The advantage of decomposing elaborate hypotheses on covariance matrices into conditionally independent hypotheses in building near-exact distributions for the test statistics

Carlos A. Coelho [*] Filipe J. Marques

fjm@fct.unl.pt *cmac@fct.unl.pt*

Mathematics Department, Faculty of Sciences and Technology
The New University of Lisbon, Portugal

Abstract

The aim of this paper is to show how the decomposition of elaborate hypotheses on the structure of covariance matrices into conditionally independent simpler hypotheses, by inducing the factorization of the overall test statistic into a product of several independent simpler test statistics, may be used to obtain near-exact distributions for the overall test statistics, even in situations where asymptotic distributions are not available in the literature and adequately fit ones are not easy to obtain.

Key words: Near-exact distributions, conditionally independent hypotheses, likelihood ratio test statistics, generalized sphericity tests, Generalized Near-Integer Gamma distribution, mixtures.

1 Introduction

The concept of near-exact distribution has already been introduced in a number of papers [5,6,9,1,10,11,8,13]. In a nutshell, near-exact distributions are new asymptotic distributions that lay closer to the exact distribution than

[*] Corresponding author. Address: The New University of Lisbon, Faculty of Sciences and Technology, Mathematics Department, Quinta da Torre, 2829-516 Caparica, Portugal; Tel:(351) 21 294 8388; Fax:(351) 21 294 8391.

Preprint submitted to Elsevier Science *16 November 2007*

the usual asymptotic distributions. They correspond to what we call near-exact c.f.'s (characteristic functions) which are c.f.'s obtained from the exact c.f. by leaving the most of it unchanged and replacing the remaining smaller part by an asymptotic result (which is intended to be asymptotic both in terms of sample size and overall number of variables, in a manner that will be more precisely stated ahead in this section). This replacement is done in such a way that the resulting c.f. corresponds to a distribution with a manageable c.d.f. (cumulative distribution function) so that the computation of near-exact quantiles is rendered feasible, easy and precise.

There are mainly two ways in which near-exact distributions, or rather, near-exact c.f.'s may be obtained. One of them is when the exact distribution is an infinite mixture, in which case we may cut short the series that corresponds to the exact c.f. at a given number of terms and replace the remainder by just one or a mixture of two or three distributions of the same kind of the ones in the infinite mixture, with parameters computed in such a way that the first few derivatives (say, usually, the first two, four or six of them) at zero of the replacing part and of the part being replaced, they match. More precisely, if the exact c.f. of the statistic W may be written as

$$\Phi_W(t) = \sum_{i=0}^{\infty} p_i \, \Phi_i(t)$$

where $p_i > 0$ $(i = 0, 1, \ldots)$ and $\sum_{i=0}^{\infty} p_i = 1$, and where $\Phi_i(t)$ are c.f.'s, then we will use instead of $\Phi_W(t)$,

$$\Phi_W^*(t) = \sum_{i=0}^{n^*} p_i \, \Phi_i(t) + \theta \, \Phi_2(t) \, ,$$

where $\theta = 1 - \sum_{i=0}^{n^*} p_i$ and $\Phi_2(t)$ is usually either a c.f. of the type of $\Phi_i(t)$ or the c.f. of the mixture of two or three such c.f.'s, defined in such a way that

$$\left. \frac{d}{dt^h} \, \theta \, \Phi_2(t) \right|_{t=0} = \left. \frac{d}{dt^h} \sum_{i=n^*+1}^{\infty} p_i \, \Phi_i(t) \right|_{t=0} \, ,$$

for $h \in H$ (where usually we will have $H = \{1, 2\}$, $H = \{1, \ldots, 4\}$ or $H = \{1, \ldots, 6\}$).

