ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

REVSTAT
Statistical Journal

Volume 16, No.3
July 2018

REVSTAT

Statistical Journal

Catalogação Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E., 2003- . - 30 cm
Semestral. - Continuação de : Revista de Estatística = ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726

CREDITS

- **PUBLISHER**
  - *Instituto Nacional de Estatística, I.P. (INE, I.P.)*
    Av. António José de Almeida, 2
    1000-043 LISBOA
    PORTUGAL
    Tel.: + 351 21 842 61 00
    Fax: + 351 21 845 40 84
    Web site: http://www.ine.pt
    Customer Support Service
    + 351 218 440 695
- **COVER DESIGN**
  - *Mário Bouçadas, designed on the stain glass window at INE by the painter Abel Manta*
- **LAYOUT AND GRAPHIC DESIGN**
  - *Carlos Perpétuo*
- **PRINTING**
  - *Instituto Nacional de Estatística, I.P.*
- **EDITION**
  - *140 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*
- **PRICE [VAT included]**
  - *€ 9,00*

INDEX

Abstracted/indexed in: *Current Index to Statistics, DOAJ, Google Scholar, Journal Citation Reports/Science Edition, Mathematical Reviews, Science Citation Index Expanded®, SCOPUS* and *Zentralblatt für Mathematic.*

REVSTAT – Statistical Journal
Volume 16, Number 3, July 2018, 279–294

TWO-FACTOR EXPERIMENTS WITH SPLIT UNITS CONSTRUCTED BY CYCLIC DESIGNS AND SQUARE LATTICE DESIGNS

Authors: Kazuhiro Ozawa
– Department of Nursing, Gifu College of Nursing, Japan
ozawa@gifu-cn.ac.jp

Iwona Mejza
– Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poland
iwona.mejza@up.poznan.pl

Stanisław Mejza
– Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poland
stanislaw.mejza@up.poznan.pl

Shinji Kuriki
– Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University, Japan
kuriki@ms.osakafu-u.ac.jp

Received: April 2016 Revised: August 2016 Accepted: August 2016

Abstract:

- We consider nested row-column designs with split units for a two-factor experiment. The most optimal design in this case is that of using for the whole plots a Latin square while for the subplot treatments with a completely randomized design for each whole plot. Such a design, in fact optimal, utilizes many experimental units and quite a large space. Hence to construct new designs of reduced size of the experiment we use a cyclic design for the whole plot treatments and a square lattice design for the subplot treatments. The proposed designs are generally balanced and they allow for giving the stratum efficiency factors, especially useful to design of experiments.

Key-Words:

- *cyclic design; general balance; square lattice design; stratum efficiency factor.*

AMS Subject Classification:

- 62K15, 62K10, 05B05.

1. INTRODUCTION

In many biological and agricultural (field) experiments, a nested row-column design with split units is often used. The design is for a two-factor experiment of split-plot type with b blocks. The first factor A has v_1 levels $A_1, A_2, ..., A_{v_1}$ and the second factor B has v_2 levels $B_1, B_2, ..., B_{v_2}$. Each block is divided into k_1 rows and k_2 columns and these k_1k_2 units are treated as whole plots. Moreover, each whole plot is divided into k_3 subplots. The levels of A and B are applied to the whole plots (called whole plot treatments) and the subplots (called subplot treatments), respectively. Such a design is called a nested row-column design with split units.

Kachlicka and Mejza (1996) considered a mixed linear model with fixed treatment effects and random block, row, column, whole plot and subplot effects for the nested row-column design with split units. The hth factorial treatment combination effect τ_h is defined by

$$\tau_h = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

for $h = (i-1)v_2 + j$, $i = 1, 2, ..., v_1$ and $j = 1, 2, ..., v_2$, where μ is the general mean, α_i denotes the main effect of the ith whole plot treatment A_i, β_j denotes the main effect of the jth subplot treatment B_j and $(\alpha\beta)_{ij}$ denotes the interaction effect of A_i and B_j. Here $\sum_{i=1}^{v_1} \alpha_i = 0$, $\sum_{j=1}^{v_2} \beta_j = 0$, $\sum_{i=1}^{v_1} (\alpha\beta)_{ij} = 0$ for $j = 1, 2, ..., v_2$ and $\sum_{j=1}^{v_2} (\alpha\beta)_{ij} = 0$ for $i = 1, 2, ..., v_1$. The mixed linear model results from a four-step randomization, i.e., the randomization of blocks, the randomization of rows within each block, the randomization of columns within each block and the randomization of subplots within each whole plot. This kind of randomization leads us to an experiment with orthogonal block structure as defined by Nelder (1965a, 1965b) and the multistratum analysis proposed by Nelder (1965a, 1965b) and Houtman and Speed (1983) can be applied to the analysis of data in the experiment. In this case, we have five strata, except zero stratum connected with the general mean only, (I) inter-block stratum, (II) inter-row stratum, (III) inter-column stratum, (IV) inter-whole plot stratum and (V) inter-subplot stratum. The statistical properties of the nested row-column design with split units are strictly connected with the eigenvalues and the eigenvectors of the stratum information matrices for the treatment combinations. The stratum information matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$ and $\mathbf{A}_5$ are given by

$$\mathbf{A}_1 = \frac{1}{k_1k_2k_3}\mathbf{N}_0\mathbf{N}_0' - \frac{r}{v}\mathbf{J}_v, \quad \mathbf{A}_2 = \frac{1}{k_2k_3}\mathbf{N}_1\mathbf{N}_1' - \frac{1}{k_1k_2k_3}\mathbf{N}_0\mathbf{N}_0', \tag{1.1}$$

$$\mathbf{A}_3 = \frac{1}{k_1k_3}\mathbf{N}_2\mathbf{N}_2' - \frac{1}{k_1k_2k_3}\mathbf{N}_0\mathbf{N}_0', \tag{1.2}$$

$$\mathbf{A}_4 = \frac{1}{k_3}\mathbf{N}_3\mathbf{N}_3' - \frac{1}{k_1k_3}\mathbf{N}_2\mathbf{N}_2' - \frac{1}{k_2k_3}\mathbf{N}_1\mathbf{N}_1' + \frac{1}{k_1k_2k_3}\mathbf{N}_0\mathbf{N}_0' \tag{1.3}$$

and

$$\mathbf{A}_5 = r\mathbf{I}_v - \frac{1}{k_3}\mathbf{N}_3\mathbf{N}_3', \tag{1.4}$$

where $v = v_1v_2$, $\mathbf{N}_0$, $\mathbf{N}_1$, $\mathbf{N}_2$ and $\mathbf{N}_3$ are the incidence matrices for the treatment combinations vs. blocks, rows, columns and whole plots, respectively, $\mathbf{I}_v$ is the identity matrix of order v and $\mathbf{J}_v$ is the $v \times v$ matrix with every element unity. Here we assume that every treatment combination A_iB_j ($i = 1, 2, ..., v_1, j = 1, 2, ..., v_2$) occurs in precisely r blocks and the treatment combinations are ordered lexicographically.

A generally balanced design was firstly introduced by Nelder (1965a, 1965b), for which the stratum information matrices are spanned by a common set of eigenvectors. Let $\boldsymbol{s}_0, \boldsymbol{s}_1, ..., \boldsymbol{s}_{v-1}$ be the mutually orthonormal common eigenvectors of the stratum information matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$ and $\mathbf{A}_5$. Since $\mathbf{A}_f\mathbf{1}_v = \mathbf{0}$ for $f = 1, 2, 3, 4, 5$, $\frac{1}{\sqrt{v}}\mathbf{1}_v$ may be chosen as the first eigenvector $\boldsymbol{s}_0$, where $\mathbf{1}_v$ is the $v \times 1$ vector of unit elements. Let ξ_{fh} be an eigenvalue of a matrix $\mathbf{A}_f^* = r^{-1}\mathbf{A}_f$ corresponding to the eigenvector $\boldsymbol{s}_h$ for $f = 1, 2, 3, 4, 5$ and $h = 1, 2, ..., v-1$. Then, a basic contrast of the treatment effects (see Pearce et al. (1974)) is defined by $\boldsymbol{s}_h'\boldsymbol{\tau}$ for $h = 1, 2, ..., v-1$, where $\boldsymbol{\tau}$ is the $v \times 1$ vector of the treatment effects. The eigenvalue ξ_{fh} can be identified as a stratum efficiency factor of the design concerning estimation of the hth basic contrast in the fth stratum for $f = 1, 2, 3, 4, 5$ and $h = 1, 2, ..., v-1$ (see, Houtman and Speed (1983)).

Many experiments require a long time or a large space (units) often making it impossible to carry out a conventional, complete (orthogonal) design of the considered type. For example, in agricultural field experiments, because of soil fertility it is difficult to find units (plots) fulfilling restrictions concerning the homogeneity of blocks, rows, columns, whole plots or subplots. Then, to satisfy the main experimental principles it is necessary to design the experiment as an incomplete (non-orthogonal) one. Such an experiment usually utilizes smaller units, with respect to size and also utilizes smaller number of units (the experiment is cheaper). The problem is to find an incomplete design proper to experimental material structure and optimal with respect to statistical properties of the design.

Kuriki et al. (2009), Mejza et al. (2009) and Mejza and Kuriki (2013) constructed nested row-column designs with split units by the Kronecker product of the incidence matrices of two designs. They used a Youden square for the whole plot treatments and various proper designs for the subplot treatments. Mejza et al. (2014) have used a balanced incomplete block design with nested rows and columns instead of the Youden square to construct a nested row-column design with split units. The designs obtained by this way need usually a large number of units. In this paper, we construct a nested row-column design with split units by a modified Kronecker product (called a semi-Kronecker product) of the incidence

matrices of two designs. We use a cyclic design for the whole plot treatments and a square lattice design for the subplot treatments. We give the stratum efficiency factors for such a nested row-column design with split units, which has the general balance property.

These designs have smaller numbers of blocks than the conventional experiments. Therefore, they would be useful in practice, for example, the reduction of the experimental expenses and effort, and the easier implementation of the experiments by using the well-known cyclic designs and square lattice designs in the literature (see, John (1987), John and Williams (1995) and Raghavarao (1971), etc.).

Other variants of incomplete split plot designs are given, for example, by Mejza and Mejza (1996), Ozawa et al. (2004), Aastveit et al. (2009), Mejza et al. (2012) and Kuriki et al. (2012).

2. A CONSTRUCTION BY A CYCLIC DESIGN AND A SQUARE LATTICE DESIGN

Firstly, we need the semi-Kronecker product (see, Khatri and Rao (1968) and Mejza, Kuriki and Mejza (2001)) of two matrices that will be used to construct nested row-column designs with split units. Suppose that two matrices $\mathbf{E}$ and $\mathbf{F}$ are divided into the same number of submatrices as follows:

$$\mathbf{E} = (\mathbf{E}_1 : \mathbf{E}_2 : \cdots : \mathbf{E}_m) \quad \text{and} \quad \mathbf{F} = (\mathbf{F}_1 : \mathbf{F}_2 : \cdots : \mathbf{F}_m).$$

Then, the semi-Kronecker product $\mathbf{E} \tilde{\otimes} \mathbf{F}$ is defined by

$$\mathbf{E} \tilde{\otimes} \mathbf{F} = (\mathbf{E}_1 \otimes \mathbf{F}_1 : \mathbf{E}_2 \otimes \mathbf{F}_2 : \cdots : \mathbf{E}_m \otimes \mathbf{F}_m),$$

where $\otimes$ denotes the usual Kronecker product.

Next, we need a cyclic design and a square lattice design. Let V be a set of v treatments and let $\mathcal{B}$ be a collection of subsets (called blocks) of V. A design $(V, \mathcal{B})$ is denoted by $\mathrm{D}(v, r, k)$ if every treatment occurs in precisely r blocks and each block contains k treatments. Let Z_v be the additive group of integers modulo v and let $(V, \mathcal{B})$ be a $\mathrm{D}(v, r, k)$ with $V = Z_v$ for which if $\{a_1, a_2, ..., a_k\}$ is a block, then $\{a_1 + 1, a_2 + 1, ..., a_k + 1\}$ is also a block. A set of blocks $\{\{a_1 + i, a_2 + i, ..., a_k + i\} \,|\, i \in Z_v\}$ is called a cyclic class and a block taken arbitrarily from each cyclic class is called an initial block. If the collection $\mathcal{B}$ of blocks is divided into some cyclic classes, then $(V, \mathcal{B})$ is said to be cyclic and it is denoted by $\mathrm{CD}(v, r, k)$. Here we consider only a case where the number of blocks in each cyclic class is v.

Let $(V, \mathcal{B})$ be a $\mathrm{D}(v, r, k)$. If the collection $\mathcal{B}$ of blocks can be grouped in such a way that every treatment occurs precisely once in every group (called a resolution class), then $(V, \mathcal{B})$ is said to be resolvable. A resolvable $\mathrm{D}(v, r, k)$ $(V, \mathcal{B})$ such that $v = s^2$, $r \leq s + 1$ and $k = s$ for a positive integer s is called a square lattice design if any two blocks from different resolution classes contain just one common treatment, and it is denoted by $\mathrm{SLD}(s^2, r, s)$. If $r = s + 1$, it is called a balanced square lattice design and it is well known that there exists a balanced square lattice design if s is a prime or a prime power (see, Raghavarao (1971)).

Now we construct a nested row-column design with split units. Let $(V_A, \mathcal{B}_A)$ be a $\mathrm{CD}(v_A, r_A, k_A)$ with $m = r_A/k_A$ initial blocks. Each cyclic class of $(V_A, \mathcal{B}_A)$ is treated as a block with k_A rows and v_A columns such that the columns are blocks of $\mathcal{B}_A$ and that every treatment of V_A occurs precisely once in each row. Such a design is denoted by $\mathcal{D}_A$. An $\mathrm{SLD}(s^2, m, s)$ is denoted by $\mathcal{D}_B$. The whole plot treatments occur in $\mathcal{D}_A$ and the subplot treatments occur in $\mathcal{D}_B$. We construct a nested row-column design, say $\mathcal{D}$, with split units embedding each block of the ith resolution class of $\mathcal{D}_B$ in every whole plot of the ith block of $\mathcal{D}_A$ for $i = 1, 2, ..., m$. The parameters of $\mathcal{D}$ are $v_1 = v_A$, $v_2 = s^2$, $b = ms$, $r = mk_A = r_A$, $k_1 = k_A$, $k_2 = v_A$ and $k_3 = s$.

Example 2.1. We use an A-efficient cyclic design $\mathrm{CD}(6, 6, 3)$ with initial blocks $\{0, 1, 2\}$ and $\{0, 1, 3\}$ given by John (1987). From two cyclic classes of this design, we have the following two blocks with 3 rows and 6 columns of $\mathcal{D}_A$:

$$\begin{array}{|c|c|c|c|c|c|}\hline 0&1&2&3&4&5\\\hline 1&2&3&4&5&0\\\hline 2&3&4&5&0&1\\\hline\end{array} \quad \text{and} \quad \begin{array}{|c|c|c|c|c|c|}\hline 0&1&2&3&4&5\\\hline 1&2&3&4&5&0\\\hline 3&4&5&0&1&2\\\hline\end{array}.$$

We also use a square lattice design $\mathrm{SLD}(9, 2, 3)$ $\mathcal{D}_B = (V_B, \mathcal{B}_B)$ with $V_B = \{1, 2, ..., 9\}$. The following columns are 6 blocks of $\mathcal{D}_B$:

$$\begin{array}{|c|}\hline 1\\\hline 2\\\hline 3\\\hline\end{array}\ \begin{array}{|c|}\hline 4\\\hline 5\\\hline 6\\\hline\end{array}\ \begin{array}{|c|}\hline 7\\\hline 8\\\hline 9\\\hline\end{array} \quad \text{and} \quad \begin{array}{|c|}\hline 1\\\hline 4\\\hline 7\\\hline\end{array}\ \begin{array}{|c|}\hline 2\\\hline 5\\\hline 8\\\hline\end{array}\ \begin{array}{|c|}\hline 3\\\hline 6\\\hline 9\\\hline\end{array},$$

where the first resolution class is constituted by the first 3 blocks and the second one is constituted by the remaining blocks. We construct a nested row-column design $\mathcal{D}$ with split units embedding each block of the first (second) resolution class of $\mathcal{D}_B$ in every whole plot of the first (second) block of $\mathcal{D}_A$, replacing the treatments $0, 1, 2, 3, 4, 5$ of $\mathcal{D}_A$ with $A_1, A_2, A_3, A_4, A_5, A_6$ and the treatments $1, 2, ..., 9$ of $\mathcal{D}_B$ with $B_1, B_2, ..., B_9$. The design $\mathcal{D}$ has 6 blocks as follows:

A_1			A_2			A_3			A_4			A_5			A_6		
B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
A_2			A_3			A_4			A_5			A_6			A_1		
B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
A_3			A_4			A_5			A_6			A_1			A_2		
B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3

A_1			A_2			A_3			A_4			A_5			A_6		
B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6
A_2			A_3			A_4			A_5			A_6			A_1		
B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6
A_3			A_4			A_5			A_6			A_1			A_2		
B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6	B_4	B_5	B_6

A_1			A_2			A_3			A_4			A_5			A_6		
B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9
A_2			A_3			A_4			A_5			A_6			A_1		
B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9
A_3			A_4			A_5			A_6			A_1			A_2		
B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9	B_7	B_8	B_9

A_1			A_2			A_3			A_4			A_5			A_6		
B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7
A_2			A_3			A_4			A_5			A_6			A_1		
B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7
A_4			A_5			A_6			A_1			A_2			A_3		
B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7	B_1	B_4	B_7

A_1			A_2			A_3			A_4			A_5			A_6		
B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8
A_2			A_3			A_4			A_5			A_6			A_1		
B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8
A_4			A_5			A_6			A_1			A_2			A_3		
B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8	B_2	B_5	B_8

A_1			A_2			A_3			A_4			A_5			A_6		
B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9
A_2			A_3			A_4			A_5			A_6			A_1		
B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9
A_4			A_5			A_6			A_1			A_2			A_3		
B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9	B_3	B_6	B_9

□

We note that if the nested row-column design with split units is constructed by the usual Kronecker product of the incidence matrices (see, Mejza et al. (2014)), then the number of blocks becomes $m^2 s = 12$. Generally, the number of blocks of a nested row-column design with split units by the Kronecker product is m times larger than those of a nested row-column design with split units by the semi-Kronecker product.

Let

$$\mathbf{N}_A = (\mathbf{N}_{A1} : \mathbf{N}_{A2} : \cdots : \mathbf{N}_{Am}) \quad \text{and} \quad \mathbf{N}_B = (\mathbf{N}_{B1} : \mathbf{N}_{B2} : \cdots : \mathbf{N}_{Bm})$$

be the incidence matrices of the cyclic design $\mathrm{CD}(v_A, r_A, k_A)$ and the square lattice design $\mathrm{SLD}(s^2, m, s)$, where $\mathbf{N}_{Ai}$ and $\mathbf{N}_{Bi}$ correspond to the ith cyclic and resolution classes, respectively. By the definition of the square lattice design,

$$\mathbf{N}'_{Bi}\mathbf{N}_{Bi} = s\mathbf{I}_s \quad \text{and} \quad \mathbf{N}'_{Bi}\mathbf{N}_{Bj} = \mathbf{J}_s \tag{2.1}$$

hold for $i, j = 1, 2, ..., m$, $i \neq j$. Then, the incidence matrix $\mathbf{N}_2$ of the nested row-column design $\mathcal{D}$ with split units is given by the semi-Kronecker product of $\mathbf{N}_A$ and $\mathbf{N}_B$, i.e.,

$$\mathbf{N}_2 = \mathbf{N}_A \tilde{\otimes} \mathbf{N}_B = (\mathbf{N}_{A1} \otimes \mathbf{N}_{B1} : \mathbf{N}_{A2} \otimes \mathbf{N}_{B2} : \cdots : \mathbf{N}_{Am} \otimes \mathbf{N}_{Bm})$$

in a suitable order of columns of $\mathcal{D}$, and the concurrence matrices $\mathbf{N}_0\mathbf{N}'_0$, $\mathbf{N}_1\mathbf{N}'_1$, $\mathbf{N}_2\mathbf{N}'_2$ and $\mathbf{N}_3\mathbf{N}'_3$ of $\mathcal{D}$ are given by

$$\mathbf{N}_0\mathbf{N}'_0 = \sum_{i=1}^{m} \left(k_A^2 \mathbf{J}_{v_A} \otimes \mathbf{N}_{Bi}\mathbf{N}'_{Bi}\right) = k_A^2 \mathbf{J}_{v_A} \otimes \mathbf{N}_B\mathbf{N}'_B, \tag{2.2}$$

$$\mathbf{N}_1\mathbf{N}'_1 = \sum_{i=1}^{m} \left(k_A \mathbf{J}_{v_A} \otimes \mathbf{N}_{Bi}\mathbf{N}'_{Bi}\right) = k_A \mathbf{J}_{v_A} \otimes \mathbf{N}_B\mathbf{N}'_B, \tag{2.3}$$

$$\mathbf{N}_2\mathbf{N}'_2 = \sum_{i=1}^{m} \left(\mathbf{N}_{Ai}\mathbf{N}'_{Ai} \otimes \mathbf{N}_{Bi}\mathbf{N}'_{Bi}\right) \tag{2.4}$$

and

$$\mathbf{N}_3\mathbf{N}'_3 = \sum_{i=1}^{m} \left(k_A \mathbf{I}_{v_A} \otimes \mathbf{N}_{Bi}\mathbf{N}'_{Bi}\right) = k_A \mathbf{I}_{v_A} \otimes \mathbf{N}_B\mathbf{N}'_B. \tag{2.5}$$

3. STRATUM EFFICIENCY FACTORS FOR $\mathcal{D}$

In this section, we give the stratum efficiency factors for the nested row-column design $\mathcal{D}$ with split units constructed in Section 2. To find the stratum efficiency factors, it is necessary to find the eigenvalues of the stratum information matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$ and $\mathbf{A}_5$ of $\mathcal{D}$. It is easy to find these eigenvalues if $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$ and $\mathbf{A}_5$ have the common eigenvectors, i.e., if $\mathcal{D}$ is generally balanced. It follows, from (2.1), that

$$\mathbf{N}_{Bi}\mathbf{N}'_{Bi}\mathbf{N}_{Bj}\mathbf{N}'_{Bj} = \mathbf{J}_{s^2} \tag{3.1}$$

holds for $i, j = 1, 2, \cdots, m$, $i \neq j$. From (3.1), it is easily verified that the concurrence matrices $\mathbf{N}_0\mathbf{N}'_0$, $\mathbf{N}_1\mathbf{N}'_1$, $\mathbf{N}_2\mathbf{N}'_2$ and $\mathbf{N}_3\mathbf{N}'_3$ given in (2.2)–(2.5) are mutually commutative. Thus, by use of (1.1)–(1.4), the stratum information matrices $\mathbf{A}_1$,

$\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$ and $\mathbf{A}_5$ are mutually commutative, which means that $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$ and $\mathbf{A}_5$ have the common eigenvectors. Therefore, $\mathcal{D}$ is generally balanced.

