NEAR-EXACT DISTRIBUTIONS AND FINITE FORM REPRESENTATIONS: PROBLEMS THEY CAN SOLVE

Full Professor, Ph.D. & Hab., Carlos A. Coelho[1]

Abstract: *Near-exact distributions are asymptotic distributions developed using a different approach. For Likelihood Ratio Test statistics used in Multivariate Analysis they are based on factorizations of the characteristic function (c.f.) of the negative logarithm of the statistic, leaving most of the factors unchanged and replacing a minor part by a sharp asymptotic result. This way we are able to overcome the undesirable problems we find when using common asymptotic distributions. It is also shown how the factorization of the above mentioned c.f. in many cases may lead to the obtention of finite forms for the exact pdf's and cdf's of such statistics. These finite forms as well as the near-exact distributions may be extremely useful in areas of application as Econometrics, helping to obtain precise inferential procedures where we can base our decisions and in finding groups of observations that may indicate the presence of phenomena worth studying. An illustration is done with an application of Wilks outlier test to recent econometric data from the 27 countries in the European Union.*

Keywords: *characteristic function, finite forms, likelihood ratio statistics, Multivariate Analysis, near-exact distributions, outlier test*

JEL: C44, C46

DOI: https://doi.org/10.58861/tae.pcesetfc.2024.22

1. Introduction

It is rather well known that the distributions of most Likelihood Ratio test (LRT) statistics used in Multivariate Analysis are not easy to handle. For this reason asymptotic distributions were developed for many of these statistics. However, it happens that most of these asymptotic distributions display behaviors that make their use to end up being very often a non-adequate choice. While the common single chi-square asymptotic distribution may fall way too far from the exact distribution, even for large samples (Brunner et al., 1977; Bai et al., 2009), other asymptotic distributions as the popular ones based on Box's approach (Box, 1949), besides still falling quite far from the exact distribution even for moderate size samples, have a quite unknown not desirable characteristic which is the one of really not being anymore a proper distribution for rather small samples, a problem that gets worse as the number of variables increases (Coelho and Marques, 2012; Coelho, 2014).

Moreover it is also a quite overlooked fact that for many LRT statistics used in Multivariate Analysis, namely for those that assume the form of a Wilks Lambda statistic (Wilks, 1932, 1935; Coelho 1988, 2022), it is really possible to

[1] Mathematics Department, NOVA School of Science and Technology, and NOVA Math – Center for Mathematics and Applications, NOVA University of Lisbon; email: cmac@fct.unl.pt

obtain their pdf's and cdf's in quite manageable finite closed forms for many situations (Coelho and Arnold, 2019). This is exactly the case of the statistic for the Wilks test for outliers (Wilks, 1963), which we will use in this paper as an example to illustrate the techniques used to obtain finite form representations for the pdf and cdf as well as the techniques used to build near-exact distributions.

2. The Wilks test for outliers

Wilks (1963) developed an interesting test for outliers, which, as he notes, although not being exactly an LRT, has quite close relations with an LRT, and which associated statistic may be given the form of a Wilks Lambda statistic. This test is addressed in more detail in section 5.1.13 of Coelho and Arnold (2019).

Let us suppose we have a sample of size n from a p-multivariate Normal population $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma\right)$, and that we want to test whether the $k\ (< n)$ observations numbered $\eta_1, \ldots, \eta_k$ (with $\eta_1, \ldots, \eta_k \in \{1, \ldots, n\}$) should be considered as outliers. Then we should consider the null hypothesis

$$H_0\text{: observations } \eta_1, \ldots, \eta_k \text{ are not outliers}$$

and Wilks suggests then the use the statistic

$$\Lambda = \frac{|A^*|}{|A|} \tag{1}$$

where A is equal to n times the usual MLE (Maximum Likelihood Estimator) 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 are not being tested for outliers.

