A near-exact distribution and exact percentage points for testing independence with missing elements in the sample correlation matrix

Evelina Veleva

Abstract: *This paper consider the likelihood ratio test for diagonality of the covariance matrix of a p-variate normal distribution, when the sample correlation matrix has missing elements. The exact null distribution of the test statistic is represented through Meijer G-functions. For small values of p, percentage points for α=0.05 and α=0.01 are computed. For calculation of quantiles when p is large, a near-exact null distribution of the test statistic is derived in the form of a Generalized Near-Integer Gamma distribution.*

Key words: *Product of beta random variables, Likelihood ratio test, Percentage points, Meijer G-function, Near-exact distribution, Generalized Near-Integer Gamma distribution, Monte Carlo simulation*

INTRODUCTION

The $[2/n]$-th power of the likelihood ratio test statistic for independence of p multivariate normal distributed random variables, based on a sample of size n, is the statistic (see [1])

$$L = \det R , \tag{1}$$

where R is the sample correlation matrix. Often in practice, the missing observations on some of the p random variables can lead to missing elements in the sample correlation matrix. For instance, if we have n_1 observations on the first $p-1$ random variables and n_2 realizations of the variables with numbers from $k+1$ to p, then in the sample correlation matrix $R=(r_{i,j})$ the elements $r_{1,p}$, …, $r_{k,p}$ will be unidentified. Another reason for missing elements in the sample correlation matrix might be a loss during the keeping or the transportation to the researcher. To check the independence of p multivariate normal distributed random variables, when the elements $r_{1,p}$, …, $r_{k,p}$, $1 \le k \le p-2$ of the sample correlation matrix $R=(r_{i,j})$ are missing, the likelihood ratio test is derived in [5]. Denoted by L_k, the $[2/(n-1)]$-th power of this test statistic is given below in (2). Under the null hypothesis that the p variables are independent, L_k is distributed as the product of $p-1$ beta random variables, as it is shown in [6].

In this paper we give a representation of the null distribution of L_k in terms of Meijer G-functions. Using both the exact mathematical formulas with a commercial mathematical software and Monte Carlo simulations, some percentage points for small values of p are computed. For calculation of quantiles when p is large, a near-exact distribution for the null distribution of L_k is derived.

THE EXACT NULL DISTRIBUTION OF L_k

Let A be a real square matrix of order n. Let α and β be nonempty subsets of the set $N_n = \{1,\dots,n\}$. By $A[\alpha,\beta]$ we denote the submatrix of A, composed of the rows with numbers from α and the columns with numbers from β. When $\beta \equiv \alpha$, $A[\alpha,\alpha]$ is denoted simply by $A[\alpha]$.

The $[2/(n-1)]$-th power of the likelihood ratio test statistic for independence of the p variables is shown in [5] to be

$$L_k = \frac{\det R[\{k+1,\dots,p\}]\det R[\{1,\dots,p-1\}]}{\det R[\{k+1,\dots,p-1\}]} , \tag{2}$$

where k is the number of missing elements in the last column of the sample correlation matrix and n is the total number of observations on the p random variables. The likelihood ratio is derived under the assumption that we do not hold the initial observations, and have only the sample correlation matrix, in which the elements $r_{1,p}$, …, $r_{k,p}$, $1 \le k \le p-2$ are missing.

From (2) It can be seen that when $k = 0$, i.e. there are no missing elements in the sample correlation matrix, L_k equals the usual test statistic (1) for independence of p random variables.

Subsequently, $X \sim Beta(\alpha, \beta)$ denotes the classical beta random variable defined on $[0,1]$, with density $f(x;\alpha,\beta) = x^{\alpha-1}(1-x)^{\beta-1} / B(\alpha,\beta)$, with $B(\alpha,\beta)$ being the beta function in (α,β).

In [6] it is shown that under the null hypothesis of independence of the p variables, L_k is distributed as a product of $p-1$ independent beta random variables

$$L_k \cong \prod_{j=1}^{p-1} X_j , \tag{3}$$

$$X_j \sim Beta\left(\frac{n-p+j}{2}, \frac{p-j-1}{2}\right),\ j = 1,\dots,k\ ,\ X_j \sim Beta\left(\frac{n-p+j-1}{2}, \frac{p-j}{2}\right),\ j = k+1,\dots,p-1 .$$

The density $g(u)$ of the product of p independently distributed beta random variables with parameters (α_j, β_j), $j = 1,\dots,p$ can be expressed in terms of Meijer G-functions (see [4], p.51) as follows

$$g(u) = \prod_{j=1}^{p} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)}\, G_{p,p}^{p,0}\left[u \left| \begin{matrix} \alpha_j + \beta_j - 1,\ j = 1,\dots,p \\ \alpha_j - 1,\quad j = 1,\dots,p \end{matrix} \right. \right],\ 0 < u < 1 . \tag{4}$$