In order to find the common eigenvectors of the stratum information matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$ and $\mathbf{A}_5$, i.e., those of the concurrence matrices $\mathbf{N}_0\mathbf{N}_0'$, $\mathbf{N}_1\mathbf{N}_1'$, $\mathbf{N}_2\mathbf{N}_2'$ and $\mathbf{N}_3\mathbf{N}_3'$, we consider the eigenvectors of $\mathbf{N}_{Ai}\mathbf{N}_{Ai}'$ for the ith cyclic class of the cyclic design $\mathrm{CD}(v_A, r_A, k_A)$ and those of $\mathbf{N}_{Bi}\mathbf{N}_{Bi}'$ for the ith resolution class of the square lattice design $\mathrm{SLD}(s^2, m, s)$ for $i = 1, 2, ..., m$. For the incidence matrix $\mathbf{N}_A$ of the $\mathrm{CD}(v_A, r_A, k_A)$, since $\mathbf{N}_{A1}\mathbf{N}_{A1}', \mathbf{N}_{A2}\mathbf{N}_{A2}', ..., \mathbf{N}_{Am}\mathbf{N}_{Am}'$ are symmetric circulant matrices, these matrices have the mutually orthonormal common eigenvectors, which are denoted by $\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_{v_A-1}$ with $\boldsymbol{x}_0 = \frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}$. The corresponding eigenvalues of $\mathbf{N}_{Ai}\mathbf{N}_{Ai}'$ are given by

$$\theta_j^{(i)} = \sum_{h=0}^{v_A-1} \lambda_h^{(i)} \cos\left(\frac{2\pi jh}{v_A}\right)$$

for $i = 1, 2, ..., m$ and $j = 0, 1, ..., v_A - 1$, where $\lambda_h^{(i)}$ $(h \neq 0)$ denotes the number of blocks containing two treatments 0 and h in the ith cyclic class of the $\mathrm{CD}(v_A, r_A, k_A)$ and $\lambda_0^{(i)} = k_A$. In particular, $\theta_0^{(i)} = k_A^2$ and the corresponding eigenvector is $\boldsymbol{x}_0 = \frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}$ (see, John (1987) and John and Williams (1995)). These eigenvalues and common eigenvectors are summarized in the following table:

Table 1: Eigenvalues and common eigenvectors of $\mathbf{N}_{Ai}\mathbf{N}_{Ai}'$ in the $\mathrm{CD}(v_A, r_A, k_A)$.

Eigenvalues	Common eigenvectors
k_A^2	$\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}$
$\theta_j^{(i)}$	$\boldsymbol{x}_j$ $(j = 1, 2, ..., v_A - 1)$

Similarly, for the incidence matrix $\mathbf{N}_B$ of the $\mathrm{SLD}(s^2, m, s)$, from (2.1), $\mathbf{N}_{Bi}\mathbf{N}_{Bi}'$ has the eigenvalues s and 0 with multiplicities s and $s(s-1)$ for each $i = 1, 2, ..., m$. From (3.1), $\mathbf{N}_{B1}\mathbf{N}_{B1}', \mathbf{N}_{B2}\mathbf{N}_{B2}', ..., \mathbf{N}_{Bm}\mathbf{N}_{Bm}'$ are mutually commutative, so these matrices have the common eigenvectors. Let $\mathbf{Q} = (\boldsymbol{q}_0, \boldsymbol{q}_1, ..., \boldsymbol{q}_{s-1})$ be an orthogonal matrix of order s with $\boldsymbol{q}_0 = \frac{1}{\sqrt{s}}\mathbf{1}_s$. For each $i = 1, 2, ..., m$, from (2.1), the mutually orthonormal eigenvectors of $\mathbf{N}_{Bi}\mathbf{N}_{Bi}'$ corresponding to the eigenvalue s are given by

$$\boldsymbol{z}_{ip} = \frac{1}{\sqrt{s}}\mathbf{N}_{Bi}\boldsymbol{q}_p$$

for $p = 0, 1, ..., s-1$. In particular, $\boldsymbol{z}_{i0} = \frac{1}{s}\mathbf{1}_{s^2}$. The eigenvectors $\boldsymbol{z}_{ip}$ are also the eigenvectors of $\mathbf{N}_{Bh}\mathbf{N}_{Bh}'$ $(h \neq i)$ for any other resolution class, and the eigenvalues

of $\mathbf{N}_{Bh}\mathbf{N}'_{Bh}$ corresponding to $\boldsymbol{z}_{i0}$ and $\boldsymbol{z}_{ip}$ $(p \neq 0)$ are s and 0, respectively. Furthermore, the mutually orthonormal common eigenvectors of $\mathbf{N}_{B1}\mathbf{N}'_{B1}, \mathbf{N}_{B2}\mathbf{N}'_{B2}$, ..., $\mathbf{N}_{Bm}\mathbf{N}'_{Bm}$ corresponding to the eigenvalue 0 are denoted by $\boldsymbol{z}^*_q$ for $q = 1, 2, ...,$ $s^2 - m(s-1) - 1$. These eigenvalues and common eigenvectors are summarized in Table 2.

Table 2: Eigenvalues and common eigenvectors of $\mathbf{N}_{Bi}\mathbf{N}'_{Bi}$ in the SLD(s^2, m, s).

Eigenvalues				Common eigenvectors
$\mathbf{N}_{B1}\mathbf{N}'_{B1}$	$\mathbf{N}_{B2}\mathbf{N}'_{B2}$	$\cdots$	$\mathbf{N}_{Bm}\mathbf{N}'_{Bm}$	
s	s	$\cdots$	s	$\frac{1}{s}\mathbf{1}_{s^2}$
s	0	$\cdots$	0	$\boldsymbol{z}_{1p}$ $(p = 1, 2, ..., s-1)$
0	s	$\cdots$	0	$\boldsymbol{z}_{2p}$ $(p = 1, 2, ..., s-1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	0	$\cdots$	s	$\boldsymbol{z}_{mp}$ $(p = 1, 2, ..., s-1)$
0	0	$\cdots$	0	$\boldsymbol{z}^*_q$ $(q = 1, 2, ..., s^2 - m(s-1) - 1)$

Combining the eigenvectors of Table 1 and Table 2, we consider the following 6 sets of vectors:

$$(1)\ \frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A} \otimes \frac{1}{s}\mathbf{1}_{s^2}, \quad (2)\ \boldsymbol{x}_j \otimes \frac{1}{s}\mathbf{1}_{s^2}, \quad (3)\ \frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A} \otimes \boldsymbol{z}_{ip},$$

$$(4)\ \frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A} \otimes \boldsymbol{z}^*_q, \quad (5)\ \boldsymbol{x}_j \otimes \boldsymbol{z}_{ip}, \quad (6)\ \boldsymbol{x}_j \otimes \boldsymbol{z}^*_q$$

for $i = 1, 2, ..., m,\ j = 1, 2, ..., v_A - 1, p = 1, 2, ..., s-1$ and $q = 1, 2, ..., s^2 - m(s-1) - 1$. The vectors of (1)–(6) are mutually orthonormal and the total number of the vectors is $v_A s^2$. We show that the vectors of (1)–(6) are the common eigenvectors of $\mathbf{N}_0\mathbf{N}'_0$, $\mathbf{N}_1\mathbf{N}'_1$, $\mathbf{N}_2\mathbf{N}'_2$ and $\mathbf{N}_3\mathbf{N}'_3$, and we find the corresponding eigenvalues of $\mathbf{N}_0\mathbf{N}'_0$, $\mathbf{N}_1\mathbf{N}'_1$, $\mathbf{N}_2\mathbf{N}'_2$ and $\mathbf{N}_3\mathbf{N}'_3$.

Firstly, we take into account the matrix $\mathbf{N}_0\mathbf{N}'_0$. For (1), we have, from (2.2), Table 1 and Table 2,

$$\begin{aligned}
&\mathbf{N}_0\mathbf{N}'_0\left(\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A} \otimes \frac{1}{s}\mathbf{1}_{s^2}\right) = \left(k_A^2\mathbf{J}_{v_A} \otimes \mathbf{N}_B\mathbf{N}'_B\right)\left(\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A} \otimes \frac{1}{s}\mathbf{1}_{s^2}\right) \\
&= \left(k_A^2\mathbf{J}_{v_A}\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}\right) \otimes \left(\mathbf{N}_B\mathbf{N}'_B\frac{1}{s}\mathbf{1}_{s^2}\right) = \left(v_A k_A^2\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}\right) \otimes \left(ms\frac{1}{s}\mathbf{1}_{s^2}\right) \\
&= mv_A k_A^2 s\left(\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A} \otimes \frac{1}{s}\mathbf{1}_{s^2}\right).
\end{aligned}$$

The corresponding eigenvalue is $mv_Ak_A^2s$.

For (2), we have

$$\mathbf{N}_0\mathbf{N}_0'\left(\boldsymbol{x}_j\otimes\frac{1}{s}\mathbf{1}_{s^2}\right)=\left(k_A^2\mathbf{J}_{v_A}\boldsymbol{x}_j\right)\otimes\left(\mathbf{N}_B\mathbf{N}_B'\frac{1}{s}\mathbf{1}_{s^2}\right)=\mathbf{0}.$$

The corresponding eigenvalue is zero for each $i=1,2,...,m$ and $j=1,2,...,v_A-1$.

For (3), we have

$$\begin{aligned}&\mathbf{N}_0\mathbf{N}_0'\left(\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}\otimes\boldsymbol{z}_{ip}\right)=\left(k_A^2\mathbf{J}_{v_A}\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}\right)\otimes\left(\mathbf{N}_B\mathbf{N}_B'\boldsymbol{z}_{ip}\right)\\&=\left(v_Ak_A^2\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}\right)\otimes\left(\sum_{h=1}^{m}\mathbf{N}_{Bh}\mathbf{N}_{Bh}'\boldsymbol{z}_{ip}\right)=\left(v_Ak_A^2\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}\right)\otimes\left(s\boldsymbol{z}_{ip}\right)\\&=v_Ak_A^2s\left(\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}\otimes\boldsymbol{z}_{ip}\right).\end{aligned}$$

The corresponding eigenvalue is $v_Ak_A^2s$ for each $i=1,2,...,m$ and $p=1,2,...,s-1$.

For (4), we have

$$\begin{aligned}&\mathbf{N}_0\mathbf{N}_0'\left(\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}\otimes\boldsymbol{z}_q^*\right)=\left(k_A^2\mathbf{J}_{v_A}\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}\right)\otimes\left(\mathbf{N}_B\mathbf{N}_B'\boldsymbol{z}_q^*\right)\\&=\left(v_Ak_A^2\frac{1}{\sqrt{v_A}}\mathbf{1}_{v_A}\right)\otimes\left(\sum_{i=1}^{m}\mathbf{N}_{Bi}\mathbf{N}_{Bi}'\boldsymbol{z}_q^*\right)=\mathbf{0}.\end{aligned}$$

The corresponding eigenvalue is zero for $q=1,2,...,s^2-m(s-1)-1$. Moreover, for (5) and (6), the eigenvalue is also zero.

Similarly, from (2.3)–(2.5), we can show that the vectors of (1)–(6) are also the eigenvectors of $\mathbf{N}_1\mathbf{N}_1'$, $\mathbf{N}_2\mathbf{N}_2'$ and $\mathbf{N}_3\mathbf{N}_3'$. The corresponding eigenvalues of $\mathbf{N}_0\mathbf{N}_0'$, $\mathbf{N}_1\mathbf{N}_1'$, $\mathbf{N}_2\mathbf{N}_2'$ and $\mathbf{N}_3\mathbf{N}_3'$ are summarized in the table below:

Table 3: Eigenvalues and common eigenvectors of $\mathbf{N}_0\mathbf{N}_0'$, $\mathbf{N}_1\mathbf{N}_1'$, $\mathbf{N}_2\mathbf{N}_2'$ and $\mathbf{N}_3\mathbf{N}_3'$.

Eigenvalues				Common eigenvectors
$\mathbf{N}_0\mathbf{N}_0'$	$\mathbf{N}_1\mathbf{N}_1'$	$\mathbf{N}_2\mathbf{N}_2'$	$\mathbf{N}_3\mathbf{N}_3'$	
$mv_Ak_A^2s$	mv_Ak_As	mk_A^2s	mk_As	(1)
0	0	$\sum_{i=1}^{m}\theta_j^{(i)}s$	mk_As	(2)
$v_Ak_A^2s$	v_Ak_As	k_A^2s	k_As	(3)
0	0	$\theta_j^{(i)}s$	k_As	(5)
0	0	0	0	(4), (6)

Here $i=1,2,...,m$ and $j=1,2,...,v_A-1$.

The vectors (1)–(6) are also the common eigenvectors of the stratum information matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$ and $\mathbf{A}_5$. By use of (1.1)–(1.4) and Table 3, the stratum efficiency factors for $\mathcal{D}$ can be calculated as in the following table:

Table 4: Stratum efficiency factors for $\mathcal{D}$.

Type of contrasts	Number of contrasts	Strata				
		I	II	III	IV	V
A	$v_A - 1$	0	0	ω_j	$1 - \omega_j$	0
B	$m(s-1)$	$1/m$	0	0	0	$1 - 1/m$
	$s^2 - m(s-1) - 1$	0	0	0	0	1
$A \times B$	$m(v_A - 1)(s-1)$	0	0	ξ_{ij}	$1/m - \xi_{ij}$	$1 - 1/m$
	$(v_A - 1)\{s^2 - m(s-1) - 1\}$	0	0	0	0	1

for $i = 1, 2, ..., m$ and $j = 1, 2, ..., v_A - 1$, where A and B denote the basic contrasts among the main effects of whole plot and subplot treatments, respectively, $A \times B$ denotes the basic contrasts among the interaction effects, $\xi_{ij} = \theta_j^{(i)}/(mk_A^2)$ and $\omega_j = \sum_{i=1}^{m} \xi_{ij}$. The eigenvectors of (2), (3)–(4) and (5)–(6) define the basic contrasts A, B and $A \times B$, respectively. We use Table 4 in order to improve the estimators for the basic contrasts of the treatment effects combining the estimators obtained from the strata I, III, IV and V. This procedure was proposed by Nelder (1965a, 1965b) and Houtman and Speed (1983). Especially, we see that some basic contrasts of B and $A \times B$ are estimable with full efficiency.

Example 3.1. For the nested row-column design $\mathcal{D}$ with split units given in Example 2.1, $m = 2$, $v_A = 6$, $k_A = 3$, $s = 3$, $\theta_1^{(1)} = 4$, $\theta_2^{(1)} = 0$, $\theta_3^{(1)} = 1$, $\theta_4^{(1)} = 0$, $\theta_5^{(1)} = 4$, $\theta_1^{(2)} = 1$, $\theta_2^{(2)} = 3$, $\theta_3^{(2)} = 1$, $\theta_4^{(2)} = 3$ and $\theta_5^{(2)} = 1$. Thus, by use of Table 4, the stratum efficiency factors can be calculated as in the following table:

Table 5: Stratum efficiency factors for $\mathcal{D}$ given in Example 2.1.

Type of contrasts	Number of contrasts	Strata				
		I	II	III	IV	V
A	1	0	0	1/9	8/9	0
	2	0	0	1/6	5/6	0
	2	0	0	5/18	13/18	0
B	4	1/2	0	0	0	1/2
	4	0	0	0	0	1
$A \times B$	4	0	0	0	1/2	1/2
	4	0	0	1/6	1/3	1/2
	4	0	0	2/9	5/18	1/2
	8	0	0	1/18	4/9	1/2
	20	0	0	0	0	1

4. REMARKS

In the design of experiments at least a few aspects play crucial roles. The first one concerns proper use of available structure of experimental units. The general rule, for example, in field agricultural experiments constitutes that smaller units better satisfy requirements concerning homogeneity of stratum units. In addition, usually smaller errors are associated after randomizations with these units.

The second aspect concerns statistical properties of designs. Using complete, orthogonal designs leads to the best unbiased estimators of the estimable functions of linear model parameters. In this work, we use a randomization-derived linear model (random block effect describing structure of units) with treatment (combination) effects being fixed. The structure of units and randomization performed lead to a design which possesses orthogonal block structure. In a complete case, the estimators of all estimable treatment effect functions are BLUEs. This means that the design is optimal from a point of view of statistical properties. Such a design can be used for our experiment if it is possible. However, many times there exist some limitations in available structure of experimental units (material). Then in our experiment some incomplete design can be applied only.

The new problem concerns how to choose an incomplete design that fits to the structure of experimental units, is optimal for the most interesting treatment effect functions, and is not so expensive (utilizes small as possible number of units of proper size). In the worse case we can use any incomplete design. Then it is difficult to describe the statistical properties of the proposed design.

The experimenter usually makes a ranking of linear functions of treatment effects (contrasts) with respect to a scientific interest and an aim of the experiment. It would be helpful to have a design with known efficiencies of all estimable treatment effect functions. This property has a generally balanced design (see, for example, Mejza (1992) and Bailey (1994)). General balance aids interpretation; the design which is generally balanced with respect to meaningful contrasts may be superior to a technically optimal design. For generally balanced designs, we can identify the meaning of the treatment effect contrasts and their efficiency factors (cf. Table 2, Table 3 and Table 4). Hence we restrict our searching in the class of generally balanced designs.

Those considered here (nested row-column designs with split units) can be characterized by a few component block designs. We are looking for methods allowing for generation of new row-column designs with split units by using some known incomplete block designs instead of component designs. The Kronecker product of the component incomplete block designs is often used for constructing

new designs with split units. The final design possesses optimal properties, but it utilizes many experimental units (high cost of the experiment). To overcome this problem (size of the experiment) we proposed to use of the semi-Kronecker product as defined in Section 2 instead of the ordinary Kronecker product. The final design is much smaller and also possesses desirable statistical properties (see Example 2.1). Moreover using the semi-Kronecker product to generate new designs leads to much smaller number of units and smaller size. In the Example 2.1, one block of the complete design will have 6 rows and 6 columns while the whole plot consists of 9 units. For example, in agricultural field experiments (where such designs are very often used) it would be difficult to find so many homogeneous plots. In these cases the use of an incomplete design is recommended. In this paper, we construct a nested row-column design with split units by the semi-Kronecker product of the incidence matrices of a cyclic design for the whole plot treatments and a square lattice design for the subplot treatments. We give the stratum efficiency factors for such a nested row-column design with split units having the general balance property.

Although we proposed the new method for constructing the design in the class of nested row-column designs with split units, we still need new methods for constructing designs in the considered class which will lead to general balanced designs with desirable statistical properties and will have reasonable size. Naturally, in the future work for construction optimal row-column designs with nested structures someone can look for new methods and for another class of incomplete block designs as considered in the paper.

ACKNOWLEDGMENTS

The authors would like to thank the referees for their valuable comments and suggestions. Moreover, the authors would like to thank to JSPS and PAN (Polish Academy of Sciences) for giving them the opportunity to carry out this joint research in Japan and in Poland. The research was partially supported by JSPS, Grant-in-Aid for Scientific Research (B) 15H03636 and Grant-in-Aid for Young Scientists (B) 25730020.

REFERENCES

[1] Aastveit, A.H.; Almøy, T.; Mejza, I. and Mejza, S. (2009). Individual control treatment in split-plot experiments, *Statistical Papers*, **50**, 697–710.

[2] Bailey, R.A. (1994). *General balance: artificial theory or practical relevance?*. In "Proceedings of the International Conference on Linear Statistical Inference LINSTAT '93" (T. Caliński and R. Kala, Eds.), Kluwer, Dordrecht, 171–184.

[3] Houtman, A.M. and Speed, T.P. (1983). Balance in designed experiments with orthogonal block structure, *The Annals of Statistics*, **11**, 1069–1085.

[4] John, J.A. (1987). *Cyclic designs*, Chapman and Hall, New York.

[5] John, J.A. and Williams, E.R. (1995). *Cyclic and Computer Generated Designs. Second edition*, Chapman and Hall, London.

[6] Kachlicka, D. and Mejza, S. (1996). Repeated row-column designs with split units, *Computational Statistics & Data Analysis*, **21**, 293–305.

[7] Khatri, C.G. and Rao, C.R. (1968). Solutions to some functional equations and their applications to characterization of probability distributions, *Sankhyā. The Indian Journal of Statistics. Series A*, **30**, 167–180.

[8] Kuriki, S.; Mejza, S.; Mejza, I. and Kachlicka, D. (2009). Repeated Youden squares with subplot treatments in a proper incomplete block design, *Biometrical Letters*, **46**, 153–162.