But, if we take X as the $n \times p$ matrix of the sample of size n from $\underline{X}$, whose first $n-k$ rows are formed by the $n-k$ observations that are not being tested for outliers, and we call this sub-matrix as X_1, it may be shown that then we may write the statistic Λ in (1) as

$$\Lambda = \frac{|A^*|}{|A^* + B|} \tag{2}$$

where

$$A^* = X_1' Q_2 X_1 \quad \text{and} \quad B = X' Q_1 X$$

with $Q_2 = I_{n-k} - \frac{1}{n-k} E_{n-k,n-k}$, and

$$Q_1 = \left[\begin{array}{c|c} \frac{k}{n(n-k)} E_{n-k,n-k} & -\frac{1}{n} E_{n-k,k} \\ \hline -\frac{1}{n} E_{k,n-k} & I_k - \frac{1}{n} E_{kk} \end{array}\right]$$

where I_k represents an identity matrix of order k and E_{kk} represents a $k \times k$ matrix of ones (for further details see section 5.1.13 of Coelho and Arnold (2019)).

In (2) A^* and B are two independent Wishart matrices, with

$$A^* \sim W_p(n-k-1, \Sigma) \quad \text{and} \quad B \sim W_p(k, \Sigma)\ ,$$

and as such we can show that the distribution of Λ is the same as that of

$$\prod_{j=1}^{p} Y_j \qquad \text{or} \qquad \prod_{h=1}^{k} Y_h^* \tag{3}$$

where

$$Y_j \sim Beta\left(\frac{n-k-j}{2}, \frac{k}{2}\right) \qquad \text{and} \qquad Y_h^* \sim Beta\left(\frac{n-p-h}{2}, \frac{p}{2}\right)$$

Form two sets of independent random variables.

3. Finite forms for the pdf's and cdf's of many LRT statistics

As already mentioned above, a not so well known fact is that for many LRT statistics used in Multivariate Analysis it is really possible to obtain their pdf's and cdf's in quite manageable finite closed forms for many of the cases. This is exactly the case of the statistic for the Wilks outlier test in (1) or (2).

From the distribution of this statistic enounced in the previous section, we may use the results in Theorem 3.2 and Corollary 4.2 in Coelho and Arnold (2019) to obtain, for even p or even k, the exact pdf and cdf of Λ given as

$$f_\Lambda(z) = \left\{\prod_{j=1}^{p+k-2} \lambda_j^{r_j}\right\} \sum_{j=1}^{p+k-2} z^{\lambda_j - 1} \sum_{h=1}^{r_j} c_{jk}(-logz)^{h-1} \tag{4}$$

and

$$F_\Lambda(z) = \left\{\prod_{j=1}^{p+k-2} \lambda_j^{r_j}\right\} \sum_{j=1}^{p+k-2} z^{\lambda_j} \sum_{h=1}^{r_j} c_{jk}(h-1)! \sum_{i=0}^{h-1} \frac{(-logz)^i}{i!\ \lambda_j^{h-i}} \tag{5}$$

where

$$r_j = \begin{cases} h_j\,, & j = 1,2 \\ r_{j-2} + h_j\,, & j = 3, \ldots, p+k-2 \end{cases} \tag{6}$$

with

$$h_j = (\#\ of\ elements\ in\ \{p,k\} \geq j) - 1 \qquad (j = 1, \ldots, p+k-2 \tag{7}$$

and

$$\lambda_j = \frac{n-2-j}{2} \qquad (j = 1, \ldots, p+k-2)$$

and where

$$c_{j,r_j} = \frac{1}{(r_j - 1)!} \prod_{\substack{i=1 \\ i \neq j}}^{p+k-2} (\lambda_i - \lambda_j)^{-r_i}\,, \qquad j = 1, \ldots, p+k-2$$

and, for $h = 1, \ldots, r_j - 1$ and $j = 1, \ldots, p+k-2$,

$$c_{j,r_j-h} = \frac{1}{h} \sum_{i=1}^{h} \frac{(r_j - h + i - 1)!}{(r_j - h - 1)!} R(i,j,p+k-2) c_{j,r_j-(h-i)}\,,$$

for

$$R(i,j,p+k-2) = \sum_{\substack{h=1 \\ h\neq j}}^{p+k-2} r_h\left(\lambda_j - \lambda_h\right)^{-i} \qquad \left(i = 1, \dots, r_j - 1\right).$$