From (3) and (4) we obtain a representation for the density of the null distribution of L_k:

$$f_{L_k}(x) = K G_{p-1,p-1}^{p-1,0}\left[x \left| \begin{matrix} \frac{n-3}{2}, & \dots & , \frac{n-3}{2}, & \frac{n-3}{2}, & \dots & , \frac{n-3}{2} \\ \frac{n-p-1}{2}, & \dots, & \frac{n-p+k-2}{2}, & \frac{n-p+k-2}{2}, & \dots, & \frac{n-4}{2} \end{matrix} \right. \right],\ 0 < x < 1 , \tag{5}$$

with $K = \dfrac{\left[\Gamma\left(\frac{n-1}{2}\right)\right]^{p-1}}{\Gamma\left(\frac{n-p+k}{2}\right)\prod_{j=1}^{p-2}\Gamma\left(\frac{n-p+j}{2}\right)}$.

The commercial mathematical softwares, like MAPLE and MATHEMATICA, are able to compute the Meijer G-functions. For calculation of the α-quantile of the null distribution of L_k, we need to solve the equation $\int_0^u f_{L_k}(x)dx = \alpha$ with respect to u. It still takes too much computer time compare to the calculation of quantiles using Monte Carlo simulation techniques.

A NEAR-EXACT NULL DISTRIBUTION OF L_k

A near-exact distribution theory for the most common likelihood ratio test statistics used in multivariate analysis is generalized in [3]. The near-exact distributions are much closer to the exact distributions than common asymptotic distributions are. They are known manageable distributions, from which quantiles and p-values may be easily computed.

To obtain a near-exact distribution for the null distribution of L_k, where k is an arbitrary integer, $1 \le k \le p-2$, we shall use an approach similar to those in [2] for the case $k = 0$.

Let $W = -\log L_k$ and $\phi_W(t)$ be the characteristic function of W. Since for $X\sim Beta(\alpha,\beta)$ we have $E(X^s) = \dfrac{\Gamma(\alpha+s)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+s)\Gamma(\alpha)}$, $s > -\alpha$, using (3) we obtain that

$$\phi_W(t) = E(e^{itW}) = \prod_{j=1}^{p-1} E(X_j^{-it}) = \left(\prod_{j=1}^{k} \frac{\Gamma\left(\frac{n-p+j}{2}-it\right)\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-p+j}{2}\right)\Gamma\left(\frac{n-1}{2}-it\right)}\right)\left(\prod_{j=k+1}^{p-1} \frac{\Gamma\left(\frac{n-p+j-1}{2}-it\right)\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-p+j-1}{2}\right)\Gamma\left(\frac{n-1}{2}-it\right)}\right)$$

$$= \frac{\Gamma\left(\frac{n-p+k}{2}-it\right)\Gamma\left(\frac{n-p}{2}\right)}{\Gamma\left(\frac{n-p+k}{2}\right)\Gamma\left(\frac{n-p}{2}-it\right)}\left(\prod_{j=1}^{p-1} \frac{\Gamma\left(\frac{n-p+j-1}{2}-it\right)\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-p+j-1}{2}\right)\Gamma\left(\frac{n-1}{2}-it\right)}\right), \tag{6}$$

where $i = (-1)^{1/2}$ is the imaginary unit. In [2], the second factor in (6) is written of the form

$$\prod_{j=1}^{p-1} \frac{\Gamma\left(\frac{n-p+j-1}{2}-it\right)\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-p+j-1}{2}\right)\Gamma\left(\frac{n-1}{2}-it\right)} = \left(\frac{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-2}{2}-it\right)}{\Gamma\left(\frac{n-1}{2}-it\right)\Gamma\left(\frac{n-2}{2}\right)}\right)^{\left\lfloor \frac{p}{2}\right\rfloor} \prod_{j=1}^{p-2}\left(\frac{\frac{n-2-j}{2}}{\frac{n-2-j}{2}-it}\right)^{\left\lfloor \frac{p-j}{2}\right\rfloor},$$

where $\lfloor x \rfloor$ denotes the largest integer not greater than x. We are using the same approach as in [2] to deal with the complementary factor in (6) and derive $\phi_W(t)$ of the form