[9] Kuriki, S.; Mejza, I. and Mejza, S. (2012). Incomplete split-plot designs supplemented by a single control, *Communications in Statistics – Theory and Methods*, **41**, 2490–2502.

[10] Mejza, I. and Mejza, S. (1996). Incomplete split plot generated by GDP-BIB(2), *Calcutta Statistical Association Bulletin*, **46**, 117–127.

[11] Mejza, I.; Kuriki, S. and Mejza, S. (2001). Balanced square lattice designs in split-block designs, *Colloquium Biometryczne*, **31**, 97–103.

[12] Mejza, I.; Mejza, S. and Kuriki, S. (2012). A method of constructing incomplete split-plot designs supplemented by control treatments and their analysis, *Journal of Statistical Theory and Practice*, **6**, 204–219.

[13] Mejza, I.; Mejza, S. and Kuriki, S. (2014). Two-factor experiment with split units constructed by a BIBRC, *SUT Journal of Mathematics*, **50**, 343–352.

[14] Mejza, S. (1992). On some aspects of general balance in designed experiments, *Statistica*, **52**, 263–278.

[15] Mejza, S.; Kuriki, S. and Kachlicka, D. (2009). Repeated Youden squares with subplot treatments in a group-divisible design, *Journal of Statistics and Applications*, **4**, 369–377.

[16] Mejza, S. and Kuriki, S. (2013). *Youden square with split units*. In "Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications" (J.L. da Silva, F. Caeiro, I. Natário and C.A. Braumann, Eds.), Springer Berlin Heidelberg, 3–10.

[17] Nelder, J.A. (1965a). The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance, *Proceedings of the Royal Society. London, A*, **283**, 147–162.

[18] Nelder, J.A. (1965b). The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance, *Proceedings of the Royal Society. London, A*, **283**, 163–178.

[19] Ozawa, K.; Mejza, S.; Jimbo, M.; Mejza, I. and Kuriki, S. (2004). Incomplete split-plot designs generated by some resolvable balanced designs, *Statistics & Probability Letters*, **68**, 9–15.

[20] Pearce, S.C.; Caliński, T. and Marshall, T.F. de C. (1974). The basic contrasts of an experimental design with special reference to the analysis of data, *Biometrika*, **61**, 449–460.

[21] Raghavarao, D. (1971). *Constructions and combinatorial problems in design of experiments*, John Wiley & Sons, New York.

REVSTAT – Statistical Journal
Volume 16, Number 3, July 2018, 295–313

WHICH EFFECT SIZE MEASURE IS APPROPRIATE FOR ONE-WAY AND TWO-WAY ANOVA MODELS? A MONTE CARLO SIMULATION STUDY

Authors: Soner Yigit
– Faculty of Agriculture, Animal Science Department, Biometry and Genetics Unit, Canakkale Onsekiz Mart University, Canakkale, Turkey
soneryigit@comu.edu.tr

Mehmet Mendes
– Faculty of Agriculture, Animal Science Department, Biometry and Genetics Unit, Canakkale Onsekiz Mart University, Canakkale, Turkey
mmendes@comu.edu.tr

Received: April 2016 Revised: September 2016 Accepted: September 2016

Abstract:

- It is very important to report some effect size measures that will show if the observed differences among the groups are also of practical significance along with statistical significance while reporting statistical analysis results. Performances of four commonly used effect size measures (Eta-Squared, Partial Eta Squared, Omega Squared and Epsilon Squared) were compared for one and two-way ANOVA models under 3000 different conditions. Results of simulation runs showed that the Epsilon and Omega-Squared estimates were quite unbiased when compared to Eta and Partial Eta-Squared which are directly reported by commonly used statistical packages while reporting ANOVA results. Thus, it could be concluded that reporting Epsilon or Omega-Squared is more appropriate to evaluate the practical significance of observed differences along with P-values.

Key-Words:

- *effect size measure; omega-squared; epsilon-squared; biased, simulation.*

AMS Subject Classification:

- 62-07.

1. INTRODUCTION

The Analysis of Variance Technique (ANOVA-F) is used in comparing the differences between two or more group means [1, 2]. However, it does not show how different the compared group means are from each other or how much of the difference occurred in the dependent variable results from the groups. In other words, while testing the statistical significance of the differences between the levels of independent variable, ANOVA-F test does not give any information about its practical significance [3]. On the other hand, in practice, there is a widespread belief that the smaller the P-value, which is used as the criterion of statistical significance, the more effective or the stronger the levels of the factor the effect of which is researched [4]. Nevertheless, statistical significance is affected by the size of the studied sample. Even very small differences could be found to be statistically significant with very large size samples, large effect sizes may not be found statistically significant with small size samples [5, 6]. Hays [7] reported that the effect size measures are as important as hypothesis testing. Recently, a significant portion of the scientific journals request reporting some effect size measures along with the P-value when reporting statistical analysis results [8] because calculating or estimating the effect size, along with helping in understanding how big the differences between the compared means are, could help in obtaining information about the practical significance of the observed difference and in determining what % of the variation of the analyzed property is described by the considered factor(s). Thus, while reporting analysis of variance results, reporting some effect size measures along with the P-values, which show statistical significance, provides significant benefits [9, 10]. For this purpose, different effect size measures are proposed [7, 11, 12, 13, 14, 15, 16]. The most popular effect size measures for analysis of variance models are found to be $\hat{\eta}^2$ (Eta-Squared), $\hat{\eta}_p^2$ (Partial Eta Squared), $\hat{\omega}^2$ (Omega Squared) and $\hat{\epsilon}^2$ (Epsilon Squared) [3, 8, 9, 17, 18]. However, it is remarkable that the performances of these effect size measures are shown for only one-way analysis of variance and that this is done under quite limited experimental conditions [8, 9, 17, 18]. Moreover, it is a reality that a significant number of the experiments conducted in practice involve in factorial designs. Thus, showing the performances of the aforementioned effect size measures in terms of in factorial design models, as well as the one-way analysis of variance model, would be beneficial. At the same time, contradicting results of the some of the limited studies comparing effect size measures (e.g. [17, 9]) could cause errors. Therefore, performances of the aforementioned effect size measures should be shown in detail under many conditions confronted in practice. Through this, it will be both possible to show the performances of the aforementioned effect size measures under many experimental conditions and to increase the opportunity of generalization of the obtained results. In the study conducted with this point of view, it is aimed to compare the performances of Eta-Squared ($\hat{\eta}^2$), Partial Eta Squared ($\hat{\eta}_p^2$), Omega Squared ($\hat{\omega}^2$) and Epsilon Squared ($\hat{\epsilon}^2$),

which are found as the most popular effect size measures in practice, for one and two-way analysis of variance models. By this means, it will be possible to determine the most convenient effect size measure or measures according to the considered experimental conditions.

2. MATERIAL AND METHOD

Materials for this study consists of random numbers generated by a Monte Carlo simulation technique. In the generation of the random numbers, the RNNOA, RNBET and RNCHI functions of IMSL library of Microsoft Fortran Power Station Developer Studio are used. In this study $\hat{\eta}^2$, $\hat{\eta}_p^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ are compared in terms of their performances (bias) under different conditions such as group number or sub-group number, distribution shape, sample size, variance ratio and population effect size. Performances of these effect sizes are determined after 1.000.000 simulation experiments for each of the considered experimental conditions. Experimental conditions considered in the study for One-Way and Two-Way Analysis of Variance models are given together on Table 1 and Table 2.

Table 1: Experimental Conditions for the One-Way Anova.

Statistical model	$Y_{ij}=\mu + \alpha_i + e_{ij}$
Number of Group (k)	3, 4, 5 and 10
Distribution	N(0,1), $\beta(10,10)$, $\beta(5,10)$, $\beta(10,5)$ and $\chi^2(3)$
μ_1:μ_2:...:μ_k	0:0:...:0.30, 0:0:...:0.60, 0:0:...:0.90 and 0:0:...:1.20
σ_1^2:σ_2^2:...:σ_k^2	1:1:...:1, 1:1:...:9, and 1:1:...:20
Number of replication (n)	5, 10, 20, 30 and 50
Number of simulation	1.000.000

Table 2: Experimental Conditions for the Two-Way Anova.

Statistical model	$Y_{ijk}=\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}$
Experimental design (rxc)	2×2, 2×3, 4×2, 3×3, 4×3 and 4×4
Distribution	N(0,1), $\beta(10,10)$, $\beta(5,10)$, $\beta(10,5)$ and $\chi^2(3)$
μ_{11}:μ_{12}:...:μ_{rc}	0:0:...:0.30, 0:0:...:0.60, 0:0:...:0.90 and 0:0:...:1.20
σ_{11}^2:σ_{12}^2:...:σ_{rc}^2	1:1:...:1, 1:1:...:9, and 1:1:...:20
Number of replication (n)	2, 3, 5, 10 and 30
Number of simulation	1.000.000

In order to compare effect size measures in terms of their performances, firstly n numbers are generated from the distributions considered in the study.

Then, generated numbers are subjected to a transformation as $(X_{ij}\text{-}\mu)/\sigma$. Afterwards, certain constant numbers (0.3, 0.6, 0.9 and 1.2) are added to the last group or sub-group in order to create differences between population means. Finally, for all considered experiment conditions and in terms of all effect sizes, population effect size is estimated 1.000.000 times, then means and standard errors are calculated.

2.1. Effect Size Measures

In order to estimate population effect size, many effect size measures are developed. In this study, Eta-squared, Partial Eta-Squared, Omega-Squared, and Epsilon-Squared, which are found as the most popular effect size measures, are taken into consideration [3, 7, 11].

$$\hat{\eta}^2 = \frac{SS_{Effect}}{SS_{Total}}, \tag{2.1}$$

$$\hat{\eta}_p^2 = \frac{SS_{Effect}}{SS_{Total} + SS_{Error}}. \tag{2.2}$$

In a One Way ANOVA-F test, $\hat{\eta}^2$ and $\hat{\eta}_p^2$ are equal [19].

$$\hat{\epsilon}^2 = \frac{SS_{Effect} - df_{Effect} MS_{Error}}{SS_{Total}} \quad [11], \tag{2.3}$$

$$\hat{\omega}^2 = \frac{SS_{Effect} - df_{Effect} MS_{Error}}{SS_{Total} + MS_{Error}} \quad [7], \tag{2.4}$$

where SS_{Total}: Total sum of squares, SS_{Effect}: Sum of squares of effect, SS_{Error}: Error sum of squares, MS_{Error}: Mean square error and df_{Effect}: Degree of freedom of effect.

2.2. Determining Population Effect Size

When determining population effect sizes, Cohen's f value is considered. The relationship between population effect size and Cohen's f value is as follows.

$$\eta^2 = \frac{f^2}{1 + f^2}, \tag{2.5}$$

$$f = \frac{\sigma_\mu}{\sigma}, \tag{2.6}$$

$$\sigma_\mu = \sqrt{\frac{\sum_{i=1}^{k}(\mu_i - \mu)^2}{k}}, \tag{2.7}$$

$$\sigma = \sqrt{\frac{\sigma_1^2 + \sigma_2^3 + \sigma_3^2}{k}}, \tag{2.8}$$

where σ_μ: Standard deviation of population means, σ: Pooled standard deviation, μ_i: i. population mean, μ: Mean of population means, k: Compared population number [20].

2.2.1. How to get population effect size in the One-Way fixed effects ANOVA model

Table 3: η^2 for the One Way ANOVA-F test.

k	$\mu_1 : \mu_2 :...: \mu_k$	$\sigma_1^2 : \sigma_2^2 : ... : \sigma_k^2$		
		1:1:...:1	1:1:...:9	1:1:...:20
3	0:0:...:0.3	0.01961	0.00543	0.00272
	0:0:...:0.6	0.07407	0.02135	0.01079
	0:0:...:0.9	0.15254	0.04679	0.02396
	0:0:...:1.2	0.24242	0.08027	0.04181
4	0:0:...:0.3	0.01660	0.00559	0.00293
	0:0:...:0.6	0.06323	0.02200	0.01160
	0:0:...:0.9	0.13185	0.04819	0.02573
	0:0:...:1.2	0.21260	0.08257	0.04485
5	0:0:...:0.3	0.01420	0.00551	0.00299
	0:0:...:0.6	0.05446	0.02167	0.01186
	0:0:...:0.9	0.11473	0.04748	0.02629
	0:0:...:1.2	0.18726	0.08140	0.04580
10	0:0:...:0.3	0.00803	0.00448	0.00279
	0:0:...:0.6	0.03138	0.01768	0.01105
	0:0:...:0.9	0.06795	0.03892	0.02452
	0:0:...:1.2	0.11473	0.06716	0.04278

If we want to compare the differences between three population means, it is found as follows
$\mu_1 = 0$, $\mu_2 = 0$ and $\mu_3 = 1.2$,
$\sigma_1^2 = 1$, $\sigma_2^2 = 1$ and $\sigma_3^2 = 20$,

$$\mu = \frac{\sum_{\mathrm{i}=1}^{\mathrm{k}}(\mu_\mathrm{i})}{\mathrm{k}} = \frac{0+0+1.2}{3} = 0.4,$$

$$\sigma_\mu = \sqrt{\frac{\sum_{i=1}^{k}(\mu_i - \mu)^2}{k}} = \sqrt{\frac{(0-0.4)^2 + (0-0.4)^2 + (1.2-0.4)^2}{3}} = 0.56568,$$

$$\sigma = \sqrt{\frac{\sigma_1^2 + \sigma_2^3 + \sigma_3^2}{k}} = \sqrt{\frac{1+1+20}{3}} = 2.70801,$$

$$f = \frac{\sigma_\mu}{\sigma} = \frac{0.56568}{2.70801} = 0.20889,$$

$$\eta^2 = \frac{f^2}{1+f^2} = \frac{0.20889^2}{1+0.20889^2} = 0.04181.$$

Population effect sizes calculated in this way for the One Way Analysis of Variance are given on Table 3.

2.2.2. How to get population effect sizes for the Two-Way Fixed Effects ANOVA model

In a 2^2 factorial design,

	c_1	c_2	μ_i σ_i^2
r_1	$\mu_{11} = 0$ $\sigma_{11}^2 = 1$	$\mu_{12} = 0$ $\sigma_{12}^2 = 1$	$\mu_{1.} = 0$ $\sigma_{1.}^2 = 1$
r_2	$\mu_{21} = 0$ $\sigma_{21}^2 = 1$	$\mu_{22} = 1.2$ $\sigma_{22}^2 = 20$	$\mu_{2.} = 0.6$ $\sigma_{2.}^2 =?$

If $X = \{x_1, x_2, ..., x_N\}$, $Y = \{y_1, y_2, ..., y_N\}$ and $Z = \{x_1, x_2, ..., x_N, y_1, y_2, ..., y_N\}$, As is known, $\mu_Z = (\mu_X + \mu_Y)/2$. If $\mu_X = \mu_Y$, $\sigma_Z^2 = (\sigma_X^2 + \sigma_Y^2)/2$. However, if $\mu_X \neq \mu_Y$, then $\sigma_Z^2 \neq (\sigma_X^2 + \sigma_Y^2)/2$.
If $\mu_X \neq \mu_Y$ then,

$$\sigma_Z^2 = \frac{\mu_X^2 + \sigma_X^2 + \mu_Y^2 + \sigma_Y^2 - 2\mu_Z^2}{2}. \tag{2.9}$$

If k population is considered as one population, then the variance of the obtained new population is calculated as follows.

$$\sigma_Z^2 = \frac{\sum_{i=1}^{k}(\mu_i^2 + \sigma_i^2) - k\mu_Z^2}{k}. \tag{2.10}$$

This formula is empirically verified. In that case, since $\mu_{21} \neq \mu_{22}$ and $\sigma_{2.}^2 \neq (\sigma_{21}^2 + \sigma_{22}^2)/2$, then

$$\sigma_{2.}^2 = \frac{\mu_{21}^2 + \sigma_{21}^2 + \mu_{22}^2 + \sigma_{22}^2 - 2\mu_{2.}^2}{2} = \frac{0^2 + 1 + 1.20^2 + 20 - 2(0.6^2)}{2} = 10.86.$$

Similarly, it is found that $\sigma^2_{.2} = 10.86$.
In factorial experiments, following equality is valid for the population effect size.

$$\eta^2_{Model} = \eta^2_r + \eta^2_c + \eta^2_{rxc}. \tag{2.11}$$

Thus, in order to find the effect size in terms of interaction (η^2_{rxc}), first η^2_{Model}, η^2_r and η^2_c should be calculated.
Calculation of η^2_{Model}

$$\mu = \frac{\sum_i^r \sum_j^c \mu_{ij}}{rc} = \frac{0+0+0+1.2}{4} = 0.3,$$

$$\sigma_\mu = \sqrt{\frac{\sum_i^r \sum_j^c (\mu_{ij}-\mu)^2}{rc}}$$

$$= \sqrt{\frac{(0-0.3)^2+(0-0.3)^2+(0-0.3)^2+(1.2-0.3)^2}{(2)(2)}} = 0.51961,$$

$$\sigma = \sqrt{\frac{\sigma^2_{11}+\sigma^2_{12}+\sigma^2_{21}+\sigma^2_{22}}{rc}} = \sqrt{\frac{1+1+1+20}{(2)(2)}} = 2.39791,$$

$$f = \frac{\sigma_\mu}{\sigma} = \frac{0.51961}{2.39791} = 0.21669,$$

$$\eta^2_{Model} = \frac{f^2}{1+f^2} = \frac{0.21669^2}{1+0.21669^2} = 0.04485.$$

Calculation of η^2_r

$$\mu = \frac{\sum_i^r \mu_{i.}}{r} = \frac{0+0.6}{2} = 0.3,$$

$$\sigma_\mu = \sqrt{\frac{\sum_i^r (\mu_{i.}-\mu)^2}{r}} = \sqrt{\frac{(0-0.3)^2+(0.6-0.3)^2}{2}} = 0.3,$$

$$\sigma = \sqrt{\frac{\sigma^2_{1.}+\sigma^2_{2.}}{r}} = \sqrt{\frac{1+10.86}{2}} = 2.43516,$$

$$f = \frac{\sigma_\mu}{\sigma} = \frac{0.3}{2.43516} = 0.12319,$$

$$\eta^2_r = \frac{f^2}{1+f^2} = \frac{0.12319^2}{1+0.12319^2} = 0.01495.$$

Calculation of η^2_c

$$\mu = \frac{\sum_j^c \mu_{.j}}{c} = \frac{0+0.6}{2} = 0.3,$$

$$\sigma_\mu = \sqrt{\frac{\sum_i^c (\mu_{.j}-\mu)^2}{c}} = \sqrt{\frac{(0-0.3)^2+(0.6-0.3)^2}{2}} = 0.3,$$

$$\sigma = \sqrt{\frac{\sigma^2_{.1}+\sigma^2_{.2}}{c}} = \sqrt{\frac{1+10.86}{2}} = 2.43516,$$

$$f = \frac{\sigma_\mu}{\sigma} = \frac{0.3}{2.43516} = 0.12319,$$
$$\eta_c^2 = \frac{f^2}{1+f^2} = \frac{0.12319^2}{1+0.12319^2} = 0.01495.$$

Calculation of η_{rxc}^2

$$\eta_{rxc}^2 = \eta_{Model}^2 - \eta_r^2 - \eta_c^2 = 0.04485 - 0.01495 - 0.01495 = 0.01495.$$

Population effect sizes calculated in this way for Interaction Effect are given on Table 4.

Table 4: η_{rxc}^2 for the Two-Way ANOVA-F test.

k	$\mu_1 : \mu_2 :...: \mu_k$	$\sigma_1^2 : \sigma_2^2 : ... : \sigma_k^2$		
		1:1:...:1	1:1:...:9	1:1:...:20
2×2	0:0:...:0.3	0.00553	0.00186	0.00098
	0:0:...:0.6	0.02108	0.00733	0.00387
	0:0:...:0.9	0.04395	0.01606	0.00858
	0:0:...:1.2	0.07087	0.02752	0.01495
2×3	0:0:...:0.3	0.00494	0.00213	0.00120
	0:0:...:0.6	0.01905	0.00839	0.00474
	0:0:...:0.9	0.04045	0.01840	0.01052
	0:0:...:1.2	0.06667	0.03158	0.01832
3×3	0:0:...:0.3	0.00441	0.00234	0.00142
	0:0:...:0.6	0.01717	0.00924	0.00565
	0:0:...:0.9	0.03704	0.02032	0.01253
	0:0:...:1.2	0.06226	0.03501	0.02186
4×2	0:0:...:0.3	0.00418	0.00210	0.00125
	0:0:...:0.6	0.01624	0.00827	0.00494
	0:0:...:0.9	0.03488	0.01818	0.01096
	0:0:...:1.2	0.05832	0.03129	0.01911
4×3	0:0:...:0.3	0.00372	0.00224	0.00145
	0:0:...:0.6	0.01460	0.00885	0.00575
	0:0:...:0.9	0.03178	0.01953	0.01276
	0:0:...:1.2	0.05405	0.03377	0.02228
4×4	0:0:...:0.3	0.00315	0.00210	0.00144
	0:0:...:0.6	0.01239	0.00832	0.00573
	0:0:...:0.9	0.02719	0.01840	0.01274
	0:0:...:1.2	0.04669	0.03195	0.02228

3. RESULTS

In this study, five different distribution shapes, three different variance ratios, five different sample sizes, four different effect size magnitudes, four group combinations (k=3, 4, 5 and 10) in one-way analysis of variance analysis, six subgroup combinations (2×2, 2×3, 3×3, 4×2, 4×3 and 4×4) in two-way analysis of variance, totally 3000 different experimental conditions are considered. Thus, all of the results could not be presented in the essay. Obtained results are given on Figure 5-24 for One-way analysis of variance, on Figure 25-54 for Two-way analysis of variance, and on Supplementary Appendix together. Furthermore, some experiment results that reflect the results significantly are summarized on Figure 1 and 2 for the One-way analysis of variance and on Figure 3 and 4 for the Two-way analysis of variance.