As an example of application we will take the data for 2023 of the per-head GNI (Gross National Income) at current prices, the UNE (Unemployment, percentage of active population) and the TFP (Total Factor Productivity, index with value 100 for 2015) obtained from the AMECO (Annual Macro Economic database of the European Commission Directorate General for Economic and Financial Affairs, at https://economy-finance.ec.europa.eu/economic-research-and-databases/economic-databases/ameco-database_en), for the 27 countries of the European Union. In Figure 1 we have two 3D plots of the cloud of points, which suggest that when considering these 3 variables Luxembourg and Ireland may be outliers, given their very high values of GNI and the very high value of TFP for Ireland and the quite low value for Luxembourg. If one wants to test the null hypothesis that these two countries should not be considered as outliers we obtain for the statistic in (1) or (2) a computed value of 0.227990, corresponding to a p-value of 7.056749×10^{-6}. We may note that in our case we have $p = 3$, since we have 3 variables, but we have $k = 2$, since we are testing for 2 countries, so that we are in a case where we have the exact pdf and cdf of Λ given by (4) and (5), and as such this p-value is obtained directly from the cdf of Λ in (5), by computing $P(\Lambda \leq 0.227990)$. The p-value is so low that we readily would reject the null hypothesis which states that these 2 countries are not outliers and take them as outliers. However, Wilks warns us that when we chose these two countries by looking at the plots, without noticing it, we actually made $\binom{27}{2}$ comparisons, by comparing them with the whole set of countries, and as such, if we want to use an α value of 0.05 for our test, we should actually reject the null hypothesis only if the p-value is smaller than $0.05/\binom{27}{2} = 0.0001424$. Well, given the very low p-value obtained, we should anyway reject the null hypothesis and consider Ireland and Luxembourg as outliers when we consider the set of the 3 variables under study.

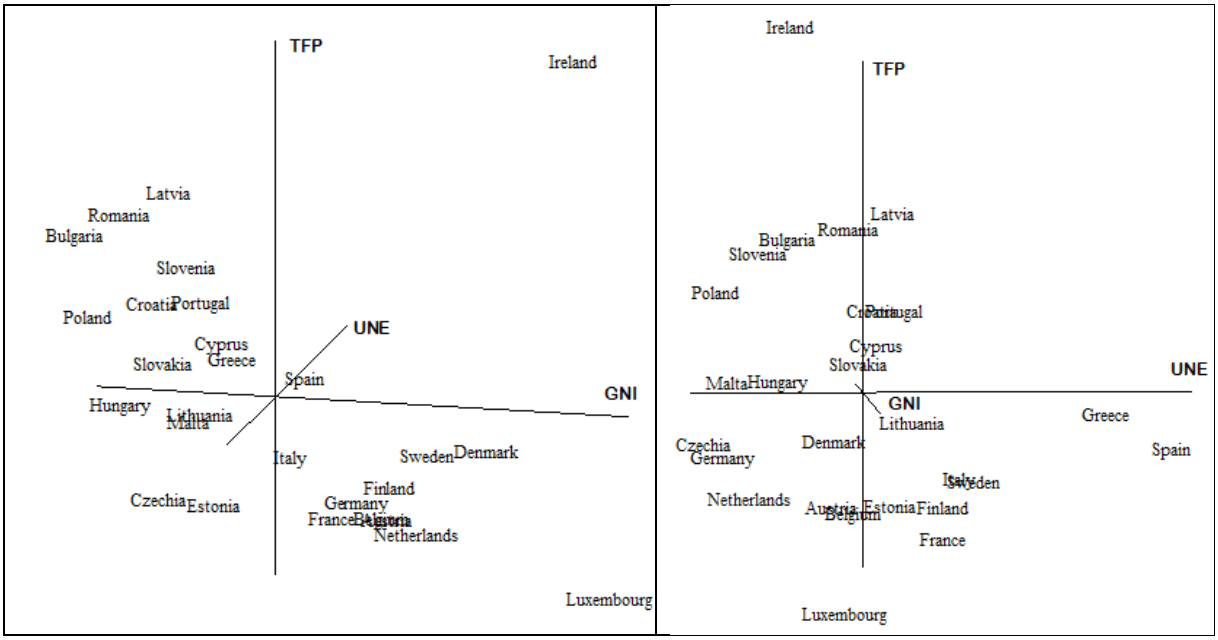

Figure 1 – 3D Plots with two different perspectives for the 27 countries in the European Union, concerning the Gross National Income (GNI), Percentage of Unemployment (UNE) and Total Factor Productivity (TFP)