The other way to obtain a near-exact c.f. is through the factorization of the exact c.f., that is, if we may write

$$\Phi_W(t) = \Phi_1(t) \, \Phi_2(t)$$

where we recognize $\Phi_1(t)$ as corresponding to a known and manageable distribution and $\Phi_2(t)$ as corresponding to a non-manageable distribution. Then we will use as near-exact c.f. for W the c.f.

$$\Phi^*_W(t) = \Phi_1(t)\,\Phi^*_2(t) \tag{1}$$

where, for $h \in H$ (usually with $H = \{1, 2\}$, $H = \{1, \ldots, 4\}$ or $H = \{1, \ldots, 6\}$),

$$\left.\frac{d}{dt^h}\,\Phi^*_2(t)\right|_{t=0} = \left.\frac{d}{dt^h}\,\Phi_2(t)\right|_{t=0} \tag{2}$$

in such a way that, if we write $\Phi^*_2(t)$ as a function of the sample size (say n) and the overall number of variables involved (say p), that is, if we write $\Phi^*_2(t) \equiv \Phi^*_2(t; n, p)$, we want

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \left|\frac{\Phi_2(t) - \Phi^*_2(t; n, p)}{t}\right| dt = 0 \quad \text{and} \quad \lim_{p\to\infty} \int_{-\infty}^{\infty} \left|\frac{\Phi_2(t) - \Phi^*_2(t; n, p)}{t}\right| dt = 0$$

with

$$\begin{aligned} \int_{-\infty}^{\infty} \left|\frac{\Phi_2(t) - \Phi^*_2(t; n, p)}{t}\right| dt &\geq \int_{-\infty}^{\infty} |\Phi_1(t)| \left|\frac{\Phi_2(t) - \Phi^*_2(t; n, p)}{t}\right| dt \\ &\geq \max_w |F_W(w) - F^*_W(w; n, p)| \,, \end{aligned}$$

where $F_W(\cdot)$ is the exact c.d.f. of W and $F^*_W(\cdot\,; n, p)$ is the near-exact c.d.f. of W, corresponding to the c.f. $\Phi^*_W(t)$ in (1).

When dealing with l.r.t. (likelihood ratio test) statistics, mainly those more commonly used in Multivariate Statistics, $\Phi_1(t)$ is the c.f. of a sum of independent Logbeta r.v.'s (random variables), that are r.v.'s whose exponential has a Beta distribution, which may be alternatively written as the c.f. of a GIG (Generalized Integer Gamma) distribution, that is, the distribution of the sum of a given number of independent Gamma distributions, all with different rate parameters and integer shape parameters (see [4]), while $\Phi_2(t)$ is the c.f. of a sum of other independent Logbeta r.v.'s, which is not possible to be written under the form of the c.f. of a GIG distribution and which will be asymptotically replaced by the c.f. of a single Gamma distribution or the c.f. of a mixture of two or three Gamma distributions with the same rate parameters. This replacement is indeed well justified, since as shown in [7], and already stated and used in [13], a single Logbeta distribution may be represented under the form of an infinite mixture of Exponential distributions, a sum of independent Logbeta random variables may thus be represented under the form of an infinite mixture of sums of Exponential distributions, which are themselves mixtures of Exponential or Gamma distributions.

The near-exact distributions we are interested in this paper are exactly the ones of this second kind.

2 How may the decomposition of a complex hypothesis into more elementary hypotheses help in building near-exact distributions for the test statistic

Let us suppose we have Λ, the l.r.t (likelihood ratio test) statistic to test a given null hypothesis H_0, and that we want to write, for $W = -\log \Lambda$,

$$\Phi_W(t) = \Phi_1(t)\,\Phi_2(t)$$

(the reason why we usually want to handle the c.f. of $W = -\log \Lambda$ instead of the c.f. of Λ is that while the moments of Λ may be relatively easy to obtain and they commonly exist for any positive integer order, and even for any order not necessarily integer just above a given negative value, the expression for the c.f. of Λ may be too hard to obtain and too hard to handle; while on the other hand once obtained a near-exact distribution for W it will then be easy to obtain the corresponding near-exact distribution for $\Lambda = e^{-W}$).

It may happen that this factorization may be too hard to obtain from scratch, given the complexity of Λ itself.