3.1. Results of The One-Way Analysis of Variance

Comparing independent group means that are taken from normal distribution and variances of which are homogeneous, while $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give quite unbiased results, $\hat{\eta}^2$ gives quite biased results. Besides, as long as the variances are homogeneous, slight [$\beta(10, 10)$] or moderate [$\beta(5, 10)$ and $\beta(10, 5)$] deviations from normality does not affect the realized estimations in terms of the three effect sizes. Under these conditions, although there is a negligible difference between $\hat{\epsilon}^2$ and $\hat{\omega}^2$, $\hat{\epsilon}^2$ gives the most unbiased estimations. When variances are homogeneous, excessive skewness and kurtosis [$\chi^2(3)$] affect estimations of the three effect sizes negatively. However, both $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give more unbiased estimations compared to $\hat{\eta}^2$. Although $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give results quite close to each other, $\hat{\omega}^2$ gives more unbiased results compared to $\hat{\epsilon}^2$ under these experimental conditions. When variances are homogeneous, regardless of the distribution shape and sample size, as the number of groups increase, estimations of $\hat{\eta}^2$ diverge from η^2, whereas estimations of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ approach to η^2. Additionally, depending on the increase in group number, differences between $\hat{\epsilon}^2$ and $\hat{\omega}^2$ decrease gradually. For example; when n=10 and k=3, 4, 5 and 10, bias of $\hat{\eta}^2$ ranges between 4.80-6.45%, 5.76-6.95%, 6.39-7.30% and 7.95-8.30%, bias of $\hat{\epsilon}^2$ ranges between 0.40-1.3%, 0.29-0.97%, 0.24-0.72% and 0.08-0.28% and bias of $\hat{\omega}^2$ ranges between 0.87-1.0%, 0.62-0.71%, 0.43-0.54% and 0.18-0.21% the difference between $\hat{\epsilon}^2$ and $\hat{\omega}^2$ when variances are heterogeneous is smaller than when variances are homogeneous. In case variances get heterogeneous too, $\hat{\eta}^2$ gives quite biased and irregular results. Choosing compared groups from symmetric distributions [N(0,1) and $\beta(10, 10)$] and heterogeneous variances do not affect estimations of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ negatively. In addition to this, the difference between $\hat{\epsilon}^2$ and $\hat{\omega}^2$ while variances are

heterogeneous is smaller than that while variances are homogeneous. However, bias of the distribution from symmetry increased the bias of the estimations made by $\hat{\epsilon}^2$ and $\hat{\omega}^2$ a little. This situation becomes much more significant especially when variances are excessively heterogeneous (20 times). However again, they give quite unbiased results compared to $\hat{\eta}^2$. When variances are heterogeneous, while an increase in group number negatively affects $\hat{\eta}^2$, but does not affect $\hat{\epsilon}^2$ and $\hat{\omega}^2$ significantly. Regardless of the compared group number, distribution shape, variance ratios and population means, depending on the increase in sample size, it is seen that estimations gradually approach to η^2 in terms of the three effect size ($\hat{\eta}^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$). Furthermore, it is seen that the most biased results are given by $\hat{\eta}^2$ under all considered experimental conditions. Additionally, as the difference between means decreases, in other words as the population effect size decreases (η^2), while bias of the estimations of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ gradually decrease, bias of the estimations of $\hat{\eta}^2$ gradually increase. Regardless of the experimental conditions, as the sample size decreases (especially when n=5), estimations show severe bias in terms of $\hat{\eta}^2$ (Figure 1). $\hat{\epsilon}^2$ and $\hat{\omega}^2$ are affected by the sample size less than $\hat{\eta}^2$ (Figure 1 and Figure 2).

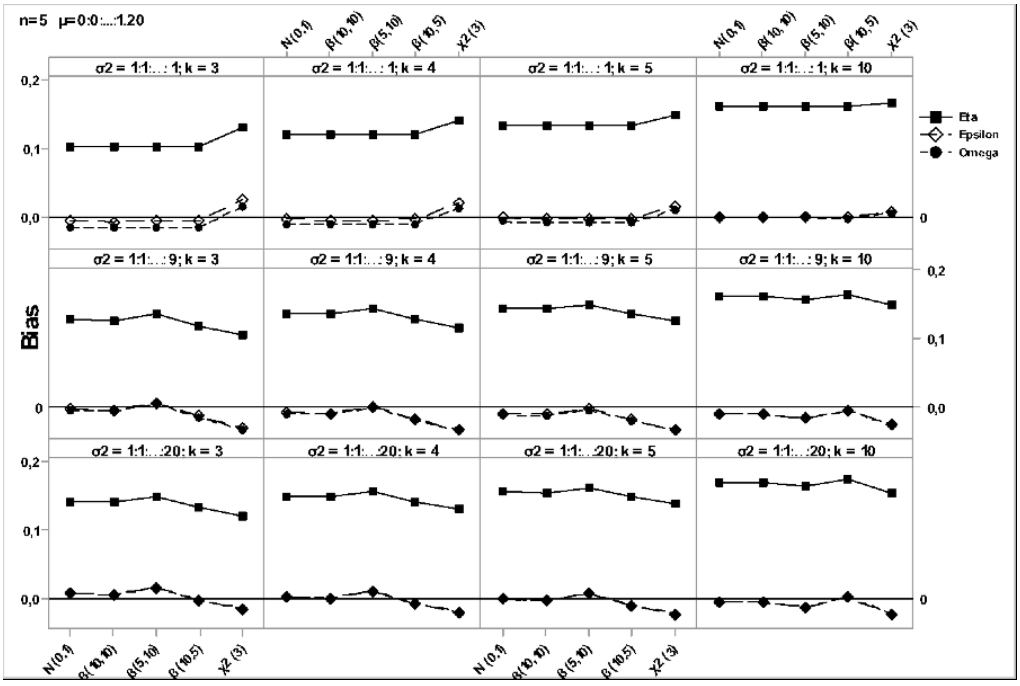

Figure 1: Bias for the One-Way ANOVA models when n=5 and μ=0:0:...:1.20.

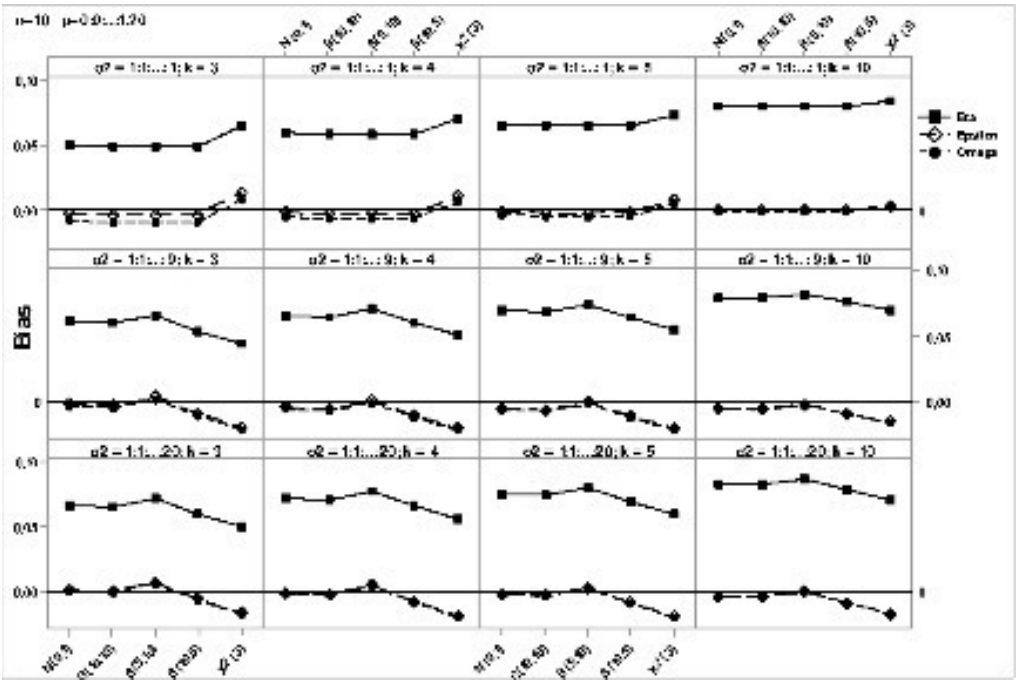

Figure 2: Bias for the One-Way ANOVA models when n=10 and μ=0:0:...:1.20.

3.2. Results of The Two-Way Analysis of Variance

With small size (n≤10) sub-groups that are taken from normal distribution and with homogeneous variances, estimations of $\hat{\eta}_p^2$ and $\hat{\eta}^2$ show excessive bias (Figure 4). However, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give significantly unbiased results. When variances are homogeneous, having slight and moderate deviance from normality does not affect the estimations of the four effect size measures. Besides, $\hat{\epsilon}^2$ gives the most unbiased results. If there is excessive skewness and kurtosis [$\chi^2(3)$], $\hat{\omega}^2$ gives the most unbiased results. However, in both cases, the difference between them is negligible. When variances are heterogeneous, regardless of the distribution of the populations they are taken from, $\hat{\eta}_p^2$ gives the most biased results under all of the considered experimental conditions, and $\hat{\eta}^2$ follows it. Additionally, as the variances get heterogeneous, $\hat{\eta}_p^2$ and $\hat{\eta}^2$ approach each other. When variances are heterogeneous, in cases where the population sub-groups are taken from are N(0,1), $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give the most unbiased results. When variances are heterogeneous, slight deviance from normality [$\beta(10, 10)$] does not affect the performances of $\hat{\epsilon}^2$ and $\hat{\omega}^2$. However, increase of the deviance of the distribution from normality increased the bias of these two effect size measures as well. This situation is seen significantly when variances are excessively heterogeneous. When variances are generally homogeneous, regardless of the experimental conditions, in cases where

the number of studied sub-groups increase, estimations of $\hat{\eta}_p^2$ and $\hat{\eta}^2$ gradually diverge from η^2, whereas estimations of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ gradually approach to η^2. Furthermore, as the number of sub-group increase, the difference between $\hat{\epsilon}^2$ and $\hat{\omega}^2$ gradually decreased. Considering the sub-groups with heterogeneous variances, regardless of the distribution shape, as the number of sub-groups increase, estimations of $\hat{\eta}_p^2$ and $\hat{\eta}^2$ gradually diverge from the population effect size. On the other hand, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ are not affected significantly from the increase in the sub-group number when variances are heterogeneous. Regardless of the experiment design (rxc), variance ratios, distribution shapes and sample size, the most biased estimations are made by $\hat{\eta}_p^2$, and $\hat{\eta}^2$ follows as a similar pattern (Figure 3 and 4).

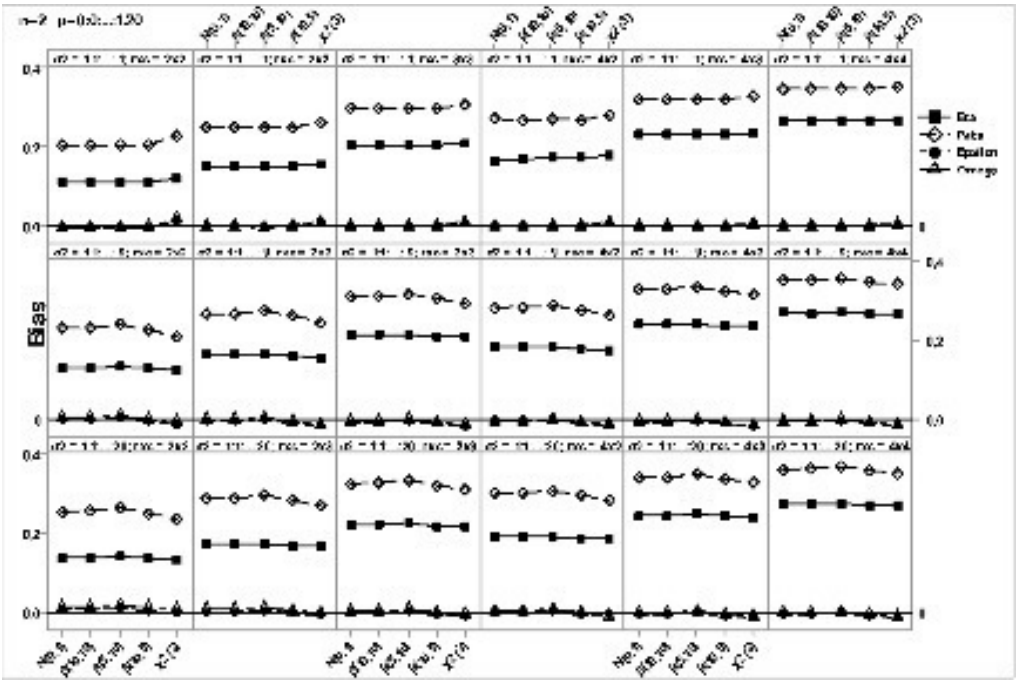

Figure 3: Bias for the Two-Way ANOVA models when n=2 and μ=0:0:...:1.20.

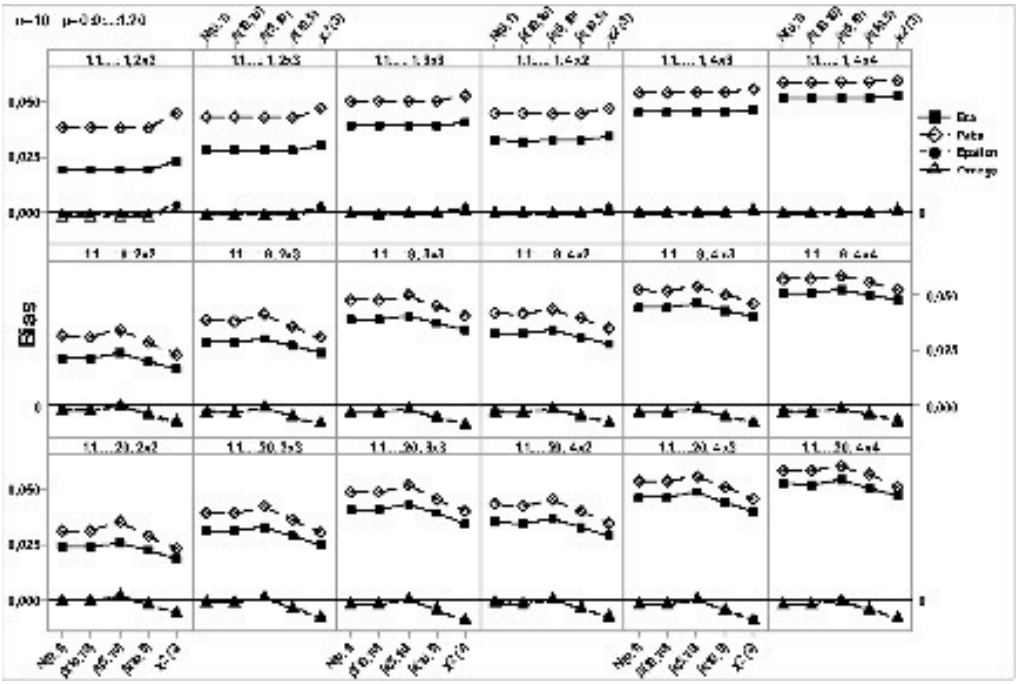

Figure 4: Bias for the Two-Way ANOVA models when n=10 and μ=0:0:...:1.20.

4. DISCUSSION

The Analysis of variance technique used most commonly in practice gives information about statistical significance only. It does not give information about practical significance of the factors and explained variance. Thus, when reporting analysis of variance results, reporting only P-values showing statistical significance will not be sufficient. Along with the P-values, some effect size measures such as $\hat{\eta}^2$, $\hat{\eta}_p^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ that show the practical significance and the share of the difference observed in the dependent variable explained by the considered factors should be reported. Thereby, understanding and interpreting the reported results in detail will be possible. There are many effect size measures developed for the purpose. However, it is an important shortcoming that performances of these effect size measures under many experimental conditions have not been shown in detail. Nevertheless, having detailed information about the performances of the effect size measures will provide insights to the researchers about which effect size measure they should report as a result of their studies. In the study conducted with this point of view, performances of $\hat{\eta}^2$, $\hat{\eta}_p^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$, which are found as the most popular effect size measures, are compared. Baguley [21] reported that

simple or unstandardized effect size measures are easier to compute and more robust than standardized effect size measures. Therefore, he has proposed to report simple effect size measures. However, in practice, standardized effect size measures have been commonly reported. For example, commonly used statistical package programs such as IBM SPSS, Minitab, Statistica and SAS report standardized effect size estimates along with P-values. We think that it is very easy to understand and interpret the effect size values for many authors and readers. That is why, in the simulation study conducted to compare performances of $\hat{\eta}^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ effect size measures for one-way analysis of variance. Keselman [17] reported that $\hat{\eta}^2$ gives similar results as $\hat{\epsilon}^2$and $\hat{\omega}^2$ in case of small population effect size. On the other hand, considering the standard deviations of estimations, he reported that $\hat{\eta}^2$ is a better estimator compared to $\hat{\epsilon}^2$ and $\hat{\omega}^2$. Nonetheless, in all other studies, it is reported that using $\hat{\eta}^2$ in estimating population effect size gives quite biased results [18, 8, 9]. In the results of our study too, it is seen that $\hat{\eta}^2$ gives quite biased results in all considered experimental conditions. In his simulation study, Keselman [17] stated that as long as the assumption of homogeneity of variances is met, selecting samples from populations with high skewness ($\gamma_1 = 2$) and kurtosis ($\gamma_2 = 6$) does not affect the performances of $\hat{\eta}^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ significantly. In their simulation study, Skidmore and Thompson [8] reported that even if the variances are heterogeneous, slight ($\gamma_1 = 0.5$ and $\gamma_2 = 0.5$ and moderate ($\gamma_1 = 1$ and $\gamma_2 = 3.75$)) level deviation from normality does not affect the performances of effect size measures ($\hat{\eta}^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$) significantly. As a result of our study too, it is seen that when variances are homogeneous as long as there is not excessive (χ^2) deviation from normality, shape of the distribution does not affect performances of the effect size measures significantly. However, in case variances are heterogeneous, it is seen that moderate [$\beta(5, 10)$ and $\beta(10, 5)$] and excessive (χ^2) deviations from normality affect effect size measures. Keselman [17] reported that while $\hat{\omega}^2$ could decisively provide estimations quite close to population effect size, $\hat{\epsilon}^2$ always produces estimations a little higher than that. However, Keselman did not report the number of replication in his study. In his simulation study aimed to compare some effect size measures, Okada [9] repeated Keselman's study with larger simulation number (1 million), at normal distribution and with different observation number combinations and stated that in all considered experimental conditions $\hat{\epsilon}^2$ gives more unbiased results than $\hat{\omega}^2$. In one of their studies, Glass and Hakstian [3] theoretically discussed whether $\hat{\epsilon}^2$ or $\hat{\omega}^2$ is unbiased and expressed that no matter how different their formulas are, both of them give similar results in practice. In the results of our study, in one-way variance analysis, as long as there is not excessive deviance from normality, it is seen that generally $\hat{\epsilon}^2$ gives the most unbiased results, and $\hat{\omega}^2$ follows it. However, in case of excessive deviances (χ^2) from normality, the most unbiased results are obtained by $\hat{\omega}^2$. On the other hand, in both situations, the difference between $\hat{\epsilon}^2$and $\hat{\omega}^2$ is negligible and confirms Glass and Hakstian [3]. As a result of the conducted simulation study, if the observation numbers in groups are equal and distributions are not excessively skewness and kurtosis, it is seen that heteroge-

neous variances do not affect $\hat{\epsilon}^2$ and $\hat{\omega}^2$ almost at all. Carrol and Nordholm [18] reported similar results. However, both Carrol and Nordholm [18] and Skidmore and Thompson [8] reported that heterogeneous variances are especially effective at unequal sample sizes (direct and inverse pairing). As long as the variances are homogeneous, regardless of the considered experimental conditions, it is seen that as the compared group numbers increase, deviances of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ in estimations approach to zero. Thus, in these experimental conditions, an increase in group numbers positively affect $\hat{\epsilon}^2$ and $\hat{\omega}^2$. However, in case of heterogeneous variances, making assessments on whether the increase in group numbers have positive or negative effects on estimations could be misleading. On the other hand, regardless of if the assumption of homogeneity of variances is met or not, is significantly affected by an increase in group numbers. Results obtained under these conditions overlap with the findings of Skidmore and Thompson [8]. It is reported that as the population effect size decreases, biasness decreases too [17]. However, as the population effect size decreases, while $\hat{\eta}^2$ gives more biased results, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ gives more unbiased results. When estimating effect sizes related to interaction effects in factorial experiments, performance of $\hat{\eta}_p^2$ is investigated in addition to $\hat{\eta}^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$. Whatever the experimental conditions considered in the study, as the sample size increases, estimations of the four effect sizes approach gradually to the population effect size. However, in all of the considered experimental conditions, it is seen that while $\hat{\eta}_p^2$ gives the most biased results, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give the most unbiased results. Furthermore, while $\hat{\eta}_p^2$ and $\hat{\eta}^2$ are negatively affected by the increase in sub-group number, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ are not negatively affected. In the meantime, it is remarkable that effects of both shape of distribution and variance rates on the considered effect size measures are generally similar to the ones in one-way variance analyses. On the other hand, our study has revealed the performances of the considered effect size measures in factorial ANOVA models. Thus, the study has fulfilled an important need in this field because factorial ANOVA design is commonly used in practice.