When we look at the plots in Figure 1 also Greece and Spain seem to be possible candidates to outliers due to their very large values on the variable UNE. For these 2 countries we obtain a computed value of Λ equal to 0.430515, to which corresponds a p-value of 3.623317×10^{-3}. Although this p-value is smaller than the values of 0.05 and 0.01, commonly used for α, if we follow the recommendation from Wilks, we should not reject the null hypothesis that states that these two countries should not be considered as outliers, since the p-value is anyway larger than $0.05/\binom{27}{2} = 0.0001424$. However, would Spain and Greece had been chosen based on some theoretical or econometric reasons, and not just by looking at the plots, we should really consider them as outliers.

4. Near-exact distributions for Λ

In the left plot in Figure 1, it also calls the attention of our eyes a quite interesting group of 3 countries formed by Bulgaria, Romania and Latvia. These 3 countries have rather low GNI but quite high TFP (mainly after Ireland is removed), which is a quite interesting combination, with Bulgaria exhibiting a below average value of UNE, Romania an average value and Latvia a value slightly above average. It would be interesting to test if these 3 countries should be considered as outliers, mainly after Ireland and Luxembourg are removed.

When carrying out such a test we will be in a situation where $p = 3$ (we have 3 variables) and $k = 3$ (the 3 countries). In this case we do not have finite forms for the pdf or the cdf of Λ and we will have to resort to the use of near-exact distributions, since although we are still dealing with a sample of 25 countries, this sample size is way too small to use any of the currently available asymptotic

distributions.

Although in the previous section we used Theorem 3.2 and Corollary 4.2 in Coelho and Arnold (2019) to obtain the result for the finite form for the pdf and cdf of Λ, since it was a more expedite way of obtaining the result, we might had worked through the c.f. (characteristic function) of $W = -log\,\Lambda$, to obtain the same result. Now in order to obtain the near-exact distribution for Λ we will indeed need to work through the c.f. of W.

Using the fact that we know the expression for the h-th moments of the random variables Y_j in (3), we may readily write the c.f. of W as

$$\Phi_W(t) = E\left(e^{\mathrm{i}tW}\right) = E\left(e^{-\mathrm{i}t\log\Lambda}\right) = E\left(\Lambda^{-\mathrm{i}t}\right) = E\left(\prod_{j=1}^{p} Y_j^{-\mathrm{i}t}\right)$$
$$= \prod_{j=1}^{p} E\left(Y_j^{-\mathrm{i}t}\right) = \prod_{j=1}^{p} \frac{\Gamma\left(\frac{n-j}{2}\right)\Gamma\left(\frac{n-k-j}{2}-\mathrm{i}t\right)}{\Gamma\left(\frac{n-k-j}{2}\right)\Gamma\left(\frac{n-j}{2}-\mathrm{i}t\right)} \,.$$

Then, after some little work we may rewrite it, for r_j given by (6) and (7), as

$$\Phi_W(t) = \underbrace{\left\{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-2}{2}-\mathrm{i}t\right)\right\}\Big/\left\{\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}-\mathrm{i}t\right)\right\}}_{\Phi_1(t)}$$
$$\times \underbrace{\prod_{j=1}^{p+k-3}\left(\frac{n-p-k-1+j}{2}\right)^{r_j}\left(\frac{n-p-k+1-j}{2}-\mathrm{i}t\right)^{-r_j}}_{\Phi_2(t)} \,.$$