But, let us suppose we may write

$$\Lambda = \prod_{j=1}^{m} \Lambda_j \tag{3}$$

where Λ_j is the l.r.t. statistic to test $H_{0j|1,\ldots,j-1}$, the j-th nested conditionally independent null hypothesis we may split H_0 into. That is, we are assuming we may write

$$H_0 \equiv H_{0m|1,\ldots,m-1} \circ \ldots \circ H_{03|1,2} \circ H_{02|1} \circ H_{01}$$

(read 'to test H_0 is equivalent to test H_{0m}, assuming $H_{0,m-1}$ through H_{01} true, after ..., after testing H_{03}, assuming H_{02} and H_{01} true, after testing H_{02}, assuming H_{01} true, after testing H_{01}) where the $H_{0j|1,\ldots,j-1}$ are all independent in the sense that under the null hypothesis H_0 it is possible to prove that the

Λ_j are independent (this is commonly the case with many l.r.t.'s, see the next section), and that further, for

$$W_j = -\log \Lambda_j \qquad (j = 1, \ldots, m)$$

we have available the factorizations

$$\Phi_{W_j}(t) = \Phi_{j1}(t)\, \Phi_{j2}(t)\,. \tag{4}$$

Then, under H_0, given the independence of the Λ_j and thus also of the W_j, we may easily write

$$\begin{aligned} \Phi_W(t) &= \prod_{j=1}^{m} \Phi_{W_j}(t) \\ &= \underbrace{\left\{ \prod_{j=1}^{m} \Phi_{j1}(t) \right\}}_{\Phi_1(t)} \underbrace{\left\{ \prod_{j=1}^{m} \Phi_{j2}(t) \right\}}_{\Phi_2(t)} \\ &= \Phi_1(t)\, \Phi_2(t)\,, \end{aligned}$$

where for the most common l.r.t. statistics used in Multivariate (and thus also in Univariate) Statistics the c.f.'s $\Phi_{j1}(t)$ may be obtained under the form of c.f.'s of GIG distributions ([5,6,13,8]) and all the $\Phi_{j2}(t)$ are c.f.'s of the sum of independent Logbeta r.v.'s. Then, $\Phi_1(t)$ itself will be the c.f. of a GIG distribution, and $\Phi_2(t)$ will be itself the c.f. of the sum of independent Logbeta r.v.'s, being thus adequately asymptotically replaced by the c.f. of a single Gamma distribution or the c.f. of the mixture of two or three Gamma distributions, verifying (2) (see [13]) and yielding this way either a GNIG (Generalized Near-Integer Gamma) distribution or a mixture of two or three GNIG distributions (the GNIG distribution is the distribution of the sum of a r.v. with a GIG distribution with an independent r.v. with a Gamma distribution with a non-integer shape parameter – for details on this distribution see [6]).

3 Examples of application

Since decompositions of the c.f.'s of the type in (4) are already available for the Wilks Λ statistic, or the l.r.t. statistic to test the independence of several sets of variables (see [5,6]), and also for the l.r.t. statistic to test sphericity (see [13]) and for the l.r.t. statistic to test the equality of several variance-covariance matrices (see [8]), we may think of l.r.t.'s whose statistic may be factorized as in (3) and where the Λ_j are the above mentioned statistics.

This was indeed what was somehow done when obtaining either the exact distribution for the generalized Wilks Λ statistic, under the form of a GIG distribution, when at most one of the sets of variables has an odd number of variables (see [4]) or when obtaining a near-exact distribution for the same statistic, under the form of a GNIG distribution, for the general case when two or more sets have an odd number of variables (see [5,6]).