5. CONCLUSION AND RECOMMENDATIONS

Reporting statistical analysis results in an understandable and informative way is very important. Therefore, when reporting statistical analysis results, along with the P-value that shows statistical significance some effect size measures should be reported. While a statistically significant difference is not necessarily practically significant, a statistically non-significant difference is not necessarily practically non-significant. Notwithstanding, majority of researchers believe that the smaller the P-value is that shows the statistical significance, the larger and the more important the difference between the groups that are compared. However, the P-value does not provide any information about practical significance. Thus, in the results of the studies, along with the statistical significance (P-value), effect

size measures that provide information about the practical significance should necessarily be reported. However, it is remarkable that majority of the researchers who report effect size report $\hat{\eta}^2$ ($\mathbb{R}^2$) and $\hat{\eta}_p^2$ [22]. This is because commonly used statistics package programs such as Minitab, IBM SPSS, NCSS, Statistica etc. directly report $\hat{\eta}^2$ ($\mathbb{R}^2$) or $\hat{\eta}_p^2$ while reporting analysis of variance results. However, the most noteworthy thing is that reported effect size measures should represent population effect size as accurately as possible (unbiased). From this point forth, performances of the most commonly known effect size measures in practice are compared under many experimental conditions in one-way and two-way analysis of variance models. In the light of the acquired findings, concluding with following results is possible:

1. In both one factor and two factor experimental conditions, $\hat{\eta}^2$ gives quite biased results. Thus, since using $\hat{\eta}^2$ to estimate population effect size at the end of analysis of variance is quite misleading, reporting $\hat{\eta}^2$ should not be recommended.

2. Although $\hat{\eta}_p^2$ is used in experimental conditions considering more than one factor as an alternative to $\hat{\eta}^2$, it is seen that $\hat{\eta}_p^2$ gives more biased results than $\hat{\eta}^2$ in two factor experiments after 1.000.000 simulation experiments. Additionally, since $\hat{\eta}_p^2$ takes every effect separately in consideration ($SS_{Effect} + SS_{Error}$), total variation explained by the model could surpass 1 (100%) [23, 24]. This is a common situation in practice [25]. Since $\hat{\eta}_p^2$ estimates of effect size are biased, reporting it should not be recommended.

3. Although Okada [9] reported that relationships among Eta, Omega and Epsilon-squared is $\hat{\omega}^2 \leq \hat{\epsilon}^2 \leq \hat{\eta}^2$, this relation is not valid for every experimental condition. For example, the relationship between Epsilon and Omega squared is $\hat{\epsilon}^2 \leq \hat{\omega}^2$ when negative estimations are obtained regardless of experimental conditions.

4. Although it is seen that in some of the experimental conditions $\hat{\epsilon}^2$ and in some of the others $\hat{\omega}^2$ gives more unbiased results, the difference between these two measures is at a negligible level. It is seen that both $\hat{\epsilon}^2$ and $\hat{\omega}^2$ estimates population effect size in a quite unbiased fashion in all experimental conditions. Thus, it could be concluded that when estimating effect size in analysis of variance models and accordingly analyzing practical significance of the observed difference, using $\hat{\epsilon}^2$ or $\hat{\omega}^2$ is much truer and one of these measures should be reported.

5. Obtaining negative estimates in some experimental conditions (i.e. small effect size magnitude) may be considered a disadvantage of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ estimates, although both measures give unbiased estimates almost all experimental conditions.

6. It is determined that it is a very important deficiency that $\hat{\epsilon}^2$ and $\hat{\omega}^2$ are not included in almost none of the commonly used statistics package programs although $\hat{\eta}^2$ and $\hat{\eta}_p^2$ has been reported to be quite biased in studies for 50 years. Thus, at least one of these two measures should be included in the libraries of commonly used package programs such as Minitab, SPSS etc.

REFERENCES

[1] Zar, H.J. (2010). *Biostatistical Analysis*, Pearson, New Jersey.

[2] Mendes, M. and Yigit, S. (2013). Comparison of ANOVA-F and ANOM tests with regard to type I error rate and test power, *Journal of Statistical Computation and Simulation*, **83**(11), 2093–2104.

[3] Glass, G.V. and Hakstian, A.R. (1969). Measures of Association in Comparative Experiments: Their Development and Interpretation, *American Educational Research Journal*, **6**(3), 403–414.

[4] Nickerson, R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy, *Psychological Methods*, **5**(2), 241–301.

[5] Fan, X. (2001). Statistical Significance and Effect Size in Education Research: Two Sides of a Coin, *The Journal of Educational Research*, **94**(5), 275–282.

[6] Mendes, M. (2012). *Uygulamali Bilimler Icin Istatistik Ve Arastirma Yontemleri*, Kriter, Canakkale.

[7] Hays, W. (1963). *Statistics for Psychologists*, Rinehart Winston, New York.

[8] Skidmore, S.T. and Thompson, B. (2012). Bias and precision of some classical ANOVA effect sizes when assumptions are violated, *Behavior Research Methods*, **45**(2), 536–546.

[9] Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA, *Behaviormetrika*, **40**(2), 129–147.

[10] Cumming, G. and Finch, S. (2001). A Primer on the Understanding, Use, and Calculation of Confidence Intervals that are Based on Central and Noncentral Distributions, *Educational and Psychological Measurement*, **61**(4), 532–574.

[11] Kelley, T.L. (1935). An Unbiased Correlation Ratio Measure, *Proceedings of the National Academy of Sciences of the United States of America*, **21**(9), 554–559.

[12] Maxwell, S.E.; Camp, C.J. and Arvey, R.D. (1981). Measures of strength of association: A comparative examination, *Journal of Applied Psychology*, **6**(5), 525–534.

[13] Keppel, G. (1982). *Design and Analysis: A Researcher's Handbook*, Prentice Hall, New Jersey.

[14] Olejnik, S. and Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations, *Contemporary Educational Psychology*, **25**(3), 241–286.

[15] Kirk, R.E. (2003). *The importance of effect magnitude*. In "Handbook of Research Methods in Experimental Psychology" (S.F. Davis, Ed.), Oxford, Blackwell, 83–105.

[16] Grissom, R. and Kim, J. (2005). *Effect Sizes for Research: A Broad Practical Approach*, Lawrence Erlbaum, New Jersey.

[17] Keselman, H.J. (1975). A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared, *Canadian Psychological Review/Psychologie Canadienne*, **16**(1), 44–48.

[18] Carroll, R.M. and Nordholm, L.A. (1975). Sampling Characteristics of Kelley's epsilon and Hays' omega, *Educational and Psychological Measurement*, **35**(3), 541–554.

[19] Levine, T.R. and Hullett, C.R. (2002). Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research, *Human Communication Research*, **28**(4), 612–615.

[20] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum, New Jersey.

[21] Baguley, T. (2009). Standardized or simple effect size: What should be reported?, *British Journal of Psychology*, **100**(3), 603–617.

[22] Fritz, C.O.; Morris, P.E. and Richler, J.J. (2012). Effect size estimates: current use, calculations, and interpretation, *Journal of Experimental Psychology: General*, **141**(1), 2–18.

[23] Cohen, J. (1973). Eta-Squared and Partial Eta-Squared in Fixed Factor Anova Designs, *Educational and Psychological Measurement*, **33**(1), 107–112.

[24] Haase, R.F. (1983). Classical and Partial Eta Square in Multifactor ANOVA Designs, *Educational and Psychological Measurement*, **43**(1), 35–39.

[25] Pierce, C.A.; Block, R.A. and Aguinis, H. (2004). Cautionary Note on Reporting Eta-Squared Values from Multifactor ANOVA Designs, *Educational and Psychological Measurement*, **64**(6), 916–924.

REVSTAT – Statistical Journal
Volume 16, Number 3, July 2018, 315–322

ON THE DISTRIBUTION OF A QUADRATIC FORM IN NORMAL VARIATES

Author: Jin Zhang
– School of Mathematics and Statistics, Yunnan University, Kunming, Yunnan, 650091, China
jinzhang@ynu.edu.cn

Received: August 2016 Revised: September 2016 Accepted: September 2016

Abstract:

- It is a well-known theorem in linear models that the idempotency of a matrix is a sufficient and necessary condition for a quadratic form in normal variates to have a chi-square distribution, but its proofs in the early literature were incorrect or incomplete. Driscoll (1999) provided an improved proof, and this article presents a simple proof. More importantly, we establish and prove a generalized theorem.

Key-Words:

- *eigenvalues; idempotent matrices; moment-generating function; normality.*

AMS Subject Classification:

- 62J05, 62H10, 62E15.

1. INTRODUCTION

There is a rich literature on the distribution and independence of quadratic forms in normal random vectors (e.g, Rao, 1973; Graybill, 1976; Driscoll and Gundberg, 1986; Mathai and Provost, 1992; Jorgensen, 1993; Driscoll, 1999; Christensen, 2002; Ravishanker and Dey, 2002; Ogawa and Olkin, 2008), which play an important role in linear models and multivariate statistical analysis.

Let $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the k-dimensional normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, and let $\chi^2_m(\lambda)$ be the noncentral chi-square distribution with m degrees of freedom and noncentrality parameter λ. The two well-known theorems below establish sufficient and necessary conditions for the independence and distributions of quadratic forms in normal variates.

Theorem 1. *Let $\boldsymbol{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$, and $\boldsymbol{A}$ and $\boldsymbol{B}$ be $k \times k$ real symmetric matrices. Then $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}$ are independently distributed if and only if $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \mathbf{0}$.*

Theorem 2. *Let $\boldsymbol{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > \mathbf{0}$, and $\boldsymbol{A}$ be a $k \times k$ real symmetric matrix. Then $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \sim \chi^2_m(\lambda)$ with $\lambda = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$ if and only if $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent of rank m.*

Unfortunately, the proofs of the two theorems in the early literature are incorrect, incomplete or misleading, especially for Theorem 1 (Driscoll and Gundberg, 1986; Driscoll, 1999; Ogawa and Olkin, 2008). Thus, many improved proofs for Theorem 1 have been obtained by Reid and Driscoll (1988), Driscoll and Krasnicka (1995), Letac and Massam (1995), Provost (1996), Olkin (1997), Marcus (1998), Li (2000), Matsuura (2003), Ogawa and Olkin (2008), Carrieu and Lassère (2009), Carrieu (2010), Bonnefond (2012), Zhang and Yi (2012), and many others. However, there is only one improved proof of Theorem 2 given by Driscoll (1999). In addition, Liu *et al.* (2009) and Duchesne and Lafaye De Micheaux (2010) discussed the computational issues in Theorem 2.

A simple proof of Theorem 2 is presented in Section 2, using elementary calculus and matrix algebra. We give a counter example of Theorem 2 in Section 3, where $\boldsymbol{\Sigma}$ is singular. Then we establish and prove its extension in Theorem 3 for the general case, where $\boldsymbol{\Sigma}$ can be singular or nonsingular.

2. A SIMPLE PROOF OF THEOREM 2

The proof of sufficiency for Theorem 2 is quite easy, but showing necessity is difficult. In fact, its proofs in the early literature were incorrect or incomplete, according to Driscoll (1999), who provided an improved proof of Theorem 2, based on the moment-generating function and cumulants. We now present a simple proof of Theorem 2, using the moment-generating function of $\chi^2_m(\lambda)$:

$$M_{\chi^2_m(\lambda)}(t) = (1-2t)^{-\frac{m}{2}} e^{\frac{2t\lambda}{1-2t}}, \quad t < 1/2.$$

Proof of Theorem 2:

Sufficiency. Suppose $(\boldsymbol{A\Sigma})^2 = \boldsymbol{A\Sigma}$ and $r(\boldsymbol{A\Sigma}) = m$, where $\boldsymbol{\Sigma} = \boldsymbol{BB}'$ and $\tilde{\boldsymbol{A}} = \boldsymbol{B}'\boldsymbol{AB}$. Then $\tilde{\boldsymbol{A}}^2 = \tilde{\boldsymbol{A}}$ and $r(\tilde{\boldsymbol{A}}) = m$. Thus, there exists an orthogonal matrix $\boldsymbol{P}$ such that

$$\tilde{\boldsymbol{A}} = \boldsymbol{P}\begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\boldsymbol{P}' = \boldsymbol{P}_1\boldsymbol{P}_1',$$

where $\boldsymbol{P} = (\boldsymbol{P}_1, \boldsymbol{P}_2)$, $\boldsymbol{P}_1'\boldsymbol{P}_1 = \boldsymbol{I}_m$ and $\boldsymbol{z} = \boldsymbol{P}_1'\boldsymbol{B}^{-1}\boldsymbol{x} \sim N_m(\boldsymbol{P}_1'\boldsymbol{B}^{-1}\boldsymbol{\mu},\ \boldsymbol{I})$. It follows that

$$\boldsymbol{x}'\boldsymbol{Ax} = \boldsymbol{z}'\boldsymbol{z} \sim \chi^2_m(\lambda),$$

where $\lambda = \frac{1}{2}(\boldsymbol{P}_1'\boldsymbol{B}^{-1}\boldsymbol{\mu})'(\boldsymbol{P}_1'\boldsymbol{B}^{-1}\boldsymbol{\mu}) = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$.

Necessity. Suppose $\boldsymbol{x}'\boldsymbol{Ax} \sim \chi^2_m(\lambda)$. Let $\boldsymbol{P} = (\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k)$ be an orthogonal matrix such that $\boldsymbol{P}'\tilde{\boldsymbol{A}}\boldsymbol{P} = \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_k)$, where $\lambda_1 \geq \cdots \geq \lambda_k$ are eigenvalues of $\tilde{\boldsymbol{A}}$. Then

$$\boldsymbol{x}'\boldsymbol{Ax} = \boldsymbol{z}'\boldsymbol{\Lambda}\boldsymbol{z} = \sum_{i=1}^{k}\lambda_i z_i^2, \quad M_{\boldsymbol{x}'\boldsymbol{Ax}}(t) = \prod_{i=1}^{k} M_{z_i^2}(t\lambda_i),$$

where $\boldsymbol{z} = \boldsymbol{P}'\boldsymbol{B}^{-1}\boldsymbol{x} \sim N_k(\boldsymbol{P}'\boldsymbol{B}^{-1}\boldsymbol{\mu},\ \boldsymbol{I})$ and $z_1, ..., z_k$ are independent. Hence,

$$(1-2t)^{-\frac{m}{2}} e^{\frac{2t\lambda}{1-2t}} = \prod_{i=1}^{k}(1-2t\lambda_i)^{-\frac{1}{2}} e^{\frac{t\lambda_i}{1-2t\lambda_i}(\boldsymbol{p}_i'\boldsymbol{B}^{-1}\boldsymbol{\mu})^2}$$

for $t < 1/2$ and $t\lambda_i < 1/2$ $(i = 1, ..., k)$. Comparing the discontinuous points of the two functions on both sides results in

$$(1-2t)^{-\frac{m}{2}} = \prod_{i=1}^{k}(1-2t\lambda_i)^{-\frac{1}{2}} = |\boldsymbol{I}_k - 2t\tilde{\boldsymbol{A}}|^{-\frac{1}{2}},$$

which implies that $\lambda_1 = \cdots = \lambda_m = 1$ and $\lambda_{m+1} = \cdots = \lambda_k = 0$.

Thus, $\tilde{\boldsymbol{A}}$ or $\boldsymbol{A\Sigma}$ is idempotent of rank m. The proof is completed. □

3. DISTRIBUTIONS OF QUADRATIC FORMS IN THE GENERAL CASE

We now discuss the distribution of quadratic form $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ in the general case, where $\boldsymbol{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ but $\boldsymbol{\Sigma}$ can be singular or not. First, it should be pointed out that Theorem 2 is not true when $\boldsymbol{\Sigma}$ is singular. Below is a counter example.

Let $\boldsymbol{A} = \boldsymbol{I}_2$ and $\boldsymbol{x} = (z, 1)'$, where $z \sim N(0, 1)$. Then $\boldsymbol{x} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

It is clear that $(\boldsymbol{A}\boldsymbol{\Sigma})^2 = \boldsymbol{A}\boldsymbol{\Sigma}$ and its rank $r(\boldsymbol{A}\boldsymbol{\Sigma}) = 1$, but

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = z^2 + 1,$$

the distribution of which is not $\chi_1^2(\lambda)$ with $\lambda = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} = \frac{1}{2}$.

To generalize Theorem 2, we have the following Theorem 3, which reduces to Theorem 2 if $\boldsymbol{\Sigma}$ is nonsingular. The proof for Theorem 3 is based on the moment-generating function of quadratic function $Q = \boldsymbol{z}'\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}'\boldsymbol{z} + c$:

$$M_Q(t) = |\boldsymbol{I} - 2t\boldsymbol{A}|^{-\frac{1}{2}} e^{ct + \frac{t^2}{2}\boldsymbol{b}'(\boldsymbol{I} - 2t\boldsymbol{A})^{-1}\boldsymbol{b}}$$

for small $|t|$ such that $\boldsymbol{I} - 2t\boldsymbol{A} > \boldsymbol{0}$, where $\boldsymbol{z} \sim N_k(\boldsymbol{0},\ \boldsymbol{I})$, $\boldsymbol{A}$ is a real symmetric matrix, $\boldsymbol{b}$ is a k-dimensional real vector, and c is a real number. In fact,

$$\begin{aligned} M_Q(t) &= \int (2\pi)^{-\frac{k}{2}} e^{t(\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}'\boldsymbol{z} + c) - \frac{1}{2}\boldsymbol{z}'\boldsymbol{z}} d\boldsymbol{z} \\ &= e^{ct + \frac{t^2}{2}\boldsymbol{b}'\boldsymbol{A}_t\boldsymbol{b}} \int (2\pi)^{-\frac{k}{2}} e^{-\frac{1}{2}(\boldsymbol{z} - t\boldsymbol{A}_t\boldsymbol{b})'\boldsymbol{A}_t^{-1}(\boldsymbol{z} - t\boldsymbol{A}_t\boldsymbol{b})} d\boldsymbol{z}, \end{aligned}$$

where $\boldsymbol{A}_t = (\boldsymbol{I} - 2t\boldsymbol{A})^{-1}$.

Theorem 3. *Let $\boldsymbol{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{A}$ be a $k \times k$ real symmetric matrix. Then $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \sim \chi_m^2(\lambda)$ with $\lambda = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$ if and only if*

$$\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma},\ r(\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}) = m,\ \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu}.$$

Proof: Let $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{z} + \boldsymbol{\mu}$, where $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}'$, $\boldsymbol{z} \sim N_k(\boldsymbol{0}, \boldsymbol{I})$ and $\tilde{\boldsymbol{A}} = \boldsymbol{B}'\boldsymbol{A}\boldsymbol{B}$.

Sufficiency. Note that $\tilde{\boldsymbol{A}}^2 = \tilde{\boldsymbol{A}}$ and $r(\tilde{\boldsymbol{A}}) = m$ due to $\boldsymbol{B} = \boldsymbol{B}\boldsymbol{B}'(\boldsymbol{B}\boldsymbol{B}')^{-}\boldsymbol{B}$ and

$$r(\tilde{\boldsymbol{A}}) \geq r(\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}) = r((\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{B})(\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{B})') = r(\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{B}) \geq r(\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}) = r(\tilde{\boldsymbol{A}}).$$

Then $||(\boldsymbol{I}-\tilde{\boldsymbol{A}})\boldsymbol{B}'\boldsymbol{A}\boldsymbol{\mu})||^2 = \boldsymbol{\mu}'\boldsymbol{A}(\boldsymbol{\Sigma}-\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma})\boldsymbol{A}\boldsymbol{\mu} = 0$, so that $\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{B} = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{B}\tilde{\boldsymbol{A}}$ and

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{z}'\tilde{\boldsymbol{A}}\boldsymbol{z} + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} + 2\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{z} = (\boldsymbol{z}+\boldsymbol{c})'\tilde{\boldsymbol{A}}(\boldsymbol{z}+\boldsymbol{c}) \sim \chi^2_m(\lambda),$$

where $\boldsymbol{c}' = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{B}$ and $\lambda = \frac{1}{2}\boldsymbol{c}'\tilde{\boldsymbol{A}}\boldsymbol{c} = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$.

Necessity. Suppose $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{z}'\tilde{\boldsymbol{A}}\boldsymbol{z} + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} + 2\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{z} \sim \chi^2_m(\lambda)$. Then

$$(1-2t)^{-\frac{m}{2}} e^{\frac{2t\lambda}{1-2t}} = |\boldsymbol{I}_k - 2t\tilde{\boldsymbol{A}}|^{-\frac{1}{2}} e^{t\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}+2t^2\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{B}(\boldsymbol{I}-2t\tilde{\boldsymbol{A}})^{-1}\boldsymbol{B}'\boldsymbol{A}\boldsymbol{\mu}}$$

for small $|t|$. Comparing the discontinuous points of the two functions on both sides gives

$$(1-2t)^{-\frac{m}{2}} = |\boldsymbol{I}_k - 2t\tilde{\boldsymbol{A}}|^{-\frac{1}{2}},$$

which implies that $\tilde{\boldsymbol{A}}^2 = \tilde{\boldsymbol{A}}$ and $r(\tilde{\boldsymbol{A}}) = m$ (see Section 2), or equivalently

$$\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma} \text{ and } r(\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}) = m.$$

It follows from above two equations that $\frac{2t\lambda}{1-2t} = t\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} + 2t^2\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{B}(\boldsymbol{I}-2t\tilde{\boldsymbol{A}})^{-1}\boldsymbol{B}'\boldsymbol{A}\boldsymbol{\mu}$, so

$$\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} = 2\lambda = (1-2t)[\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} + 2t\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{B}(\boldsymbol{I}-2t\tilde{\boldsymbol{A}})^{-1}\boldsymbol{B}'\boldsymbol{A}\boldsymbol{\mu}],$$

which and $(\boldsymbol{I}-2t\tilde{\boldsymbol{A}})^{-1} = \sum_{n=0}^{\infty}(2t\tilde{\boldsymbol{A}})^n$ imply that for small $|t|$,

$$\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} + 2t\boldsymbol{\mu}'(\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{A}-\boldsymbol{A})\boldsymbol{\mu} + 4t^2\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{B}(\tilde{\boldsymbol{A}}-\boldsymbol{I})\boldsymbol{B}'\boldsymbol{A}\boldsymbol{\mu} + \cdots$$

By the theory of power series, $\boldsymbol{\mu}'(\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{A}-\boldsymbol{A})\boldsymbol{\mu} = 0 = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{B}(\tilde{\boldsymbol{A}}-\boldsymbol{I})\boldsymbol{B}'\boldsymbol{A}\boldsymbol{\mu}$. That is,

$$\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu}.$$

The proof is completed. □

ACKNOWLEDGMENTS

This research is supported by the Natural Science Foundation of China (NSFC, Grant No. 11561073, 11461079).