We will thus keep $\Phi_2(t)$ unchanged (since it corresponds to a sum of independent Gamma distributions with integer shape parameters, which has a closed finite form for the pdf and cdf), and we will approximate asymptotically $\Phi_1(t)$ by a finite mixture of $\Gamma(1/2+k,(n-2)/2)$ distributions $(k=0,\ldots,m^*)$, with weights determined in such a way that this mixture matches the first m^* derivatives of $\Phi_1(t)$ at $t=0$. The near-exact distributions obtained in this way will match the first m^* exact moments of W and will be mixtures of m^*+1 GNIG (Generalized Near-Integer Gamma) distributions, which will yield near-exact distributions for Λ with pdf's and cdf's respectively given by

$$f_\Lambda(z) = \sum_{h=0}^{m^*} \pi_h \, f^{GNIG}\left(-\log z \,\middle|\, \{r_j\}_{j=1:p+k-3}, \frac{1}{2}+h; \left\{\frac{n-2}{2}\right\}_{j=1}\right.$$
$$\left.\left\{\frac{n-p-k-1+j}{2}\right\}_{j=1:p+k-3}, \frac{n-2}{2}; p+k-3\right)\frac{1}{z}$$

and

$$F_{\Lambda}(z) = 1 - \sum_{h=0}^{m^*} \pi_h \; F^{GNIG}\left(-\log z \;\middle|\; \{r_j\}_{j=1:p+k-3}, \frac{1}{2} + h; \left\{\frac{n-2}{2}\right\}_{j=1} \right.$$
$$\left. \left\{\frac{n-p-k-1+j}{2}\right\}_{j=1:p+k-3}, \frac{n-2}{2}; p+k-3 \right)$$

where $f^{GNIG}(\cdot)$ and $F^{GNIG}(\cdot)$ represent respectively the pdf and cdf of the GNIG distribution (see Appendix C in Coelho (2021) for further details on the notation used and on the GNIG distribution).

The computed value of the statistic Λ, when testing the 3 countries Bulgaria, Romania and Latvia for outliers, considering the 25 countries that remain after removing Ireland and Luxembourg from the study, is 0.423534, to which corresponds a p-value of 6.239164×10^{-3}, obtained from the near-exact cdf of Λ that matches 4 exact moments of W, that is, the one with $m^* = 4$. This p-value although being rather small, and smaller than the values of 0.05 and 0.01, commonly used for α, is still quite larger than $0.05/\binom{25}{3} = 2.17391 \times 10^{-5}$. Since the 3 countries were indeed chosen by looking at the plot, we should not reject the null hypothesis that they are not outliers, although they stand out from the group of the other 22 countries and may be a group of countries worth studying since they show, at least for the year 2023 an interesting pattern.

5. Brief conclusion and final remarks

The exact form obtained for the cdf of Λ when either p or k are even, and the near-exact distribution obtained for the case when they are both odd enable us to obtain very sharp p-values on which we can base our inference.

References

Bai, Z., Jiang, D., Yao, J.-F., Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT, *Ann. Statist.*, 37, 3822-3840.

Box, G. E. P. (1949) A general distribution theory for a class of likelihood criteria, *Biometrika*, 36, 317–346.

Brunner, E., Dette, H. & Munk, A. (1977). Box-Type Approximations in Nonparametric Factorial Designs, *J. Amer. Stat. Assoc.*, 92, 1494-1502.

Coelho, C. A. (1998). The Generalized Integer Gamma Distribution – a Basis for Distributions in Multivariate Statistics. *J. Multiv. Analysis*, 64, 86-102.

Coelho, C. A. (2014). Near-exact distributions – what are they and why do we need them? *Proceedings 59th ISI World Statistics Congress, Special Topics Session 084* "Is distribution theory still relevant?", 2879-2884.

Coelho, C. A. (2021). Testing equality of mean vectors with block-circular and block compound-symmetric covariance matrices, in 'Multivariate, Multilinear and Mixed Linear Models', Filipiak, K., Markiewicz, A., von Rosen, D. (eds.), Springer series in Contributions to Statistics, 157–201.

Coelho, C. A. (2022). Likelihood ratio tests for elaborate covariance structures and for MANOVA models with elaborate covariance structures – a review, *J. Indian Inst. Science*, 102, 1219-1257.

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.

Coelho, C. A., Marques, F. J. (2012). Near-exact distributions for the likelihood ratio test statistic to test equality of several variance-covariance matrices in elliptically contoured distributions, *Computational Statistics*, 27, 627-659.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance, *Biometrika*, 24, 471-494.

Wilks, S. S. (1935). On the independence of *k* sets of normally distributed statistical variables. *Econometrica*, 3, 309-326.

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