In the subsections ahead we will use the following notation:

- $\Lambda_1(q,p;N_j)$, with $\Lambda_1(q,p;N_j) = \frac{\prod_{j=1}^{q}|A_j|^{N_j/2}}{|A|^{N/2}}\frac{N^{Np/2}}{\prod_{j=1}^{q}N_j^{N_jp/2}}$, to denote the l.r.t. statistic used to test $H_{01}:\Sigma_1=\ldots=\Sigma_q$, based on samples of size N_j $(j=1,\ldots,q)$ from $N_p(\underline{\mu}_j,\Sigma_j)$, with $A_j=\hat{\Sigma}_j$, $A=A_1+\ldots+a_q$ and $N=N_1+\ldots+N_q$;
- $\Lambda_2(N;p_1,\ldots,p_k)$, with $\Lambda_2(N;p_1,\ldots,p_k)=\left(\frac{|A|}{\prod_{i=1}^{k}|A_{ii}|}\right)^{N/2}$, to denote the l.r.t. statistic used to test $H_{02}:\Sigma=diag(\Sigma_{11},\ldots,\Sigma_{ii},\ldots,\Sigma_{kk})$, based on a sample of size N from $N_p(\underline{\mu},\Sigma)$, with $A=\hat{\Sigma}$, $A_{ii}=\hat{\Sigma}_{ii}$ and

$$\underline{\mu}=\left[\underline{\mu}_1,\ldots,\underline{\mu}_i,\ldots,\underline{\mu}_h\right]',\quad \Sigma=\begin{bmatrix}\Sigma_{11} & \ldots & \Sigma_{1i} & \ldots & \Sigma_{1k}\\ \vdots & \ddots & \vdots & & \vdots\\ \Sigma_{i1} & \ldots & \Sigma_{ii} & \ldots & \Sigma_{ik}\\ \vdots & & \vdots & \ddots & \vdots\\ \Sigma_{k1} & \ldots & \Sigma_{ki} & \ldots & \Sigma_{kk}\end{bmatrix};$$

- $\Lambda_3(p;N)$, with $\Lambda_3(p;N)=\left(\frac{|A|}{\left(tr\,\frac{1}{p}\,A\right)^p}\right)^{N/2}$, to denote the l.r.t. statistic used to test $H_{03}:\Sigma=\sigma^2I_p$, based on a sample of size N from $N_p(\underline{\mu},\Sigma)$, with $A=\hat{\Sigma}$.

3.1 The test of equality of several multivariate Normal distributions

It may the best well known example of the situation we are trying to illustrate, since we may write in this case ([2,12,14])

$$H_0:\underline{\mu}_1=\ldots=\underline{\mu}_q\,,\quad \Sigma_1=\ldots=\Sigma_q \tag{5}$$

with

$$H_0 \equiv H_{02|1} \circ H_{01}$$

where

$$H_{01} : \Sigma_1 = \ldots = \Sigma_q$$

and

$$H_{02|1} : \underline{\mu}_1 = \ldots = \underline{\mu}_q \\ \text{given that } \Sigma_1 = \ldots = \Sigma_q \,,$$

so that, using the notation defined at the beginning of this section, we may write the l.r.t. statistic to test H_0 in (5) as

$$\Lambda = \Lambda_2(N; p, q)\ \Lambda_1(q, p; N_j) \qquad (\text{with } N = N_1 + \ldots + N_q)\,,$$

where under H_0 in (5), $\Lambda_2(N; p, q)$ and $\Lambda_1(q, p; N_j)$ are independent.

Since factorizations of the type in (4) for the c.f. of $-\log \Lambda_2(N; p, q)$ and $-\log \Lambda_1(q, p; N_j)$ are available ([5,6,8]), the process of obtaining near-exact distributions for the l.r.t. statistic for this test may be quite easily implemented by using the approach proposed in this paper.

3.2 The sphericity test

The sphericity test itself, whose null hypothesis may be written as

$$H_0 : \Sigma = \sigma^2 I_p \tag{6}$$

may be seen as another example of a l.r.t. whose null hypothesis may be written as the composition of two conditionally independent null hypotheses (see Ch. 10, subsec. 10.7.3 in [2]), since we may indeed write H_0 in (6) once again as

$$H_0 \equiv H_{02|1} \circ H_{01}$$

where

$$H_{01} : \Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$$

and

$$H_{02|1} : \sigma_1^2 = \ldots = \sigma_p^2$$
$$\text{given that } \Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)\,,$$

so that, using the notation defined at the beginning of this section, we may write the l.r.t. statistic to test H_0 in (6) as

$$\Lambda = \Lambda_2(N; \underbrace{1, \ldots, 1}_{p})\ \Lambda_1(p, 1; N)\,,$$

where, under H_0 in (6), $\Lambda_2(N; 1, \ldots, 1)$ and $\Lambda_1(p, 1; N)$ are independent.