REFERENCES

[1] Bonnefond, X. (2012). An analytical proof of Ogawa's determinantal theorem, *Aequationes Mathematicae*, **83**, 127–130.

[2] Carrieu, H. (2010). Close to the Craig-Sakamoto theorem, *Linear Algebra Appl.*, **432**, 777–779.

[3] Carrieu, H. and Lassère, P. (2009). One more simple proof of the Craig-Sakamoto theorem, *Linear Algebra Appl.*, **431**, 1616–1619.

[4] Christensen, R. (2002). *Plane Answers to Complex Questions: The Theory of Linear Models*, 3rd ed., Springer, New York.

[5] Driscoll, M.F. (1999). An improved result relating quadratic forms and chi-square distributions, *American Statistician*, **53**, 273–275.

[6] Driscoll, M.F. and Gundberg Jr., W.R. (1986). A history of the development of Craig's theorem, *American Statistician*, **42**, 139–142.

[7] Driscoll, M.F. and Krasnicka, B. (1995). An accessible proof of Craig's theorem in the general case, *American Statistician*, **49**, 59–62.

[8] Duchesne, P. and Lafaye De Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods, *Computational Statistics and Data Analysis*, **54**, 858–862.

[9] Graybill, F.A. (1976). *Theory and Application of the Linear Model*, North Scituate, MA: Duxbury Press.

[10] Jorgensen, B. (1993). *The Theory of Linear Models*, New York: Chapman & Hall.

[11] Letac, G. and Massam, H. (1995). Craig-Sakamoto's theorem for the Wishart distributions on symmetric cones, *Ann. Inst. Statist. Math.*, **47**, 785–799.

[12] Li, C.K. (2000). A simple proof of the Craig-Sakamoto theorem, *Linear Algebra Appl.*, **321**, 281–283.

[13] Liu, H.; Tang, Y. and Zhang, H.H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables, *Computational Statistics and Data Analysis*, **53**, 853–856.

[14] Marcus, M. (1998). On a determinant result of I. Olkin, *Linear Algebra Appl.*, **277**, 237–238.

[15] Mathai, M.A. and Provost, S.B. (1992). *Quadratic Forms in Random Variables, Theory and Applications*, Marcell Dekker, New York.

[16] Matsuura, M. (2003). On the Craig-Sakamoto theorem and Olkin's determinantal result, *Linear Algebra Appl.*, **364**, 321–323.

[17] Ogawa, J. and Olkin, I. (2008). A tale of two countries: The Craig-Sakamoto-Matusita theorem, *J. Statist. Plann. Infer.*, **138**, 3419–3428.

[18] Olkin, I. (1997). A determinantal proof of the Craig-Sakamoto theorem, *Linear Algebra Appl.*, **264**, 217–223.

[19] Provost, S.B. (1996). On Craig's theorem and its generalizations, *J. Statist. Plann. Infer.*, **53**, 311–321.

[20] Ravishanker, N. and Dey, D.K. (2002). *A First Course in Linear Model Theory*, Chapman & Hall/CRC, Boca Raton.

[21] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications* (2nd ed.), New York: Wiley.

[22] Reid, J.G. and Driscoll, M.F. (1988). An accessible proof of Carig's theorem in the non-central case, *American Statistician*, **42**, 139–142.

[23] Searle, S.R. (1972). *Linear Models*, New York: John Wiley & Sons.

[24] Zhang, J. and Yi, J. (2012). A simple proof of the generalized Craig-Sakamoto theorem, *Linear Algebra Appl.*, **437**, 781–782.

REVSTAT – Statistical Journal
Volume 16, Number 3, July 2018, 323–348

MODEL-ASSISTED AND MODEL-CALIBRATED ESTIMATION FOR CLASS FREQUENCIES WITH ORDINAL OUTCOMES

Authors: Maria del Mar Rueda
– Department of Statistics and Operational Research, University of Granada, Spain
mrueda@ugr.es

Antonio Arcos
– Department of Statistics and Operational Research, University of Granada, Spain
arcos@ugr.es

David Molina
– Department of Statistics and Operational Research, University of Granada, Spain
dmolinam@ugr.es

Manuel Trujillo
– Institute of Social Studies of Andalusia (IESA), Spanish National Research Council (CSIC), Spain
mtrujillo@iesa.csic.es

Received: February 2016 Revised: September 2016 Accepted: September 2016

Abstract:

- This paper considers new techniques for complex surveys in the case of estimation of proportions when the variable of interest has ordinal outcomes. Ordinal model-assisted and ordinal model-calibrated estimators are introduced for class frequencies in a population, taking two different approaches. Theoretical properties and numerical methods are investigated. Simulation studies using data from a real macro survey are considered to evaluate the performance of the proposed estimators. The empirical coverage and the length of confidence intervals are computed using several techniques in variance estimation. We also use data from an opinion survey to show the behavior of the proposed estimators in real applications.

Key-Words:

- *complex surveys; model calibration; ordinal data; weighted least squares; weighted maximum likelihood.*

AMS Subject Classification:

- 62D05.

1. INTRODUCTION

Questions with categorical outcomes are quite common in surveys, especially in health, marketing, public opinion and official surveys. In the simplest case, questions have only two possible responses, which are often used to represent the "success" or the "failure" of an experiment (such as the occurrence or nonoccurrence of an event or the presence or absence of a characteristic). These items can be modelled statistically using binary logit regression ([1]; [25]).

In more complex situations, items have three or more possible options and respondents must select one of them. When analyzing a polytomous variable it is necessary to determine whether its categories can be ordered according to an intrinsic characteristic of the categories themselves. If so, the number of outcomes attributable to each category is modelled by an ordinal distribution, which gives rise to an ordinal logit model. Otherwise, we should use a multinomial logit model, which is based on the multinomial probability distribution.

Most studies related to binary, multinomial or ordinal logit regression are based on the assumption of a simple random sample drawn from a large population. However, this scenario is not always present in practice: many surveys assume a finite population with samples extracted from complex sampling designs. For example, the Educational Longitudinal Study developed by the National Center for Education Statistics, the Post Enumeration Survey conducted by the Portuguese Statistical Office ([6]) and the Programme for International Student Assessment (PISA) study conducted by the Organisation for Economic Co-operation and Development, all applied complex sampling survey designs. These designs have in common the use of strata, clusters and unequal probabilities of selection in data collection. In this respect, it has been shown that ignoring weights, clusters and strata can lead to biased parameter estimates and erroneous standard errors in ordinal logistic regression analysis [24].

[23] used binary and multinomial logistic regressions in the context of survey sampling. In this context, [38] used a logistic regression model to obtain a calibration estimator for the finite population distribution function under a general sampling design, while [22] developed point and variance estimators for the total of finite population characteristics from a clustered sample assisted by a logistic regression model.

Ordinal regression models have been used extensively in sociological, medical and educational research but have a very sparse presence in parameter estimation in finite population sampling, which motivated this work. Therefore, the objective of this paper is to introduce new ordinal model-assisted estimators and ordinal model-calibrated estimators for the proportions of the categories of a response variable with ordinal outcomes.

This article proceeds as follows: Section 2 reviews the estimation methods that have been suggested to determine the proportion of categories of an ordinal response variable in finite population sampling. In section 3, the use of ordinal regression models in survey sampling is introduced. In section 4 we propose several estimators for the proportions of categories: the first of these is based on the procedure used by [23] for the case of a nominal variable, and the second is defined using calibration techniques ([8]; [31]). A brief discussion is then offered of numerical methods for parameter estimation. The main theoretical properties of the proposed estimators are studied in section 5. Variance estimation is addressed in section 6. Section 7 describes how the performance of the proposed estimators is measured through simulation experiments. Section 8 presents the results obtained from the different estimation strategies with respect to an opinion survey dataset. Finally, section 9 summarizes the conclusions drawn.

2. ESTIMATORS FOR CLASS RELATIVE FREQUENCIES OF A DISCRETE RESPONSE VARIABLE UNDER A GENERAL SAMPLING DESIGN

Let U denote a finite population with N units, $U = \{1, ..., k, ..., N\}$. Assume that data are collected from respondents who provide a single choice from a list of alternatives coded $1, 2, ..., i, ..., m$. Consider a discrete m-valued survey variable Y and denote the value observed for the kth individual of the population as y_k. Our aim is to estimate the frequency distribution of Y in the population U. To do so, we define a class of indicators z_i $(i = 1, ..., m)$ such that for each unit $k \in U$ $z_{ki} = 1$ if $y_k = i$ and $z_{ki} = 0$ otherwise. The problem thus, is to estimate the proportions $P_i = 1/N \sum_{k \in U} z_{ki}$, $i = 1, 2, ..., m$.

Let s be a probability sample of size n drawn from population U using a sampling design p_d. The sampling design considered induces first-order inclusion probabilities π_k, second-order inclusion probabilities π_{kl} and design weights $d_k = 1/\pi_k$, for $k, l = 1, ..., N$.

The customary design unbiased estimator of P_i is given by

$$\widehat{P}_{\mathrm{HT}i} = \frac{1}{N} \sum_{k \in s} \frac{z_{ki}}{\pi_k} = \frac{1}{N} \sum_{k \in s} z_{ki} d_k, \tag{2.1}$$

where the subindex HT refers to the Horvitz–Thompson estimator [20]. The design weights d_k are commonly thought of as the number of population units represented by unit k in the sample. [10] discussed the estimation of proportions using Bernoulli sampling and stratified designs.

In sample surveys, the use of auxiliary variables has been widely discussed by survey practitioners since this approach can increase the efficiency of the es-

timates in different contexts (see e.g. [9]). Thus, it is common practice to use auxiliary information on a character x related to the main variable y. A variety of approaches are available to construct more efficient estimators including design-based and model-based methods (see e.g. [35]; [32]).

Let us now consider a general situation where the auxiliary variable can be either numeric or binary. Let x_k be the value of the study variable x for the kth population element, available for all of U. For the sample s, the values of the two variables (y_k, x_k), $k \in s$, are observed. Under this scenario, we can consider the use of superpopulation models for sampling surveys. A superpopulation model is a way of formalising the relationship between a target variable and the auxiliary data. In previous research, superpopulation models have been used in sociological and electoral studies. For example, [5] used the superpopulation approach to estimate average customer satisfaction and [29] used superpopulation models to analyze electoral polls. Traditionally, linear regression models have been used to incorporate auxiliary information but (as is well known in sociological literature, see e.g. [36]) for qualitative variables a linear model might be unrealistic.

A first procedure is to consider the superpopulation multinomial logistic model given in [23]: we assume that the population under study $\boldsymbol{y} = (y_1, ..., y_N)^\top$ constitutes a body of superpopulation random variables $\boldsymbol{Y} = (Y_1, ..., Y_N)^\top$, containing a superpopulation model, ξ, such that

$$\mu_i(x_k) = P(Y_k = i|x_k) = \mathrm{E}_\xi(Z_{ki}|x_k) = \frac{\exp(\alpha_i + \beta_i x_k)}{\sum_{j=1}^{m} \exp(\alpha_j + \beta_j x_k)},$$

$i = 1, ..., m$, $k = 1, ..., N$ (E_ξ denotes the expected value with respect to the model) and assume that Y_k are conditionally independent given x_k.

Usually, population parameters α_i and β_i involved in the model ξ are unknown and should be estimated from the sample. Considering $\hat{\alpha}_i$ and $\hat{\beta}_i$ as the maximum likelihood estimations of α_i and β_i, we can define an estimator for probabilities for each category as follows:

$$p_{ki}^M = \hat{\mu}_i(x_k) = \frac{\exp(\hat{\alpha}_i + \hat{\beta}_i x_k)}{\sum_{j=1}^{m} \exp(\hat{\alpha}_j + \hat{\beta}_j x_k)}, \quad i = 1, ..., m, \;\; k = 1, ..., N.$$

[23] used the values p_{ki}^M as auxiliary information to define an estimator of class frequencies for nominal response variables. This estimator is in the form

$$\hat{F}_{\mathrm{LV}i} = \sum_{k \in U} p_{ki}^M + \sum_{k \in s} d_k (z_{ki} - p_{ki}^M), \quad i = 1, ..., m, \tag{2.2}$$

where the subindex LV refers to the Lehtonen and Veijanen estimator. An estimator of class proportions can be obtained simply by dividing in (2.2) by population size, N, which is assumed to be known, as follows:

$$\hat{P}_{\mathrm{LV}i} = \frac{1}{N}\hat{F}_{\mathrm{LV}i} = \frac{1}{N}\left(\sum_{k\in U} p_{ki}^{M} + \sum_{k\in s} d_k(z_{ki} - p_{ki}^{M})\right), \quad i = 1, ..., m. \tag{2.3}$$

The sum $\sum_{k\in U} p_{ki}^{M}$ implies that auxiliary information is known for every element in the population. However, when categorical variables (such as gender or the professional status of the individual) or quantitative categorized variables (such as the age of the individual, grouped in classes) are used as auxiliary information in a survey, we may not have a complete list of individuals. Nevertheless, the proposed estimators can still be computed since the population information needed can be found in the databases of national statistical agencies.

3. THE USE OF ORDINAL REGRESSION MODELS IN SURVEY SAMPLING

Let us now assume that the m possible values of Y can be sorted, such that $1 < \cdots < m$. A disadvantage of using multinomial models for ordinal data is that information about the ordering is discarded. Ordinal regression provides a better fit and hence more accurate results. Within ordinal regression models, the most popular is the cumulative logit model, which assumes a linear model for the logit of cumulative probabilities for categories of Y. Given a particular point, the cumulative probability can be defined as the probability that Y falls at or below that point. For the ith category, its cumulative probability can be expressed as

$$P(Y \leq i) = \mu_1 + \cdots + \mu_i, \quad i = 1, ..., m,$$

with $\mu_i = P(Y = i)$. Logit transformations of the cumulative probabilities are, for $i = 1, ..., m-1$,

$$\mathrm{logit}(P(Y \leq i)) = \log\left(\frac{P(Y \leq i)}{1 - P(Y \leq i)}\right) = \log\left(\frac{P(Y \leq i)}{P(Y > i)}\right) = \log\left(\frac{\mu_1 + \cdots + \mu_i}{\mu_{i+1} + \cdots + \mu_m}\right).$$

Note that no logit transformation can be defined for the mth category since, in this case, $P(Y \leq m) = 1$, and so $1 - P(Y \leq m) = 1 - 1 = 0$ and therefore the denominator would be cancelled out. An important property that is usually assumed to be satisfied is that of proportional odds, according to which the effects of the predictors are the same across categories. This implies that β parameters associated with the independent variables are fixed and independent of the category in question. Let us consider

$$P(Y \leq i | X = x_k) = \frac{\exp(\alpha_i + \beta x_k)}{1 + \exp(\alpha_i + \beta x_k)}, \quad i = 1, ..., m-1, \;\; k = 1, ..., N.$$

The cumulative probability for the last category, $P(Y \leq m | X = x_k)$, is always equal to 1. The probability for each category can, then, be calculated as the difference of the cumulative probabilities.

Thus, we propose a superpopulation model ξ with random variables $\boldsymbol{Y} = (Y_1, ..., Y_N)^\top$ such that

$$\mu_1(x_k) = \mathrm{E}_\xi(Z_{k1}|x_k) = \frac{\exp(\alpha_1 + \beta x_k)}{1 + \exp(\alpha_1 + \beta x_k)},$$
$$\mu_i(x_k) = \mathrm{E}_\xi(Z_{ki}|x_k) = \frac{\exp(\alpha_i + \beta x_k)}{1+\exp(\alpha_i + \beta x_k)} - \frac{\exp(\alpha_{i-1} + \beta x_k)}{1+\exp(\alpha_{i-1} + \beta x_k)}, \quad i = 2, ..., m-1,$$
$$\mu_m(x_k) = \mathrm{E}_\xi(Z_{km}|x_k) = 1 - \frac{\exp(\alpha_{m-1} + \beta x_k)}{1 + \exp(\alpha_{m-1} + \beta x_k)}.$$

To define a new estimator for a proportion, using this regression model, we estimate the superpopulation parameter $\theta = (\alpha_1, ..., \alpha_{m-1}, \beta)$ from the units of sample s. After calculating the optimal estimators of the m parameters involved in the model, we can define estimators for individual probabilities

(3.1)

$$p_{k1} = \hat{\mu}_1(x_k) = \frac{\exp(\hat{\alpha}_1 + \hat{\beta} x_k)}{1 + \exp(\hat{\alpha}_1 + \hat{\beta} x_k)},$$
$$p_{ki} = \hat{\mu}_i(x_k) = \frac{\exp(\hat{\alpha}_i + \hat{\beta} x_k)}{1 + \exp(\hat{\alpha}_i + \hat{\beta} x_k)} - \frac{\exp(\hat{\alpha}_{i-1} + \hat{\beta} x_k)}{1 + \exp(\hat{\alpha}_{i-1} + \hat{\beta} x_k)}, \quad i = 2, ..., m-1,$$
$$p_{km} = \hat{\mu}_m(x_k) = 1 - \frac{\exp(\hat{\alpha}_{m-1} + \hat{\beta} x_k)}{1 + \exp(\hat{\alpha}_{m-1} + \hat{\beta} x_k)}.$$

Now, we consider the question of estimating the model parameters. Two general approaches can be adopted to find the optimal estimations of these parameters: (1) by minimizing the sum of the squared distances between the observed and the predicted values (i.e., least squares estimation); or (2) by maximizing the likelihood function (i.e., maximum likelihood estimation or ML estimation).

Weighted least squares method. One way to estimate the parameters of the ordinal logistic regression model is that of least squares. However, in our case, instead of using ordinary least squares, weighted least squares (WLS) must be used. The main difference between the two is that in WLS each observation is weighted using its corresponding survey weight (see e.g. [37]). In this context, WLS involves minimizing, with respect to the residual standard squared error, the weighted distance between the observed outcome (or a function of the observed outcome) and non-linear estimates. In the present case, the function to minimize is

$$S = \sum_{i=1,...,m} \sum_{k \in s} d_k r_{ki}^2,$$

with $r_{ki} = \log\left(P(Y \leq i)/(1 - P(Y \leq i))\right) - \alpha_i - \beta x_k$. This typically requires a numerical procedure, such as the Gauss-Newton method with the Levenberg-Marquardt adjustment (see [19]), which uses derivatives or estimates of derivatives to select the optimal fit. In an iterative fitting process for WLS, assuming ordinal data, at some settings of explanatory variables, the estimated mean may fall below the lowest score or above the highest one and then the fit fails (see [1]).

Maximum likelihood method. Ordered regression models are usually implemented using ML. For ML estimation, the ordinal likelihood function must be numerically maximized to find the parameter values below which the observed data were most likely produced. In theory, these estimates might have the properties of asymptotic efficiency and invariance under parameterization, which makes ML estimation [28] an attractive option in general.

The Nelder Mead simplex [27] is a popular and powerful direct search procedure for likelihood-based optimization. The attraction of this method is that it does not use any derivatives and does not assume that the objective function being optimized has continuous derivatives. In cases such as the present, we expect continuity in the first derivatives and so the latter advantage is not so important. However, this method may be much less efficient or even highly unstable, compared to derivative-based ML estimation methods when sample sizes are as large as the datasets commonly found in complex survey designs.

Let us now examine the logistic likelihood function for modelling ordinal outcomes. As the available data are limited to the sample s, the likelihood function is defined as:

$$L(\boldsymbol{\theta}) = \prod_{i=1,\ldots,m} \prod_{k \in s} \mu_i(x_k)^{z_{ki} d_k}.$$

The pseudolikelihood ([17]; [32]), which is more convenient for use in optimization procedures is given by

$$\log\left(L(\boldsymbol{\theta})\right) = \sum_{i=1,\ldots,m} \sum_{k \in s} d_k z_{ki} \log\left(\mu_i(x_k)\right).$$

ML estimates are obtained by solving a system of m nonlinear equations. Traditionally, two alternatives can be used to address the solution of these equations numerically: Fisher scoring or Newton-Raphson algorithms. Since the results obtained by either method are nearly the same, the decision as to which one to use is trivial (see e.g. [18]).

Various statistical packages can be used to compute the ML estimates of an ordinal logistic model, such as SAS (PROC SURVEYLOGISTIC) or library ordinal for R, but all of them use the Newton-Raphson algorithm to solve the weighted ML equations. The SAS SURVEYLOGISTIC procedure also implements the Fisher scoring algorithm.

4. PROPOSED ESTIMATORS FOR ITEMS WITH ORDINAL OUTCOMES

The estimated individual probabilities (3.1) may be used to define new estimators. We consider a model-assisted approach and a model-calibrated approach to define the following ordinal estimators:

The model-assisted ordinal estimator. Using the idea of the generalized difference predictor given in [5], we define an estimator for proportions of the ordered categories of the response variable as follows

$$\hat{P}_{\mathrm{MA}i} = \frac{1}{N}\left(\sum_{k\in U} p_{ki} + \sum_{k\in s} d_k(z_{ki} - p_{ki})\right), \quad i = 1, ..., m, \tag{4.1}$$

where the subindex MA stands for Model-Assisted.

