hypothesis that the covariance matrices for all 3 populations are circular.

where

$$r_j = 1 + \left\lfloor \frac{p-j}{2} \right\rfloor, \quad j = 2, \dots, p.$$

An application of this test will be done by testing the circularity for the population of “controls”, which sample is in Table 3.9.7 of Timm (2002) and also for the populations of the low and high short-term memory groups, which samples are in Table 3.9.8 of the same reference.

We can use the same file **Response_times_all.dat** that we used for the previous test and just run a command like

```
> PvalDataCircOddp("Response_times_all.dat")
```

three times, choosing each time one of the samples in the file.

For population 1 we get a p-value of 0.982209 and for populations 2 and 3 p-values of 0.443470 and 0.389817 respectively. And as such, we should not reject the null hypothesis that the covariance matrices for all 3 populations are circular.

One should not be surprised by the extremely high p-value for the test to population 1, since the sample covariance matrix for the five response times in the sample corresponding to Table 3.9.7 of Timm (2002), is

$$\begin{bmatrix} 65.0909 & 33.6455 & 47.5909 & 36.7727 & 25.4273 \\ 33.6455 & 46.0727 & 28.9455 & 40.3364 & 28.3636 \\ 47.5909 & 28.9455 & 60.6909 & 37.3727 & 41.1273 \\ 36.7727 & 40.3364 & 37.3727 & 62.8182 & 31.6818 \\ 25.4273 & 28.3636 & 41.1273 & 31.6818 & 58.2182 \end{bmatrix}$$

(See subsec 5.2.2 in Coelho and Arnold (2019))

Let us suppose that, as for the previous test, we have a sample of size n from a population $\underline{X} \sim N_P(\underline{\mu}, \Sigma)$ and that now we want to test the null hypothesis

$$H_0 : \Sigma \text{ is circular, } \underline{\mu} = aE_{p1} \text{ for some unspecified } a \in \mathbb{R}.$$

Then, the $(2/n)$ -th power of the LRT statistic is (Olkin and Press (1969))

$$\Lambda = \frac{|V|}{v_1^*} \prod_{j=2}^p \frac{1}{v_j^* + w_j}$$

where the v_j^* are defined as in the previous test and, for $m = \lfloor p/2 \rfloor$,

$$w_j = \begin{cases} y_j^2, & j = 1, \text{ and also } j = m + 1 \text{ if } p \text{ is even} \\ y_j^2 + y_{p-j+2}^2, & j = 2, \dots, p - m; m + 2, \dots, p \end{cases}$$

where $\underline{Y} = [y_j] = \sqrt{n}U\bar{\underline{X}}$ for a matrix U defined as in the two previous tests and $\bar{\underline{X}}$ as the sample mean vector.

Then

$$\Lambda \sim \prod_{j=2}^p Y_j$$

where

$$Y_j \sim \begin{cases} \text{Beta}\left(\frac{n-j}{2}, \frac{j}{2}\right) & j = 2, \dots, m+1 \\ \text{Beta}\left(\frac{n-j}{2}, \frac{j+1}{2}\right) & j = m+2, \dots, p. \end{cases}$$

As such, the exact distribution of Λ is given, for odd p , by Theorem 3 with

$$m^* = 1, \quad n_1 = 2, \quad k_1 = \frac{p-1}{2}, \quad a_1 = \frac{n-1}{2}, \quad \text{and} \quad s_1 = 1,$$

so that, in this case, the exact PDF and CDF of Λ are given by Corollary 3, through the PDF and CDF of the EGIG distribution as

$$f_{\Lambda}(z) = f^{EGIG}\left(z \mid \{r_j\}_{j=1:p}; \left\{\frac{n-j}{2}\right\}_{j=1:p}; p\right)$$

and

$$F_{\Lambda}(z) = f^{EGIG}\left(z \mid \{r_j\}_{j=1:p}; \left\{\frac{n-j}{2}\right\}_{j=1:p}; p\right),$$

with

$$r_j = \frac{p-1}{2} - \left\lfloor \frac{|j-2|}{2} \right\rfloor, \quad j = 1, \dots, p.$$