$$\phi_W(t) = \underbrace{\left(\frac{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-2}{2}-it\right)}{\Gamma\left(\frac{n-1}{2}-it\right)\Gamma\left(\frac{n-2}{2}\right)}\right)^{\left\lfloor \frac{p}{2}\right\rfloor+\delta}}_{\phi_1(t)} \prod_{j=1}^{p-3}\left(\frac{\frac{n-2-j}{2}}{\frac{n-2-j}{2}-it}\right)^{\left\lfloor \frac{p-j-1}{2}\right\rfloor+\varepsilon_j}, \tag{7}$$

where

$$\delta = \begin{cases} -1, & if\ k\ is\ odd\ and\ p\ is\ even \\ 1, & if\ k\ is\ odd\ and\ p\ is\ odd \\ 0, & if\ k\ is\ even \end{cases}, \quad \varepsilon_j = \begin{cases} 1, & if\ p-k-j\ is\ positive\ and\ even \\ 0, & otherwise \end{cases}. \tag{8}$$

Let us denote the first factor in the right hand side of (7) by $\phi_1(t)$. Replacing in (7) $\phi_1(t)$ with the function $\mu^r(\mu - it)^{-r}$, which is the characteristic function of a Gamma distributed random variable with parameters μ and r, we obtain the characteristic function of a near-exact distribution of W. For

$$\mu = \frac{m_1}{m_2 - m_1^2},\quad r = \frac{m_1^2}{m_2 - m_1^2}, \text{ where } m_1 = \frac{1}{i}\left(\frac{d}{dt}\phi_1(t)\right)\bigg|_{t=0},\ m_2 = -\left(\frac{d^2}{dt^2}\phi_1(t)\right)\bigg|_{t=0}, \tag{9}$$

the first two moments of the exact and near-exact distribution of W will coincide (see [3]). The values of m_1 and m_2 can be computed numerically. The obtained near-exact distribution of W is a Generalized Near – Integer Gamma (GNIG) distribution (see [2] and [3]), $GNIG(r_1,\ldots,r_{p-3},r;\lambda_1,\ldots,\lambda_{p-3},\mu;p-2)$, where μ and r are given by (9), $r_j = \lfloor (p-j-1)/2 \rfloor + \varepsilon_j$ with ε_j given by (8) and $\lambda_j = (n-2-j)/2$, $j = 1,\ldots,p-3$. This is the distribution of the sum of $p-2$ independent Gamma distributed random variables X_1, …, X_{p-2}, where X_j has integer shape parameter r_j and rate parameter $\lambda_j > 0$, with $\lambda_j \neq \lambda_{j'}$, for all $j, j' \in \{1,\ldots,p-3\}$ and X_{p-2} has noninteger shape parameter r and rate parameter

$\mu \neq \lambda_j$, $j = 1,\ldots,p-2$. The probability density function and the cumulative distribution function of a GNIG distribution can be found in [3]. If for a given α, $W^*_{(1-\alpha)}$ is the $(1-\alpha)$ quantile of the obtained GNIG distribution, then since $W = -\log L_k$, $L^*_{k(\alpha)} = e^{-W^*_{(1-\alpha)}}$ is the α near-exact quantile of L_k.

NUMERICAL EXAMPLES AND SOME PERCENTAGE POINTS

Let us consider the case $n = 10$, $p = 5$ and $k = 2$. We compute the density of L_k, using three different methods:

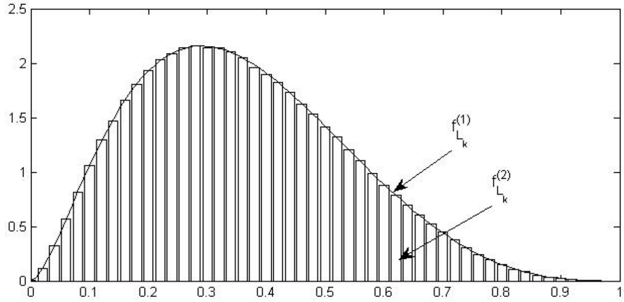
(1) the expression of f_{L_k}, as a G-function distribution given by equation (5), and using the mathematical software MAPLE. Its graph is denoted by $f^{(1)}_{L_k}$.

(2) We simulate 1 000 000 values of L_k as the product $\prod_{j=1}^{4} X_j$, with $X_j \sim Beta((5+j)/2,(4-j)/2)$, $j = 1,2$, $X_j \sim Beta((4+j)/2,(5-j)/2)$, $j = 3,4$ as given by (3), and derive the density denoted by $f^{(2)}_{L_k}$.

(3) We simulate 1 000 000 values of the near-exact distribution of W, i.e. the $GNIG(2,1,1.0042;3.5,3,3.7824;3)$ distribution, as the distribution of the sum of three independent Gamma random variables, with shape parameters 2, 1 and 1.0042 and corresponding rate parameters 3.5, 3 and 3.7824. The values 1.0042 and 3.7824 for r and μ are obtained numerically from (9). Using the formula $L_k = e^{-W}$ we compute 1 000 000 "near-exact" realizations of L_k and derive a near-exact density denoted by $f^{(3)}_{L_k}$.