Although we may question the usefulness of this approach, moreover given that very accurate near-exact distributions have already been obtained for the sphericity l.r.t. statistic by [13], interestingly enough, these same authors are in the process of obtaining other near-exact distributions for this statistic exactly through the use of this approach, which besides having a simpler formulation may even be more accurate than the previously developed ones namely for larger sample sizes and larger number of variables.

3.3 Generalized sphericity tests

Other tests onto which the authors intend to apply the approach proposed in this paper are extended versions of the sphericity test, which we may call the 'multi-sample scalar-block sphericity test' and the 'multi-sample matrix-block sphericity test'.

For the multi-sample scalar-block sphericity test we have

$$H_0 : \Sigma_1 = \ldots = \Sigma_q = \begin{bmatrix} \sigma_1^2 I_{p_1} & & 0 \\ & \ddots & \\ 0 & & \sigma_k^2 I_{p_k} \end{bmatrix} \tag{7}$$

with

$$H_0 \equiv H_{03|1,2} \circ H_{02|1} \circ H_{01}$$

where

$$H_{01} : \Sigma_1 = \ldots = \Sigma_q\,,$$

$$\begin{aligned} H_{02|1} &: \Sigma = diag(\Sigma_{11}, \ldots, \Sigma_{kk}) \\ &\qquad \text{given that } \Sigma_1 = \ldots = \Sigma_q = \Sigma\,, \end{aligned}$$

where Σ is of dimensions $p{\times}p$, while Σ_{ii} $(i = 1, \ldots, k)$ of dimensions $p_i{\times}p_i$ is the i-th diagonal block of Σ, with $p = \sum_{i=1}^{k} p_i$, and

$$\begin{aligned} H_{03|1,2} &: \Sigma_{ii} = \sigma_i^2 I_{p_i}\,, \quad \text{for } i = 1, \ldots, k \\ &\qquad \text{given that } \Sigma = diag(\Sigma_{11}, \ldots, \Sigma_{kk}) \text{ and} \\ &\qquad \text{given that } \Sigma_1 = \ldots = \Sigma_q = \Sigma\,, \end{aligned}$$

so that we may write the l.r.t. statistic to test H_0 in (7), for $N = N_1 + \ldots + N_q$ as

$$\Lambda = \Lambda_1(q, p; N_j)\ \Lambda_2(N; p_1, \ldots, p_k)\ \prod_{i=1}^{k} \Lambda_3(p_i; N)\,. \tag{8}$$

Since, under H_0 in (7), it is possible to prove the independence of all the statistics in (8) and since factorizations of the type in (4) are available for the c.f.'s of the logarithms of all the statistics in (8) ([5,6,8,13]), the process of obtaining near-exact distributions for the l.r.t. statistic for this test may be implemented by using the approach proposed in this paper.

Concerning the multi-sample matrix-block sphericity test, its null hypothesis may be written as

$$H_0 : \Sigma_1 = \ldots = \Sigma_q = \Delta \otimes I_k \left(= \begin{bmatrix} \Delta & & 0 \\ & \ddots & \\ 0 & & \Delta \end{bmatrix} \right) \tag{9}$$

with

$$H_0 \equiv H_{03|1,2} \circ H_{02|1} \circ H_{01}\,,$$

where

$$H_{01} : \Sigma_1 = \ldots = \Sigma_q \,(= \Sigma)\,,$$

$$\begin{aligned} H_{02|1} &: \Sigma = diag(\Sigma_{11}, \ldots, \Sigma_{kk}) \\ &\qquad \text{given that } \Sigma_1 = \ldots = \Sigma_q \,(= \Sigma)\,, \end{aligned}$$