Then

$$\Lambda \sim \prod_{j=2}^p Y_j$$

where

$$Y_j \sim \begin{cases} \text{Beta}\left(\frac{n-j}{2}, \frac{j}{2}\right) & j = 2, \dots, m+1 \\ \text{Beta}\left(\frac{n-j}{2}, \frac{j+1}{2}\right) & j = m+2, \dots, p. \end{cases}$$

As such, the exact distribution of Λ is given, for odd p , by Theorem 3 with

$$m^* = 1, \quad n_1 = 2, \quad k_1 = \frac{p-1}{2}, \quad a_1 = \frac{n-1}{2}, \quad \text{and} \quad s_1 = 1,$$

so that, in this case, the exact PDF and CDF of Λ are given by Corollary 3, through the PDF and CDF of the EGIG distribution as

$$f_{\Lambda}(z) = f^{EGIG}\left(z \mid \{r_j\}_{j=1:p}; \left\{\frac{n-j}{2}\right\}_{j=1:p}; p\right)$$

and

$$F_{\Lambda}(z) = f^{EGIG}\left(z \mid \{r_j\}_{j=1:p}; \left\{\frac{n-j}{2}\right\}_{j=1:p}; p\right),$$

with

$$r_j = \frac{p-1}{2} - \left\lfloor \frac{|j-2|}{2} \right\rfloor, \quad j = 1, \dots, p.$$

To illustrate the application of this test we will use once again the data in Tables 3.9.7 and 3.9.8 of Timm (2002), to test simultaneously the circularity of the covariance matrix and the equality of the expected values for all $p = 5$ variables for the population of “controls” (Table 3.9.7) and the populations of the low and high short-term memory groups (Table 3.9.8).

We will use the same file `Response_times_all.dat` that we used for the two previous tests and run a command like

```
> PvalDataCircMeanOddp("Response_times_all.dat")
```

three times, choosing each time one of the samples in the file.

For population 1 we get a p-value of 0.937302 and for populations 2 and 3 p-values of 0.433794 and 0.238907 respectively. And as such, we should not reject the null hypothesis that the covariance matrices for all 3 populations are circular and simultaneously the means for the 5 variables are all equal, meaning that the placement of the token word seems to not affect the response time.

To illustrate the application of this test we will use once again the data in Tables 3.9.7 and 3.9.8 of Timm (2002), to test simultaneously the circularity of the covariance matrix and the equality of the expected values for all $p = 5$ variables for the population of “controls” (Table 3.9.7) and the populations of the low and high short-term memory groups (Table 3.9.8).

We will use the same file **Response_times_all.dat** that we used for the two previous tests and run a command like

```
> PvalDataCircMeanOddp("Response_times_all.dat")
```

three times, choosing each time one of the samples in the file.

For population 1 we get a p-value of 0.937302 and for populations 2 and 3 p-values of 0.433794 and 0.238907 respectively. And as such, we should not reject the null hypothesis that the covariance matrices for all 3 populations are circular and simultaneously the means for the 5 variables are all equal, meaning that the placement of the token word seems to not affect the response time.

To illustrate the application of this test we will use once again the data in Tables 3.9.7 and 3.9.8 of Timm (2002), to test simultaneously the circularity of the covariance matrix and the equality of the expected values for all $p = 5$ variables for the population of “controls” (Table 3.9.7) and the populations of the low and high short-term memory groups (Table 3.9.8).

We will use the same file **Response_times_all.dat** that we used for the two previous tests and run a command like

```
> PvalDataCircMeanOddp("Response_times_all.dat")
```

three times, choosing each time one of the samples in the file.

For population 1 we get a p-value of 0.937302 and for populations 2 and 3 p-values of 0.433794 and 0.238907 respectively. And as such, we should not reject the null hypothesis that the covariance matrices for all 3 populations are circular and simultaneously the means for the 5 variables are all equal, meaning that the placement of the token word seems to not affect the response time.