Figure 1 show these three densities, which can be seen as almost identical.

1.a



1.b

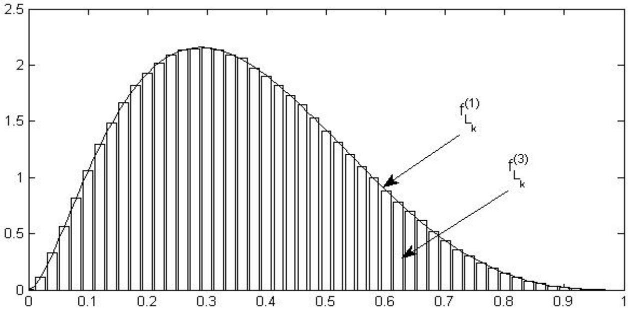


Figure 1. Density of L_k, with $n = 10$, $p = 5$ and $k = 2$, by three different approaches.

The 0.05 and 0.01 significance points of the null distribution of L_k were computed for various values of n, p and k. The results are given in Table 1.

Table 1. 1% and 5% points of L_k

n \ p	4		5		6		7		8	
	k=1	*k*=2	*k*=1	*k*=2	*k*=1	*k*=2	*k*=1	*k*=2	*k*=1	*k*=2
1 % points										
10	0.1195	0.1606	0.0382	0.0524	0.0087	0.0126	0.0012	0.0019	0.08E-03	0.15E-03
15	0.2874	0.3385	0.1563	0.1833	0.0748	0.0883	0.0308	0.0369	0.0106	0.0130
20	0.4139	0.4633	0.2730	0.3036	0.1669	0.1856	0.0937	0.1045	0.0478	0.0536
25	0.5053	0.5507	0.3684	0.3988	0.2545	0.2751	0.1657	0.1792	0.1012	0.1095
30	0.5732	0.6144	0.4443	0.4734	0.3303	0.3512	0.2347	0.2494	0.1589	0.1689
50	0.7259	0.7551	0.6289	0.6514	0.5332	0.5512	0.4415	0.4560	0.3568	0.3684
5 % points										
10	0.2103	0.2706	0.0785	0.1031	0.0215	0.0296	0.0038	0.0057	0.35E-03	0.62E-03
15	0.4006	0.4611	0.2349	0.2701	0.1219	0.1412	0.0549	0.0643	0.0209	0.0249
20	0.5235	0.5770	0.3629	0.3985	0.2336	0.2567	0.1384	0.1526	0.0747	0.0829
25	0.6060	0.6529	0.4586	0.4920	0.3290	0.3527	0.2226	0.2389	0.1415	0.1522
30	0.6647	0.7060	0.5307	0.5614	0.4065	0.4296	0.2978	0.3147	0.2079	0.2199
50	0.7906	0.8181	0.6961	0.7183	0.5998	0.6182	0.5049	0.5200	0.4148	0.4272

CONCLUSIONS

The computation of quantiles of a G-function distribution is still very slow when using precise numerical integration on the distribution tails. It is shown that by Monte Carlo simulation, quick and at the same time precise calculations can be achieved, especially when the number of variables p is not very large. A near-exact distribution for the null distribution of L_k is derived, which can be used for quick and accurate calculation of quantiles for all values of p.

REFERENCES

[1] Anderson,T. An Introduction to Multivariate Statistical Analysis. 3rd ed., New York: J.Wiley & Sons, 2003.

[2] Coelho,C., F.Marques. Near-exact distributions for the independence and sphericity likelihood ratio test statistics. Journal of Multivariate Analysis, Elsevier, vol. 101(3), 2010, 583-593.

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

[4] Mathai,A., H.Haubold. Special Functions for Applied Scientists. New York: Springer Science + Business Media, LLC, 2008.

[5] Veleva,E. Test for independence of the variables with missing elements in the same column of the empirical correlation matrix. Serdica Math. J., 34 (2008), 509 – 530.

[6] Veleva,E. The Exact Distribution Of The Ump Test For Diagonality Of Covariance Matrices With Missing Elements. Journal of the Technical University at Plovdiv “Fundamental Sciences and Applications”, vol. 14, 2009, 449-454.

ABOUT THE AUTHOR

Principal Assistant, Evelina Veleva, Department of Numerical Methods and Statistics, University of Rousse, e-mail: eveleva@uni-ruse.bg

Докладът е рецензиран.