where Σ is of dimensions $p{\times}p$, while Σ_{ii} $(i = 1, \dots, k)$ of dimensions $p^*{\times}p^*$ is the i-th diagonal block of Σ, with $p = kp^*$, and

$$\begin{aligned} H_{03|1,2} : \Sigma_{11} = \dots = \Sigma_{kk} (= \Delta) & \\ & \text{given that } \Sigma = diag(\Sigma_{11}, \dots, \Sigma_{kk}) \text{ and} \\ & \text{given that } \Sigma_1 = \dots = \Sigma_q = \Sigma\,, \end{aligned}$$

so that we may write the l.r.t. statistic to test H_0 in (9), for $N = N_1 + \dots + N_q$ as

$$\Lambda = \Lambda_1(q, p; N_j)\ \Lambda_2(N; \underbrace{p^*, \dots, p^*}_{k})\ \Lambda_1(k, p^*; N)\,. \tag{10}$$

Since, under H_0 in (9), it is possible to prove the independence of all the statistics in (10) and since factorizations of the type in (4) are available for the c.f.'s of the logarithms of all the statistics in (10) ([5,6,13]), the process of obtaining near-exact distributions for the l.r.t. statistic for this test may once again be implemented by using the approach proposed in this paper.

4 Final remark

We may even plan to go beyond the tests presented in the previous section, namely those in subsection 3, with virtually no limit, by adequately nesting the three elementary multivariate likelihood ratio tests (referred at the beginning of the previous section). In every case we will still be able to obtain near-exact distributions for the overall test statistics, in situations where the construction of well-fit asymptotic distributions, using the usual techniques, is too complicated or even virtually impossible.

Acknowledgment

This research was financially supported by the Portuguese Foundation for Science and Technology, through the Center for Mathematics and its Applications (CMA) from the New University of Lisbon.

References

[1] R. P. Alberto, C. A. Coelho, C. A., 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 (2007) 1612-1626.

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

[3] C. A. Coelho, Generalized Canonical Analysis, Ph.D. Thesis, The University of Michigan, Ann Arbor, MI, 1992.

[4] C. A. Coelho, The Generalized Integer Gamma distribution - a basis for distributions in Multivariate Statistics, J. Multivariate analysis 64 (1998) 86-102.

[5] C. A. Coelho, The Generalized Integer Gamma distribution as an asymptotic replacement for the logBeta random variable - Applications, American Journal of Mathematical and Management Sciences, 23 (2003) 383-399.

[6] C. A. Coelho, 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 independent Beta random variables, J. Multivariate Analysis, 89 (2004) 191-218.

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

[8] C. A. Coelho, F. J. Marques, Near-exact distributions for the likelihood ratio test statistic for testing equality of several variance-covariance matrices, The New University of Lisbon, Mathematics Department, Technical Report #12/2007 (submitted for publication).

[9] C. A. Coelho, J. T. Mexia, On the exact and near-exact distributions of statistics used in generalized F tests, The New University of Lisbon, Mathematics Department, Technical Report #23/2006 (submitted for publication).

[10] L. M. Grilo, C. A. Coelho, 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 (2007) 1560-1575.

[11] L. M. Grilo, C. A. Coelho, The exact and near-exact distributions for the Wilks Lambda statistic used in the test of independence of two sets of variables. The New University of Lisbon, Mathematics Department, Technical Report #15/2007 (submitted for publication).

[12] A. M. Kshirsagar, Multivariate Analysis, Marcel Dekker, Inc., New York, 1972.

[13] F. J. Marques, C. A. Coelho, Near-exact distributions for the sphericity likelihood ratio test statistic, Journal of Statistical Planning and Inference (2008) (in print).

[14] R. J. Muirhead, Aspects of Multivariate Statistical Theory, J. Wiley & Sons, New York, 1986.

[15] J. W. Mauchly, Significance test for sphericity of a normal n-variate distribution, Annals of Mathematical Statistics, 11 (1940) 204-209.