For even p (See subsec 5.3.1 in Coelho and Arnold (2019))

The exact distribution of Λ is given by a combination of the products in Theorems 1 or 2 and Theorem 3, and its PDF and CDF are thus given by Corollaries 4.4 or 4.5 in Coelho and Arnold (2019), with

$$m^* = 1, \quad n_1 = 1, \quad k_1 = 2, \quad a_1 = \frac{n}{2}, \quad m_1 = 1, \quad \text{and} \\ m^{**} = 1, \quad n_1^* = 2, \quad k_1^* = \frac{p}{2} - 1, \quad a_1^* = \frac{n-2}{2}, \quad s_1^* = 2,$$

this way giving exact PDFs and CDFs for Λ , through the EGIG PDF and CDF, with exactly similar expressions as for the case of odd p , now with

$$r_j = \begin{cases} \frac{p}{2} - 1, & j = 1 \\ \frac{p}{2} - \left\lfloor \frac{j-1}{2} \right\rfloor, & j = 2, \dots, p. \end{cases}$$

To illustrate the application of the test in this case, we will once again use the 3 samples in file **Response_times_all.dat**, where in each one of them we can choose for example any 4 variables on which to carry out the test.

We may for example use the command

```
> PvalDataCircMeanEvenp("Response_times_all.dat",1)
```

and for example choose the first 4 variables in each sample.

The p-values obtained (0.980742, 0.476920, 0.352923) come with no surprise, given the ones obtained for the test with all 5 variables, and once again would lead us to not reject the null hypothesis that the population covariance matrices are circular and the means for the 4 variables considered are all equal.

You may want to run the test for other combinations of even numbers of variables.

To illustrate the application of the test in this case, we will once again use the 3 samples in file **Response_times_all.dat**, where in each one of them we can choose for example any 4 variables on which to carry out the test.

We may for example use the command

```
> PvalDataCircMeanEvenp("Response_times_all.dat",1)
```

and for example choose the first 4 variables in each sample.

The p-values obtained (0.980742, 0.476920, 0.352923) come with no surprise, given the ones obtained for the test with all 5 variables, and once again would lead us to not reject the null hypothesis that the population covariance matrices are circular and the means for the 4 variables considered are all equal.

You may want to run the test for other combinations of even numbers of variables.

To illustrate the application of the test in this case, we will once again use the 3 samples in file **Response_times_all.dat**, where in each one of them we can choose for example any 4 variables on which to carry out the test.

We may for example use the command

```
> PvalDataCircMeanEvenp("Response_times_all.dat",1)
```

and for example choose the first 4 variables in each sample.

The p-values obtained (0.980742, 0.476920, 0.352923) come with no surprise, given the ones obtained for the test with all 5 variables, and once again would lead us to not reject the null hypothesis that the population covariance matrices are circular and the means for the 4 variables considered are all equal.

You may want to run the test for other combinations of even numbers of variables.

To illustrate the application of the test in this case, we will once again use the 3 samples in file **Response_times_all.dat**, where in each one of them we can choose for example any 4 variables on which to carry out the test.

We may for example use the command

```
> PvalDataCircMeanEvenp("Response_times_all.dat",1)
```

and for example choose the first 4 variables in each sample.

The p-values obtained (0.980742, 0.476920, 0.352923) come with no surprise, given the ones obtained for the test with all 5 variables, and once again would lead us to not reject the null hypothesis that the population covariance matrices are circular and the means for the 4 variables considered are all equal.

You may want to run the test for other combinations of even numbers of variables.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets with even number of variables)

[IndCircMeans]

125

(See subsec 5.3.4 in Coelho and Arnold (2019))

Let us suppose a similar setup as the one in the test of independence of several sets of variables, where we have $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, with

$$\underline{X} = [\underline{X}'_1, \dots, \underline{X}'_k, \dots, \underline{X}'_m]'$$

where $\underline{X}_k \sim N_{p_k}(\underline{\mu}_k, \Sigma_{kk})$, with

$$\underline{\mu} = [\underline{\mu}'_1, \dots, \underline{\mu}'_k, \dots, \underline{\mu}'_m]', \text{ and } \Sigma = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1k} & \cdots & \Sigma_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \Sigma_{k1} & \cdots & \Sigma_{kk} & \cdots & \Sigma_{km} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \Sigma_{m1} & \cdots & \Sigma_{mk} & \cdots & \Sigma_{mm} \end{bmatrix}.$$

and that we are interested in testing the null hypothesis

$$H_0 : \Sigma = \text{diag}(\Sigma_{11}, \dots, \Sigma_{kk}, \dots, \Sigma_{mm}), \text{ with} \\ \Sigma_{kk} \text{ circular and } \mu_k = a_k E_{p_k 1} \quad (k = 1, \dots, m)$$