This estimator is similar to the $\hat{P}_{\mathrm{LV}i}$ estimator proposed by [23] but changes the p_{ki}^M values to p_{ki} values.

The model-calibrated ordinal estimator. A new calibration estimator, let us say P_{MC} (the subindex MC stands for Model-Calibrated), can be defined using the probabilities calculated in (3.1). This estimator is in the form

$$\hat{P}_{\mathrm{MC}i} = \frac{1}{N}\sum_{k\in s} w_k z_{ki}, \quad i = 1, ..., m, \tag{4.2}$$

where, in this case, the weights w_k minimize $G(w_k, d_k)$, and where $G(\cdot, \cdot)$ is a particular distance measure, subject to

$$\sum_{k\in s} w_k p_{ki} = \sum_{k\in U} p_{ki}. \tag{4.3}$$

This is an extension of the model calibration approach proposed by [39]. The distance measure that is usually considered is the chi square

$$\chi = \sum_{k\in s} \frac{(\omega_k - d_k)^2}{d_k q_k}, \tag{4.4}$$

where the q_k's are known positive weights unrelated to d_k. Following [31], section 4.2, and using p_{ki} as an auxiliary variable with a known total $\sum_{k\in U} p_{ki}$, by minimizing (4.4) subject to (4.3) we obtain new weights w_k. By substituting these weights in (4.2) we obtain the following analytic expression for the chi-square calibration estimator:

$$\hat{P}_{\mathrm{MC}i} = \frac{1}{N}\sum_{k\in s} d_k z_{ki} + \frac{1}{N}\left(\sum_{k\in U} p_{ki} - \sum_{k\in s} d_k p_{ki}\right)\hat{B}_i,$$

where $\hat{B}_i = (\sum_{k\in s} d_k p_{ki}^2)^{-1}(\sum_{k\in s} d_k p_{ki} z_{ki})$.

From calibration theory (see [8]), it is well known that all other calibration estimators that use different distance functions are asymptotically equivalent to the chi-square calibration estimator, under additional regularity conditions concerning the shape of the distance function.

So far, we have considered only one auxiliary variable when defining the estimators. These estimators can be easily extended to the general case of p auxiliary variables $\boldsymbol{x} = (x_1, ..., x_p)^\top$ observed for each individual in the population U.

5. PROPERTIES OF THE PROPOSED ESTIMATORS

The most significant properties of the proposed estimators $\hat{P}_{\mathrm{MA}i}$ and $\hat{P}_{\mathrm{MC}i}$ are summarized in this section. To illustrate the asymptotic properties of the proposed classes of estimators, we consider the asymptotic framework of [21], in which the finite population U and the sampling design $p_d(\cdot)$ are embedded into a sequence of populations and designs indexed by N, $\{U_N, p_{d_N}\}$, with $N \to \infty$. We assume therefore, that n tends to infinity as $N \to \infty$. We further assume that $N > 0$. The subscript N may be discarded for ease of notation, although all limiting processes are understood as $N \to \infty$. We denote by E_p the expected value with respect to the sampling design.

The following assumptions are imposed for the sampling design p_d and for the variables:

- **i**) Let $\boldsymbol{\theta_U}$ be the census level parameter estimate obtained by maximizing the likelihood $L(\boldsymbol{\theta})$. Assume that $\boldsymbol{\theta} = \lim_{N\to\infty} \boldsymbol{\theta_U}$ exists and that the pseudomaximum likelihood estimator is $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta_U} + O_p(n^{-1/2})$. [1]
- **ii**) For any study variable h the sampling designs are such that the Horvitz–Thompson estimator for $\overline{h}_N = N^{-1} \sum_{k\in U} h_k$ is asymptotically normal distributed.
- **iii**) Let $B_{iU} = \sum_{k\in U} (\mu_i(x_k)^2)^{-1} \sum_{k\in U} \mu_i(x_k) z_{ki}$. Assume that $B_i = \lim_{N\to\infty} B_{Ui}$ exists, and the sampling design is such that B_i are consistently estimated by $\hat{B}_i$ for $i = 1, ..., m$.

Theorem 5.1. *Under conditions i) and ii) the estimator $\hat{P}_{MAi}$ is approximately design unbiased for P_i, asymptotically normal distributed and the asymptotic design variance is given by*

$$AV_p(\hat{P}_{MAi}) = \frac{1}{N^2} \sum_{k\in U} \sum_{l\in U} \Delta_{kl} (d_k c_{ki})(d_l c_{li}),$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$; $c_{ki} = z_{ki} - \mu_i(x_k, \boldsymbol{\theta_U})$.

[1] This is true under certain regularity conditions given by [3].

Proof: The proof for unbiasedness is very similar to the one presented in [32], page 223, for the difference estimator.

Let us consider the parametric vector $\boldsymbol{t} = (t_1, t_2, ..., t_m)$ and the function $\mu_i(x_k, t_1, t_2, ..., t_m) = \exp(t_i + t_m x_k)/(1 + \exp(t_i + t_m x_k)) - \exp(t_{i-1} + t_m x_k)/(1 + \exp(t_{i-1} + t_m x_k))$. This function has partial derivatives, for each x_k, $\partial\mu_i(x_k, t)/\partial t_j$, which are continuous in $\boldsymbol{t}$ and

$$\frac{\partial\mu_i(x_k, \boldsymbol{t})}{\partial t_i}|_{\boldsymbol{t}=(x_k,\alpha_1,...,\beta)} \leq 1,$$

$$\frac{\partial\mu_i(x_k, \boldsymbol{t})}{\partial t_{i-1}}|_{\boldsymbol{t}=(x_k,\alpha_1,...,\beta)} \leq 1,$$

$$\frac{\partial\mu_i(x_k, \boldsymbol{t})}{\partial t_m}|_{\boldsymbol{t}=(x_k,\alpha_1,...,\beta)} \leq x_k, \text{ and}$$

$$\frac{\partial\mu_i(x_k, \boldsymbol{t})}{\partial t_j}|_{\boldsymbol{t}=(x_k,\alpha_1,...,\beta)} = 0 \text{ for } j \neq i, i-1, m.$$

Thus, by applying the Taylor series expansion at $\boldsymbol{t} = \boldsymbol{\theta_U}$

$$p_{ki} = \mu_i(x_k, \hat{\boldsymbol{\theta}}) = \mu_i(x_k, \boldsymbol{\theta_U}) + \sum_{j=1,...,m} \partial\mu_i(x_k, \boldsymbol{t})/\partial t_j|_{\boldsymbol{t}=(x_k,\alpha_1,...,\beta)}(\hat{\theta}_j - \theta_{Uj}).$$

Under condition i)

$$p_{ki} = \mu_i(x_k, \boldsymbol{\theta_U}) + O_p(n^{-1/2}),$$

and then

$$\frac{1}{N}\sum_{k\in U} p_{ki} - \frac{1}{N}\sum_{k\in U} \mu_i(x_k, \boldsymbol{\theta_U}) = O_p(n^{-1/2}), \text{ and}$$

$$\frac{1}{N}\sum_{k\in s} d_k p_{ki} - \frac{1}{N}\sum_{k\in s} d_k \mu_i(x_k, \boldsymbol{\theta_U}) = O_p(n^{-1/2}).$$

Thus

$$\hat{P}_{\mathrm{MA}i} = \frac{1}{N}\left(\sum_{k\in s} d_k z_{ki} - \sum_{k\in s} d_k \mu_i(x_k, \boldsymbol{\theta_U})\right) + \frac{1}{N}\sum_{k\in U} \mu_i(x_k, \boldsymbol{\theta_U}) + O_p(n^{-1/2}),$$

and the asymptotic design variance of $\hat{P}_{\mathrm{MA}i}$ is the same as that the Horvitz–Thompson estimator $\hat{C}_{\mathrm{HT}i} = 1/N \sum_{k\in s} d_k(z_{ki} - \mu_i(x_k, \boldsymbol{\theta_U}))$.

Condition ii) ensures that estimator $\hat{C}_{\mathrm{HT}i}$ is asymptotically normal distributed and, therefore, estimator $\hat{P}_{\mathrm{MA}i}$ is also asymptotically normal distributed. □

Theorem 5.2. *Under conditions i), ii) and iii) the calibration estimator* $\hat{P}_{MCi}$ *is approximately design unbiased for* P_i*, asymptotically normal distributed and the asymptotic design variance is given by*

$$AV_p(\hat{P}_{MCi}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k e_{ki})(d_l e_{li}),$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ *and* $e_{ki} = z_{ki} - \mu_i(x_k, \boldsymbol{\theta_U})B_{iU}$.

Proof:

$$\hat{P}_{\mathrm{MC}i} = \frac{1}{N}\sum_{k\in s} d_k z_{ki} + \frac{1}{N}\left(\sum_{k\in U} p_{ki} - \sum_{k\in s} d_k p_{ki}\right) B_{iU} + \frac{1}{N}\left(\sum_{k\in U} p_{ki} - \sum_{k\in s} d_k p_{ki}\right)\left(\hat{B}_i - B_{iU}\right).$$

Under condition iii) $\hat{B}_i - B_{iU} = o(1)$; under conditions i) and ii)

$$\frac{1}{N}\sum_{k\in U} p_{ki} - \frac{1}{N}\sum_{k\in s} d_k p_{ki} = \frac{1}{N}\sum_{k\in U}\mu_i(x_k, \boldsymbol{\theta_U}) - \frac{1}{N}\sum_{k\in s} d_k\mu_i(x_k, \boldsymbol{\theta_U}) + O_p(n^{-1/2}).$$

Thus

$$\hat{P}_{\mathrm{MC}i} = \frac{1}{N}\sum_{k\in s} d_k z_{ki} + \frac{1}{N}\left(\sum_{k\in U}\mu_i(x_k, \boldsymbol{\theta_U}) - \sum_{k\in s} d_k\mu_i(x_k, \boldsymbol{\theta_U})\right) B_{iU} + o_p(n^{-1/2}),$$

and consequently

$$\mathrm{E}_p(\hat{P}_{\mathrm{MC}i}) \to \mathrm{E}_p\left(\frac{1}{N}\sum_{k\in s} d_k z_{ki}\right) = P_i,$$

and

$$\mathrm{V}_p(\hat{P}_{\mathrm{MC}i}) \to \mathrm{V}_p\left(\frac{1}{N}\sum_{k\in s} d_k\left(z_{ki} - \mu_i(x_k, \boldsymbol{\theta_U})\right) B_{iU}\right).$$

Under condition ii), estimator $(1/N)\sum_{k\in s} d_k(z_{ki} - \mu_i(x_k, \boldsymbol{\theta_U}))B_{iU}$ is asymptotically normal distributed, and therefore we conclude that estimator $\hat{P}_{\mathrm{MC}i}$ is also asymptotically normal distributed.

□

6. ESTIMATION FOR THE VARIANCE OF ORDINAL ESTIMATORS

The next theorem gives analytic expressions for the estimators of the design variances $\mathrm{V}_p(\hat{P}_{\mathrm{MA}i})$ and $\mathrm{V}_p(\hat{P}_{\mathrm{MC}i})$, obtained using the linearization method.

Theorem 6.1. *Under conditions i), ii) and iii) and assuming that all second order probabilities are non null,*

$$\widehat{V}(\hat{P}_{MAi}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}(d_k\tilde{c}_{ki})(d_l\tilde{c}_{li}) \quad (Lin) \tag{6.1}$$

is approximately design unbiased for $V_p(\hat{P}_{MAi})$ *and*

$$\widehat{V}(\hat{P}_{MCi}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}(d_k\tilde{e}_{ki})(d_l\tilde{e}_{li}) \quad (Lin) \tag{6.2}$$

is approximately design unbiased for $V_p(\hat{P}_{MCi})$ *where* $\tilde{c}_{ki} = z_{ki} - p_{ki}$ *and* $\tilde{e}_{ki} = z_{ki} - p_{ki}\widehat{B}_i$.

Proof: We denote by I_k the sample membership indicator of element k. Thus, for each $i = 1, ..., m$:

$$\mathrm{E}_p(\widehat{\mathrm{V}}(\hat{P}_{\mathrm{MA}i})) = \frac{1}{N^2}\mathrm{E}_p\sum_{k\in U}\sum_{l\in U}\frac{\Delta_{kl}}{\pi_{kl}}(d_k\tilde{c}_{ki})(d_l\tilde{c}_{li})I_k(s)I_l(s) =$$

$$= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\frac{\Delta_{kl}}{\pi_{kl}}(d_k\tilde{c}_{ki})(d_l\tilde{c}_{li})\pi_{kl} \to \mathrm{V}_p(\hat{P}_{\mathrm{MA}i}),$$

using the theorem 5.1. From the theorem 5.2, the estimator of the design variance $\mathrm{V}_p(\hat{P}_{\mathrm{MC}i})$ can be derived. □

These variance estimators require knowledge of second-order inclusion probabilities, which are often impossible to compute or unavailable to data analysts for complex sampling designs. A simple alternative is to use with-replacement variance estimators (see [32], page 99). For the $\hat{P}_{\mathrm{MA}i}$ estimator, the with-replacement variance estimator is

$$\hat{\mathrm{v}}_{\text{W-R}}(\hat{P}_{\mathrm{MA}i}) = \frac{1}{N^2}\frac{1}{n(n-1)}\sum_{k\in s}\left(\frac{\tilde{c}_{ki}}{pr_k} - \frac{1}{n}\sum_{j\in s}\frac{\tilde{c}_{ji}}{pr_j}\right)^2 \quad \text{(W-R)},$$

where $pr_k = \pi_k/n$ when we have a simple random sampling without replacement design. For other sampling designs, the relationship between pr_k and π_k is $\pi_k = 1 - (1 - pr_k)^n$ according to expression (2.9.5), page 51 in [32].

The with-replacement variance estimator for $\hat{P}_{\mathrm{MC}i}$ is obtained in a similar way:

$$\hat{\mathrm{v}}_{\text{W-R}}(\hat{P}_{\mathrm{MC}i}) = \frac{1}{N^2}\frac{1}{n(n-1)}\sum_{k\in s}\left(\frac{\tilde{e}_{ki}}{pr_k} - \frac{1}{n}\sum_{j\in s}\frac{\tilde{e}_{ji}}{pr_j}\right)^2 \quad \text{(W-R)}.$$

These with-replacement variance estimators are not without bias. An expression for the bias can be obtained using the theory of sampling with probability proportional to size (see [11] or [32]).

Alternative variance estimators can be obtained using implicit differentiation [3] or replicated sampling methods (see [37] for a detailed description of these techniques in finite population sampling). The replicated methods estimate the variance of a parameter by generating replicated subsamples and examining the variability of the subsample estimates. The replicated methods, also referred to as resampling methods, include balanced repeated replication (BRR), jackknife repeated replication (JRR) [34] and the bootstrap method [12]. This article focuses on jackknife techniques due to their simplicity and because they are implemented in general purpose software packages, such as R (see for example the packages sampling [33], samplingVarEst [14] and samplingEstimates [13]).

For a non stratified design, the jackknife estimator of the variance for any of the model-assisted estimators, $\hat{P}_{\mathrm{MA}i}$ is given by

$$\hat{\mathrm{v}}_{\mathrm{J}}(\hat{P}_{\mathrm{MA}i}) = \frac{n-1}{n}\sum_{j\in s}(\hat{P}_{\mathrm{MA}i}(j) - \overline{P}_{\mathrm{MA}i})^2 \quad \text{(Tukey)}, \tag{6.3}$$

where $\hat{P}_{\mathrm{MA}i}(j)$ is the value of the estimator $\hat{P}_{\mathrm{MA}i}$ after dropping unit j from s and where $\overline{P}_{\mathrm{MA}i}$ is the mean of values $\hat{P}_{\mathrm{MA}i}(j)$.

The jackknife estimator may present an important bias when designs without replacement are used in finite populations. In such a case, an approximated finite-population correction could be incorporated into the estimation in order to achieve unbiasedness. A modified jackknife estimator of variance, $\hat{\mathrm{v}}^*_{\mathrm{J}}(\hat{P}_i)$, can be calculated by replacing $\hat{P}_{\mathrm{MA}i}(j)$ in (6.3) with $\hat{P}^*_{\mathrm{MA}i}(j) = \hat{P}_{\mathrm{MA}i} + \sqrt{1-\overline{\pi}}(\hat{P}_{\mathrm{MA}i}(j) - \overline{P}_{\mathrm{MA}i})$, where $\overline{\pi} = \sum_{k\in s}\pi_k/n$.

Using the idea of the unequal probability jackknife variance estimator given by [4], we can obtain a new estimator $\hat{\mathrm{v}}_{\mathrm{JC}}(\overline{P}_{\mathrm{MA}i})$ by replacing $\tilde{c}_{ki}$ in (6.1) with $c\tilde{m}_{ki} = 1 - \widetilde{d}_k(\widetilde{C}_{\mathrm{HT}i} - \widetilde{C}_{\mathrm{HT}i}(k))$ where $\widetilde{C}_{\mathrm{HT}i} = 1/N\sum_{k\in s} d_k\tilde{c}_{ki}$, $\widetilde{C}_{\mathrm{HT}i}(k)$ is the Horvitz–Thompson estimator dropping the unit k of the sample and $\widetilde{d}_k = d_k/\sum_{l\in s} d_l$. The design consistency of this type of variance estimator was highlighted in [2].

More recently, [15] formulated a new design-consistent variance estimator for the population mean. Based on this idea, we can obtain a new variance estimator $\hat{\mathrm{v}}_{\mathrm{JEB}}(\overline{P}_{\mathrm{MA}i})$ by replacing $\tilde{c}_{ki}$ in (6.1) with $\widetilde{ce}_{ki} = d_k^{\alpha_k}(\widetilde{C}_{\mathrm{HT}i} - \widetilde{C}_{\mathrm{HT}i}(k))$. The authors propose d the use of $\alpha_k = 1\ \forall k \in s$.

Similarly, we define jackknife variance estimators for the ordinal calibration estimator.

7. MONTE CARLO SIMULATION EXPERIMENTS

To determine the behaviour of the estimators when they are applied to real data obtained through complex sampling designs, we consider data from the 2012 PISA survey. This is a macro-surveying procedure that is conducted every three years to collect information about 15-year-old students in each of the 65 countries participating. The main aim of the survey is to determine how well students are prepared to meet the challenges of the future. To do so, their performance and attitudes are measured in three key areas: mathematics, reading and science.

The 2012 PISA survey was focused on mathematics in particular, and so the students were asked to indicate their degree of agreement with various statements related to mathematics. The population considered for our study was composed of $N = 15,499$ 15-year-old Spanish students who responded to the survey, and who attended $C = 838$ different schools. We chose the question "How strongly do you agree with the statement: I enjoy reading about mathematics?" as the main variable, where the possible options were `1` = strongly agree, `2` = agree, `3` = disagree and `4` = strongly disagree. The population percentages obtained for these categories were 0.03, 0.149, 0.413 and 0.408, respectively. We then considered the degree of agreement (expressed as Strongly agree, agree, disagree and strongly disagree) with to the following sentences: "Making an effort in mathematics is worth it because it will help me in the work that I want to do later on", "Learning mathematics is worthwhile for me because it will improve my career" and "I will learn many things in mathematics that will help me get a job" as auxiliary variables.

With these data as population, we used a stratified design, selecting a sample of schools with probabilities proportional to their size within each stratum. Then, the values of all the students at the selected schools were observed. The population was divided into five different strata depending on the type of location of each school: villages (fewer than 3,000 people), small towns (3,000 to 15,000 people), towns (15,000 to 100,000 people), cities (100,000 to 1,000,000 people) and large cities (over 1,000,000 people). The number of schools (C_h) and students (N_h) by stratum is detailed in Table 1.

Table 1: Strata population data.

	Villages	Small towns	Towns	Cities	Large cities	Total
C_h	48	239	254	269	28	838
N_h	831	4,312	4,795	5,046	515	15,499

Two sample sizes for schools ($c = 25$ and $c = 50$) are included in the study. A sample of schools using a Midzuno sampling scheme was drawn from each

stratum considering probabilities proportional to the school size (taken as the number of students enrolled in the school).

The free statistical software R ([30]) was used to perform this simulation study. The library ordinal of R ([7]) was used, where necessary, to estimate the parameters of the ordinal model. We have developed new R-code implementing the proposed estimators. The R libraries samplingVarEst ([14]) and samplingEstimates ([13]) were used to estimate the variance of the estimators according to the different methods discussed. For each estimator, we computed the percent relative bias $\text{RB}\% = \text{E}_{\text{MC}}(\hat{P} - P)/P * 100\%$ and the percent relative mean squared error $\text{RMSE}\% = \text{E}_{\text{MC}}[(\hat{P} - P)^2]/P^2 * 100\%$ for each category of the main variable Y based on 1,000 simulation runs. We used RMSE% to calculate the percent relative efficiency gain with respect to the HT estimator for the three remaining estimators. The minimum, maximum and mean percent over the categories are also calculated (in absolute values for the relative bias).

The results for relative bias and relative efficiency based on 1,000 simulated samples are shown in Table 2. Additionally, the mean number of students finally observed in each scenario, $\bar{n}$, is included for informative purposes.

Table 2: Relative bias (in % and Italics) and Relative efficiency (with respect to the HT estimator) of the estimators. Auxiliary variables: "Making an effort in...", "Learning mathematics is...", "I will learn many...".