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets with even number of variables)

[IndCircMeans]

126

Then, for a sample of size n from \underline{X} , the $(2/n)$ -th power of the LRT statistic is

$$\Lambda = |V| \prod_{k=1}^m \frac{1}{v_{1k}^* \prod_{j=2}^{p_k} (w_{jk} + v_{jk}^*)}$$

where $V = nUAU'$ for U the orthogonal matrix we already used before and A the MLE of Σ and the v_{jk}^* and w_{jk} defined as before, now defined for each one of the m sets of variables, and thus indexed in k , with $k = 1, \dots, m$.

We may then show that

$$\Lambda \sim \left\{ \prod_{k=1}^{m-1} \prod_{j=1}^{p_k} Y_{jk} \right\} \left\{ \prod_{k=1}^m \prod_{j=2}^{p_k} Y_{jk}^* \right\}$$

where, for $m_k = \lfloor p_k/2 \rfloor$,

$$Y_{jk} \sim \text{Beta} \left(\frac{n - q_k - j}{2}, \frac{q_k}{2} \right) \quad \text{and} \quad Y_{jk}^* \sim \begin{cases} \text{Beta} \left(\frac{n-j}{2}, \frac{j}{2} \right), & j = 2, \dots, m_k + 1 \\ \text{Beta} \left(\frac{n-j}{2}, \frac{j+1}{2} \right), & j = m_k + 2, \dots, p_k. \end{cases}$$

are all independent.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets with even number of variables)

[IndCircMeans]

126

Then, for a sample of size n from \underline{X} , the $(2/n)$ -th power of the LRT statistic is

$$\Lambda = |V| \prod_{k=1}^m \frac{1}{v_{1k}^* \prod_{j=2}^{p_k} (w_{jk} + v_{jk}^*)}$$

where $V = nUAU'$ for U the orthogonal matrix we already used before and A the MLE of Σ and the v_{jk}^* and w_{jk} defined as before, now defined for each one of the m sets of variables, and thus indexed in k , with $k = 1, \dots, m$.

We may then show that

$$\Lambda \sim \left\{ \prod_{k=1}^{m-1} \prod_{j=1}^{p_k} Y_{jk} \right\} \left\{ \prod_{k=1}^m \prod_{j=2}^{p_k} Y_{jk}^* \right\}$$

where, for $m_k = \lfloor p_k/2 \rfloor$,

$$Y_{jk} \sim \text{Beta} \left(\frac{n - q_k - j}{2}, \frac{q_k}{2} \right) \quad \text{and} \quad Y_{jk}^* \sim \begin{cases} \text{Beta} \left(\frac{n-j}{2}, \frac{j}{2} \right), & j = 2, \dots, m_k + 1 \\ \text{Beta} \left(\frac{n-j}{2}, \frac{j+1}{2} \right), & j = m_k + 2, \dots, p_k. \end{cases}$$

are all independent.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets with even number of variables)

[IndCircMeans]

127

As such, if all p_k are even, the exact distribution of Λ is given by a combination of the products in Theorems 1 or 2 and Theorem 3, and its PDF and CDF are thus given by Corollaries 4.4 or 4.5 in Coelho and Arnold (2019), with

$$m^* = 2m - 1, \quad k_v = 2 \quad (v = 1, \dots, 2m - 1), \quad n_v = \begin{cases} \frac{p_v}{2}, & v = 1, \dots, m - 1 \\ 1, & v = m, \dots, 2m - 1, \end{cases}$$

$$m_v = \begin{cases} q_v, & v = 1, \dots, m - 1 \\ 1, & v = m, \dots, 2m - 1, \end{cases} \quad a_v = \begin{cases} \frac{m - q_v}{2}, & v = 1, \dots, m - 1 \\ \frac{n}{2}, & v = m, \dots, 2m - 1, \end{cases}$$

and

$$m^{**} = m, \quad n_v^* = 2, \quad a_v^* = \frac{n - 2}{2}, \quad k_v^* = \frac{p_v}{2} - 1, \quad s_v^* = 2 \quad (v = 1, \dots, m^{**})$$

for

$$q_v = p_{v+1} + \dots + p_m,$$

where m is the number of sets of variables.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets with even number of variables)

[IndCircMeans]

128

The exact PDF and CDF of Λ are thus given, in this case, through the PDF and CDF of the EGIG distribution as

$$f_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j^*\}_{j=1:p}; \left\{ \frac{n-j}{2} \right\}_{j=1:p}; p \right)$$

and

$$F_{\Lambda}(z) = f^{EGIG} \left(z \mid \{r_j^*\}_{j=1:p}; \left\{ \frac{n-j}{2} \right\}_{j=1:p}; p \right),$$

with

$$r_j^* = \begin{cases} \sum_{k=1}^m r_{jk}^*, & j = 1, 2 \\ r_{j-2} + \sum_{k=1}^m r_{jk}^*, & j = 3, \dots, p \end{cases}$$

where

$$r_{jk}^* = \begin{cases} r_{jk}, & j = 1, \dots, p_k \\ 0, & j = p_k + 1, \dots, p, \end{cases} \quad \text{for } p = \sum_{k=1}^m p_k,$$

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets with even number of variables)

[IndCircMeans]

129

with

$$r_{jk} = \begin{cases} \frac{p_k}{2} - 1, & j = 1 \\ \frac{p_k}{2} - \left\lfloor \frac{j-1}{2} \right\rfloor, & j = 2, \dots, p_k, \end{cases} \quad (3)$$

and

$$r_j = \begin{cases} h_j, & j = 1, 2 \\ r_{j-2} + h_j, & j = 3, \dots, p-2, \end{cases}$$

for

$$h_j = (\# \text{ of } p_k\text{'s } (k = 1, \dots, m) \geq j) - 1, \quad j = 1, \dots, p-2.$$

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets with even number of variables)

[IndCircMeans]

130

As an example of application of this test we will use the data on response times for the five token words in a given sentence in Tables 3.9.7 and 3.9.8 of Timm (2002). Since we need a sample of a given size on several sets of variables, we decided to use the data on the first ten individuals for the five response times in Table 3.9.7 of Timm (2002) and all ten observations for the five response times for the two groups of individuals in Table 3.9.8 of the same reference, as if they would refer to three sets of five response times for three different sentences, collected on the same ten individuals, making up this way a sample of size ten for an overall set of 15 variables (the 15 response times, five for each sentence). Then we may be interested in testing the independence for example of the response times for the last four token words of the first sentence and the response times for the last two of the other two sentences, at the same time that we also test for the circularity of the covariance matrices and the equality of the mean response times.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets with even number of variables)

[IndCircMeans]

131

The data is in file **Response_times_1set.dat**, and we would then use a command like

```
> PvalDataIndCircMeans("Response_times_1set.dat",c(4,2,2),1)
```

and we would obtain a p-value of 0.445729, leading us to not reject the null hypothesis of independence of the 3 sets of variables together with the assumption of circular covariance matrices for the 3 sets of variables and equality of the mean values for the variables in each set.

We should anyway note that we are using a very small sample, with size ten, for a total of eight variables, which may give us a rather limited power for rejection of the null hypothesis.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets with even number of variables)

[IndCircMeans]

131

The data is in file **Response_times_1set.dat**, and we would then use a command like

```
> PvalDataIndCircMeans("Response_times_1set.dat",c(4,2,2),1)
```

and we would obtain a p-value of 0.445729, leading us to not reject the null hypothesis of independence of the 3 sets of variables together with the assumption of circular covariance matrices for the 3 sets of variables and equality of the mean values for the variables in each set.

We should anyway note that we are using a very small sample, with size ten, for a total of eight variables, which may give us a rather limited power for rejection of the null hypothesis.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets but one with even number of variables)

[IndCircMeans10dd]

132

(See subsec 5.3.5 in Coelho and Arnold (2019))

Let us suppose a similar setup to the one for the last test, and the test to a similar null hypothesis. The only difference being that now all but one of the sets of variables have an even number of variables, that is, one of the sets of variables has an odd number of variables.

Then the exact distribution of the LRT statistic would still be the same in terms of a product of independent Beta r.v.'s, but in terms of the EGIG distribution, the exact distribution of Λ will be given by a combination of the products in Theorem 3 and Theorem 1 or 2, and its p.d.f. and c.d.f. given by Corollaries 4.4 or 4.5 in Coelho and Arnold (2019), with

$$m^* = 2m - 2, \quad k_v = 2 \quad (v = 1, \dots, 2m - 2), \quad n_v = \begin{cases} 1, & v = 1, \dots, m - 1 \\ \frac{p_v}{2}, & v = m, \dots, 2m - 2, \end{cases}$$
$$m_v = \begin{cases} 1, & v = 1, \dots, m - 1 \\ q_v, & v = m, \dots, 2m - 2, \end{cases} \quad a_v = \begin{cases} \frac{n+1}{2}, & v = 1, \dots, m - 1 \\ \frac{n-q_v}{2}, & v = m, \dots, 2m - 2, \end{cases}$$

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets but one with even number of variables)

[IndCircMeans1Odd]

132

(See subsec 5.3.5 in Coelho and Arnold (2019))

Let us suppose a similar setup to the one for the last test, and the test to a similar null hypothesis. The only difference being that now all but one of the sets of variables have an even number of variables, that is, one of the sets of variables has an odd number of variables.

Then the exact distribution of the LRT statistic would still be the same in terms of a product of independent Beta r.v.'s, but in terms of the EGIG distribution, the exact distribution of Λ will be given by a combination of the products in Theorem 3 and Theorem 1 or 2, and its p.d.f. and c.d.f. given by Corollaries 4.4 or 4.5 in Coelho and Arnold (2019), with

$$m^* = 2m - 2, \quad k_v = 2(v = 1, \dots, 2m - 2), \quad n_v = \begin{cases} 1, & v = 1, \dots, m - 1 \\ \frac{p_v}{2}, & v = m, \dots, 2m - 2, \end{cases}$$

$$m_v = \begin{cases} 1, & v = 1, \dots, m - 1 \\ q_v, & v = m, \dots, 2m - 2, \end{cases} \quad a_v = \begin{cases} \frac{n+1}{2}, & v = 1, \dots, m - 1 \\ \frac{n-q_v}{2}, & v = m, \dots, 2m - 2, \end{cases}$$

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets but one with even number of variables)

[IndCircMeans10dd]

133

and

$$m^{**} = m, \quad n_v^* = 2, \quad a_v^* = \frac{n-1}{2}, \quad (v = 1, \dots, m^{**})$$

$$k_v^* = \begin{cases} \frac{p_v}{2} - 1, & v = 1, \dots, m-1 \\ \frac{p_m-1}{2}, & v = m \end{cases}, \quad s_v^* = \begin{cases} 2, & v = 1, \dots, m-1 \\ 1, & v = m, \end{cases}$$

for

$$q_v = p_{v+1} + \dots + p_m,$$

where m is the number of sets of variables.

Then we may show that the exact PDF and CDF of Λ are given, in terms of the PDF and CDF of the EGIG distribution by a similar formulation as for the previous case, now with the r_{jk} still given by (3) for $k = 1 \dots, m-1$ and by

$$r_{jk} = \frac{p_k - 1}{2} - \left\lfloor \frac{|j-2|}{2} \right\rfloor$$

for $k = m$.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets but one with even number of variables)

[IndCircMeans1Odd]

133

and

$$m^{**} = m, \quad n_v^* = 2, \quad a_v^* = \frac{n-1}{2}, \quad (v = 1, \dots, m^{**})$$

$$k_v^* = \begin{cases} \frac{p_v}{2} - 1, & v = 1, \dots, m-1 \\ \frac{p_m-1}{2}, & v = m \end{cases}, \quad s_v^* = \begin{cases} 2, & v = 1, \dots, m-1 \\ 1, & v = m, \end{cases}$$

for

$$q_v = p_{v+1} + \dots + p_m,$$

where m is the number of sets of variables.

Then we may show that the exact PDF and CDF of Λ are given, in terms of the PDF and CDF of the EGIG distribution by a similar formulation as for the previous case, now with the r_{jk} still given by (3) for $k = 1 \dots, m-1$ and by

$$r_{jk} = \frac{p_k - 1}{2} - \left\lfloor \frac{|j-2|}{2} \right\rfloor$$

for $k = m$.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets but one with even number of variables)

134

To exemplify the application of this test we will use once again the data in file **Response_times_1set.dat**, used in our previous test, now using the last 3 variables for the 1st set, and the last 2 for the 2nd and 3rd sets, by using the command

```
> PvalDataIndCircMeans10dd("Response_times_1set.dat",c(3,2,2),1)
```

which would lead to a p-value of 0.315999, or using the last 4 variables in the 1st set and the last 3 for the 2nd set and the last 2 for the 3rd set, with a command like

```
> PvalDataIndCircMeans10dd("Response_times_1set.dat",c(4,3,2),1)
```

which would lead to a p-value of 0.667459, in both cases leading to the non-rejection of the null hypothesis of independence of the 3 sets of variables and the assumption of circular covariance matrices of each set and the equality of the means of the variables in each set.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets but one with even number of variables)

134

To exemplify the application of this test we will use once again the data in file **Response_times_1set.dat**, used in our previous test, now using the last 3 variables for the 1st set, and the last 2 for the 2nd and 3rd sets, by using the command

```
> PvalDataIndCircMeans10dd("Response_times_1set.dat",c(3,2,2),1)
```

which would lead to a p-value of 0.315999, or using the last 4 variables in the 1st set and the last 3 for the 2nd set and the last 2 for the 3rd set, with a command like

```
> PvalDataIndCircMeans10dd("Response_times_1set.dat",c(4,3,2),1)
```

which would lead to a p-value of 0.667459, in both cases leading to the non-rejection of the null hypothesis of independence of the 3 sets of variables and the assumption of circular covariance matrices of each set and the equality of the means of the variables in each set.

The LRT for simultaneous test of Independence of several sets of variables, the Circularity of the covariance matrices and equality of mean values (all sets but one with even number of variables)

[IndCircMeans1Odd]

134



To exemplify the application of this test we will use once again the data in file **Response_times_1set.dat**, used in our previous test, now using the last 3 variables for the 1st set, and the last 2 for the 2nd and 3rd sets, by using the command

```
> PvalDataIndCircMeans1Odd("Response_times_1set.dat",c(3,2,2),1)
```


which would lead to a p-value of 0.315999, or using the last 4 variables in the 1st set and the last 3 for the 2nd set and the last 2 for the 3rd set, with a command like

```
> PvalDataIndCircMeans1Odd("Response_times_1set.dat",c(4,3,2),1)
```


which would lead to a p-value of 0.667459, in both cases leading to the non-rejection of the null hypothesis of independence of the 3 sets of variables and the assumption of circular covariance matrices of each set and the equality of the means of the variables in each set.


All  functions for the tests addressed in this presentation are available in the  workspace **EGIG.RData** in

<https://github.com/Carlos-Coelho/SASA-2024>


These  functions, as well as modules programmed in Mathematica[®] and Maxima for the tests addressed in Coelho and Arnold (2019) are also available from

<https://github.com/Carlos-Coelho/MeijerFoxFiniteforms>


-  workspace **EGIG.RData**
- Mathematica[®] package **EGIG.m**
- Maxima workspace **EGIG.wxmx**

All  functions for the tests addressed in this presentation are available in the workspace **EGIG.RData** in

<https://github.com/Carlos-Coelho/SASA-2024>

These  functions, as well as modules programmed in Mathematica[®] and Maxima for the tests addressed in Coelho and Arnold (2019) are also available from

<https://github.com/Carlos-Coelho/MeijerFoxFiniteforms>

-  workspace **EGIG.RData**
- Mathematica[®] package **EGIG.m**
- Maxima workspace **EGIG.wxmx**

Thanks for listening!