Estimator	1	2	3	4	min	max	mean
	$c = 25$ $(\bar{n} = 482.88)$						
HT	*0.35* 100.00	*−0.02* 100.00	*−0.34* 100.00	*−0.25* 100.00	*0.02* 100.00	*0.35* 100.00	*0.24* 100.00
LV	*0.74* 111.37	*0.30* 141.94	*−0.10* 329.21	*−0.06* 318.49	*0.06* 111.37	*0.74* 329.21	*0.30* 225.25
MA	*0.73* 111.71	*0.43* 143.76	*−0.08* 372.79	*−0.13* 367.04	*0.08* 111.71	*0.73* 372.79	*0.34* 248.82
MC	*0.78* 110.65	*0.48* 144.05	*−0.11* 381.46	*−0.12* 374.35	*0.11* 110.65	*0.78* 381.46	*0.37* 252.62
	$c = 50$ $(\bar{n} = 966.43)$						
HT	*−0.40* 100.00	*−0.37* 100.00	*−0.58* 100.00	*−0.45* 100.00	*0.37* 100.00	*0.58* 100.00	*0.45* 100.00
LV	*−0.02* 111.92	*0.09* 131.05	*−0.11* 333.83	*0.07* 313.76	*0.02* 111.92	*0.11* 333.83	*0.07* 222.64
MA	*0.03* 112.31	*0.17* 131.56	*−0.09* 389.68	*0.03* 381.50	*0.03* 112.31	*0.17* 389.68	*0.08* 253.76
MC	*−0.02* 112.83	*0.19* 132.42	*−0.09* 397.13	*0.03* 389.41	*0.02* 112.83	*0.19* 397.13	*0.08* 257.94

1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree

Relative bias is below 1% in all cases and can be considered negligible. Both model-assisted and model-calibrated estimators show good performance in terms of efficiency, with the first of these showing slightly better results. Whatever the estimator, the most accurate estimations are achieved in categories 3 and 4, those with the largest population sizes.

The efficiency of these estimators is greater than that of the HT estimator in all cases, and is especially high for categories 3 and 4. As the sample size increases, so does the relative efficiency of the ordinal estimators, with values close to 400% in categories 3 and 4 for $c = 50$.

An alternative model was then fitted for the same variable response, taking the student's gender and the educational level of the father and mother as auxiliary variables. With these covariates it was not possible to obtain a good model fit, since they achieved a very low association with the main variable in the population. Indeed, the Akaike Information Criterion (AIC) in this case was noticeably higher than the value obtained for the previous model fit.

Table 3 shows the results of relative bias and relative efficiency of the estimators for these variables.

Table 3: Relative bias (in % and Italics) and relative efficiency (with respect to the HT estimator) of the estimators. Auxiliary variables: Sex of the student, educational level of of the father and that of the mother.

Estimator	1	2	3	4	min	max	mean
	$c = 25$ $(\bar{n} = 460.03)$						
HT	*0.56* 100.00	*−0.08* 100.00	*−0.41* 100.00	*−0.13* 100.00	*0.08* 100.00	*0.56* 100.00	*0.29* 100.00
LV	*1.01* 108.00	*0.37* 126.32	*−0.24* 329.57	*0.04* 293.39	*0.04* 108.00	*1.01* 329.57	*0.41* 214.32
MA	*0.79* 110.00	*0.27* 128.24	*−0.22* 333.88	*0.06* 288.83	*0.06* 110.00	*0.79* 333.88	*0.33* 215.23
MC	*0.68* 110.11	*0.31* 127.97	*−0.24* 335.45	*0.08* 291.91	*0.08* 110.11	*0.68* 335.45	*0.33* 216.36
	$c = 50$ $(\bar{n} = 920.96)$						
HT	*−0.36* 100.00	*−0.38* 100.00	*−0.61* 100.00	*−0.40* 100.00	*0.36* 100.00	*0.61* 100.00	*0.44* 100.00
LV	*0.53* 107.71	*0.36* 113.25	*−0.20* 340.53	*0.03* 284.22	*0.03* 107.71	*0.53* 340.53	*0.28* 211.42
MA	*0.13* 109.11	*0.19* 114.43	*−0.15* 344.48	*0.07* 278.70	*0.07* 109.11	*0.19* 344.48	*0.14* 211.68
MC	*0.06* 109.68	*0.24* 113.39	*−0.17* 345.99	*0.08* 280.24	*0.06* 109.68	*0.24* 345.99	*0.14* 212.32

1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree

The results for this population presented a similar pattern: both model-assisted and model-calibrated ordinal estimators achieved very good performance, but in this case, the differences between the estimators are not significant. Efficiency gains with respect to the HT estimator are smaller in this scenario than in the previous one. However, once again, the largest efficiency gains are obtained in categories 3 and 4.

Alternative scenarios were also considered, and these yielded similar results. Specifically, even in the case in which tests of the proportional odds assumption provided evidence of the non-proportional odds context, the efficiency results were comparable.

In a similar way, we computed confidence intervals using different methods to estimate the variance of the estimators. Tables 4 and 5 show the relative length (length / parameter) in % and the empirical coverage of the confidence intervals, in the first case for a good model fit and in the second for a bad one.

Table 4: Relative length in % (LEN) and empirical coverage (COV) of confidence intervals for the estimators using different estimators for variance. Auxiliary variables: "Making an effort in...", "Learning mathematics is...", "I will learn many...". Nominal level 95%.

Estimator		LV		MA		MC		LV		MA		MC	
		LEN	COV	LEN	COV	LEN	COV	LEN	COV	LEN	COV	LEN	COV
		$c = 25$						$c = 50$					
Lin	1	115.00	89.8	115.83	90.3	115.75	90.7	84.69	92.0	84.97	91.8	84.95	92.0
	2	49.82	92.8	49.22	93.9	49.23	93.2	35.42	92.6	35.03	92.5	35.02	92.7
	3	25.45	91.4	23.96	91.5	23.94	91.8	18.09	92.2	17.13	92.5	17.12	93.1
	4	26.68	91.7	24.22	93.0	24.22	93.2	18.99	93.0	17.35	91.8	17.36	92.0
	mean	54.24	91.4	53.31	92.0	53.29	92.2	39.30	92.4	38.62	92.2	38.61	92.5
W-R	1	111.24	89.8	111.97	90.4	111.74	90.4	81.96	92.3	82.20	92.2	82.12	92.1
	2	46.77	91.9	46.36	92.0	46.32	91.9	33.21	90.8	32.92	91.1	32.91	90.8
	3	24.57	92.7	24.05	93.6	24.05	94.1	17.38	92.7	16.98	94.0	16.98	94.2
	4	23.98	91.1	22.98	92.6	22.98	93.1	16.94	90.6	16.23	91.2	16.23	91.7
	mean	51.64	91.4	51.34	91.2	51.27	92.4	37.37	91.6	37.08	92.1	37.06	92.2
EB	1	114.31	89.4	115.35	90.6	115.17	90.6	84.74	91.7	85.11	91.6	85.10	91.6
	2	49.98	93.3	48.94	93.2	48.91	93.3	35.71	92.7	34.99	92.7	34.98	92.8
	3	24.90	91.1	23.81	91.4	23.81	91.9	17.92	91.8	17.13	93.1	17.13	93.6
	4	26.13	92.0	24.10	92.9	24.10	93.1	18.80	91.7	17.36	92.1	17.37	92.1
	mean	53.83	91.5	53.05	92.0	53.00	92.2	39.29	92.0	38.65	92.4	38.65	92.5
CBS	1	114.30	89.4	115.34	90.6	115.16	90.6	84.73	91.7	85.11	91.6	85.10	91.6
	2	49.98	93.3	48.94	93.2	48.90	93.3	35.71	92.7	34.99	92.7	34.98	92.8
	3	24.90	91.1	23.81	91.4	23.81	91.9	17.92	91.8	17.13	93.1	17.13	93.6
	4	26.13	92.0	24.10	92.9	24.10	93.1	18.80	91.7	17.36	92.1	17.37	92.1
	mean	53.83	91.5	53.05	92.0	52.99	92.2	39.29	92.0	38.65	92.4	38.64	92.5
Tukey	1	110.94	90.0	111.68	90.3	111.44	90.3	82.08	92.3	82.32	92.4	82.25	92.2
	2	46.74	92.1	46.30	92.4	46.26	92.6	33.31	91.7	32.99	91.3	32.98	90.9
	3	24.47	92.8	23.97	93.6	23.98	94.4	17.40	92.5	17.01	93.9	17.01	94.4
	4	23.88	90.7	22.91	92.4	22.90	92.7	16.95	90.6	16.25	91.7	16.25	92.3
	mean	51.51	91.4	51.22	92.2	51.14	92.5	37.43	91.8	37.14	92.3	37.12	92.5

1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree

Table 5: Relative length in % (LEN) and empirical coverage (COV) of confidence intervals for compared estimators using different estimators for variance. Auxiliary variables: Sex of the student, educational level of father and educational level of mother. Nominal level 95%.

Estimator		LV		MA		MC		LV		MA		MC	
		LEN	COV	LEN	COV	LEN	COV	LEN	COV	LEN	COV	LEN	COV
		$c = 25$						$c = 50$					
Lin	1	123.09	90.1	122.30	89.1	122.31	88.6	90.16	91.7	89.86	91.4	89.87	91.9
	2	53.27	93.3	53.04	92.9	53.03	93.2	37.78	93.1	37.60	92.8	37.62	93.0
	3	25.45	90.6	25.43	90.7	25.43	91.4	18.13	91.8	18.13	92.3	18.13	92.1
	4	27.84	91.6	27.84	92.6	27.84	92.3	20.00	92.5	20.05	92.6	20.05	92.3
	mean	57.41	91.4	57.15	91.3	57.15	91.3	41.52	92.3	41.41	92.3	41.42	92.3
W-R	1	119.22	89.2	118.28	89.6	118.36	89.9	87.81	91.6	87.14	92.1	87.18	91.9
	2	49.39	91.8	48.94	92.2	48.91	91.5	34.96	91.6	34.60	91.8	34.59	91.8
	3	25.18	92.9	25.26	93.0	25.26	92.5	17.80	92.6	17.83	93.2	17.83	93.1
	4	25.52	91.4	25.34	91.4	25.34	91.0	18.07	90.7	17.95	90.6	17.95	90.4
	mean	54.82	91.3	54.45	91.5	54.47	91.2	39.66	91.6	39.38	91.9	39.38	91.8
EB	1	122.34	89.5	121.73	89.0	121.82	88.9	90.16	91.9	89.97	91.9	90.02	92.1
	2	52.43	92.9	52.81	93.7	52.78	93.2	37.40	92.6	37.64	92.9	37.62	93.1
	3	25.19	90.9	25.25	91.0	25.25	91.1	18.09	91.6	18.13	92.3	18.13	92.5
	4	27.62	91.6	27.70	92.2	27.69	92.6	19.95	92.4	20.06	92.4	20.06	92.8
	mean	56.89	91.2	56.87	91.4	56.89	91.4	41.40	92.1	41.45	92.4	41.45	92.6
CBS	1	122.33	89.5	121.73	89.0	121.81	88.9	90.15	91.9	89.97	91.9	90.01	92.1
	2	52.43	92.9	52.81	93.7	52.78	93.2	37.40	92.6	37.63	92.9	37.62	93.1
	3	25.19	90.9	25.25	91.0	25.25	91.1	18.09	91.6	18.13	92.3	18.13	92.5
	4	27.61	91.6	27.69	92.2	27.69	92.6	19.95	92.4	20.06	92.4	20.06	92.8
	mean	56.89	91.2	56.87	91.4	56.88	91.4	41.40	92.1	41.45	92.4	41.45	92.6
Tukey	1	118.85	89.5	117.96	88.5	118.04	88.5	87.93	91.9	87.27	92.3	87.31	92.1
	2	49.28	92.6	48.87	92.7	48.84	91.8	35.01	91.7	34.69	91.6	34.67	92.1
	3	25.08	92.8	25.17	93.1	25.17	92.9	17.82	92.2	17.86	93.0	17.85	93.0
	4	25.43	91.6	25.26	91.8	25.25	91.4	18.09	91.2	17.98	90.9	17.98	91.0
	mean	54.66	91.6	54.31	91.5	54.33	91.1	39.72	91.7	39.45	91.9	39.45	92.0

1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree

It is no easy matter to compare all the estimators and all the variance estimation techniques over all the categories. However, the tables obtained show that the lengths of the EB (Escobar-Berger, [15]) and CBS (Campbell, [4]; Berger and Skinner, [2]) intervals are practically the same, and also that the intervals with the LV estimator have longer lengths, while those with the MC estimators have shorter ones, for both sample sizes. Obviously, the length of the confidence intervals decreases as the sample size increases.

The coverage is below the nominal value in every case. The MC estimator obtains the closest coverage to the nominal level, but with small differences with respect to the other estimators.

8. APPLICATION TO AN OPINION SURVEY

In this section, the ordinal regression approach is illustrated using a real survey, deriving the proposed estimates and comparing these to alternative ones.

This population-based survey was conducted by the Institute of Social Studies of Andalusia, a public scientific research institute specialising in the social sciences. Its aim is to reflect the opinions of the population of Andalusia, a region in Southern Spain, with regard to various aspects of policies. Taking into account the time and budget available, 1,890 interviews were performed by qualified interviewers, specially trained in survey techniques. The interviews were carried out by the Statistics and Surveys sections of the institute using Computer Assisted Telephone Interviewing data input techniques. A stratified random sampling design with eight strata, each one corresponding to a municipality in the region, was considered. In each stratum, a simple random sampling without replacement design was considered. The design weights were modified to adjust for coverage and non-response bias. The two main variables included in this study, related to "education" and "housing", are the answers to the following questions:

- *Do you think that education issues have improved, remain the same or have worsened in recent years?*
- *Do you think that housing issues have improved, remain the same or have worsened in recent years?*

each one with three possible response categories. As in the simulation study, R software and the library ordinal were used to analyze the data. Together with the ordinal model-assisted MA (4.1) and the ordinal model-calibrated MC (4.2) estimators, the HT estimator (2.1) and the LV estimator (2.3) were computed for comparison purposes. As auxiliary information we took into account the sex of the respondents, a categorical variable with two possible outcomes, and their age, categorized into four age ranges. The population information for the auxiliary variables needed to compute the LV estimator and the two proposed estimators is shown in Table 6.

Table 6: Population information for the auxiliary variables (Sex and Age).

Sex	*Age*			
	18–29	30–44	45–59	≥ 60
Male	411,501	699,378	636,061	578,775
Female	460,834	649,434	615,057	731,410

Table 7 shows the point and 95% confidence interval estimation of proportions of each category of the main variables.

Table 7: Point (PROP) and 95% confidence level estimation (lower bound, LB, upper bound, UB, and length, LEN) of percentages. Auxiliary variables: Sex and Age.

Estimator	*In recent years, education issues...*				*In recent years, housing issues...*			
	PROP	LB	UB	LEN	PROP	LB	UB	LEN
	... have improved				*... have improved*			
HT	3.88	2.61	5.16	2.55	7.15	5.74	8.55	2.81
LV	4.49	3.11	5.87	2.76	7.65	6.10	9.20	3.10
MA	3.92	2.67	5.18	2.51	7.08	5.69	8.47	2.78
MC	3.94	2.68	5.19	2.51	7.07	5.68	8.47	2.79
	... remain the same				*... remain the same*			
HT	17.69	15.44	19.94	4.50	9.42	7.80	11.04	3.24
LV	18.12	15.87	20.37	4.50	9.87	8.12	11.63	3.51
MA	17.84	15.64	20.03	4.39	9.35	7.74	10.96	3.22
MC	17.82	15.63	20.02	4.39	9.36	7.76	10.97	3.21
	... have worsened				*... have worsened*			
HT	78.41	74.59	82.24	7.65	83.42	79.52	87.32	7.80
LV	77.38	74.76	79.99	5.23	82.47	80.00	84.93	4.93
MA	78.22	75.82	80.63	4.81	83.56	81.52	85.59	4.07
MC	78.23	75.83	80.63	4.80	83.55	81.52	85.59	4.07

Whatever the category of either of the two main variables, the lengths of the confidence intervals of the proposed estimators are shorter than that of the corresponding confidence interval associated with the LV estimator, which uses the same amount of auxiliary information. In part, these differences are due to the better fit of the ordinal logistic model than the multinomial logistic model in both cases. Indeed, for the two response variables, the AIC is larger for the multinomial model than for the ordinal model. To highlight these discrepancies, we computed the relative length reduction of the confidence intervals of the proposed estimators with respect to the corresponding confidence intervals of the LV estimator. The results are shown in Table 8.

The length reductions are significant in all categories of the response variables (6.5% on average for the first variable and 12% for the second).

Tables 9 and 10 show the point estimation for the proposed estimators, classified by sex and age. In the first of these respects, it is noticeable that more men than women believe that education and housing issues have improved or remain the same, while the women are slightly more pessimistic.

The general perception that these issues have worsened is common to all age groups, with the highest such proportion being found among respondents aged 45-59 years.

Table 8: Relative length reduction in % of the 95% confidence intervals of the proposed estimators with respect to the LV estimator.

In recent years, education issues...				
Estimator	REDUCTION			
	... have improved	*... remain the same*	*... have worsened*	MEAN
MA	9.326725	2.359543	8.013819	6.566695
MC	9.349046	2.371333	8.014386	6.578255

In recent years, housing issues...				
Estimator	REDUCTION			
	... have improved	*... remain the same*	*... have worsened*	MEAN
MA	10.12919	8.518985	17.41836	12.022178
MC	10.10099	8.540746	17.42298	12.021572

Table 9: Point estimation of percentages by sex.

Estimator	*In recent years, education issues...*			*In recent years, housing issues...*		
	ALL	MEN	WOMEN	ALL	MEN	WOMEN
	... have improved			*... have improved*		
HT	3.88	4.65	3.15	7.15	10.00	4.39
LV	4.49	4.96	4.05	7.65	10.33	5.11
MA	3.92	4.69	3.21	7.08	9.97	4.34
MC	3.94	4.69	3.16	7.07	9.99	4.38
	... remain the same			*... remain the same*		
HT	17.69	20.71	14.80	9.42	11.59	7.33
LV	18.12	21.27	15.15	9.87	11.97	7.89
MA	17.84	20.83	15.01	9.35	11.49	7.32
MC	17.82	20.72	14.84	9.36	11.58	7.32
	... have worsened			*... have worsened*		
HT	78.41	74.63	82.05	83.42	78.40	88.27
LV	77.38	73.76	80.80	82.47	77.69	86.99
MA	78.22	74.47	81.78	83.56	78.52	88.32
MC	78.23	74.58	81.99	83.55	78.42	88.28

Table 10: Point estimation of percentages by age groups.

Estimator	*In last years, education issues...*					*In last years, housing issues...*				
	ALL	18–29	30–44	45–59	≥60	ALL	18–29	30–44	45–59	≥60
	... have improved					*... have improved*				
HT	3.88	3.99	4.24	1.92	5.61	7.15	8.66	8.34	4.94	6.85
LV	4.49	4.20	4.15	1.87	7.54	7.65	8.90	8.24	4.86	8.87
MA	3.92	4.01	4.26	1.84	5.51	7.08	8.52	8.50	4.99	6.66
MC	3.94	4.09	4.22	1.89	5.66	7.07	8.59	8.29	4.92	6.76
	... remain the same					*... remain the same*				
HT	17.69	16.75	17.02	17.06	20.07	9.42	10.16	11.47	8.90	7.07
LV	18.12	16.87	16.97	16.99	21.23	9.87	10.44	11.33	8.74	9.07
MA	17.84	16.31	17.17	17.35	20.02	9.35	9.68	11.61	9.13	7.01
MC	17.82	16.84	17.08	17.22	20.16	9.36	10.11	11.43	8.86	7.02
	... have worsened					*... have worsened*				
HT	78.41	79.25	78.73	81.01	74.31	83.42	81.17	80.17	86.14	86.07
LV	77.38	78.92	78.87	81.13	71.21	82.47	80.65	80.42	86.38	82.04
MA	78.22	79.67	78.56	80.79	74.46	83.56	81.78	79.88	85.87	86.32
MC	79.06	78.69	80.87	80.78	74.16	83.55	81.29	80.27	86.21	86.20

9. CONCLUSIONS

Data collected from surveys are often organized into discrete categories. Analyzing variables with ordinal outcomes, obtained from a complex survey, often requires specialised techniques. To improve the accuracy of estimation procedures, a survey statistician often makes use of the auxiliary data available from administrative registers and other sources.

In this paper, we present estimation techniques applied to the results of complex surveys when the variable of interest has ordinal outcomes, and describe the joint distribution of the class indicators by an ordinal model. Ordinal model-assisted estimators and ordinal model-calibrated estimators are introduced for class frequencies, using two different approaches to estimation.

We show that the proposed estimators are asymptotically normal distributed and we derive expressions for their asymptotic variances. Resampling techniques are obtained when joint inclusion probabilities are unavailable to data analysts.

We used the weighted ML estimation procedure to obtain the estimators for the model parameters because in the iterative fitting process for WLS, assuming ordinal data, at some settings of explanatory variables the estimated mean may fall below the lowest score or above the highest one and then the fit will fail [1]. When numerical maximization for the pseudolikelihood is feasible, good estimates may be obtained in certain cases by WLS. This approach is usable when working with discrete predictors [23].

We also include a limited simulation study with a real population, finding that the ordinal logistic formulation yields better results than the classical estimators that implicitly assume individual linear models for the variables.

The effective use of auxiliary information from survey data depends on the population quantities to be estimated and on the actual relation between the response variable and the covariates. The simulation results obtained show that these estimators are robust against misspecified models.

Ordinal model-assisted and model-calibrated estimators also have some drawbacks: they require a sampling frame, complete with all the explanatory variables used in the assisting model, for all units in the population. This situation frequently arises, for example, when categorical variables (such as gender or the professional status of the individual) or quantitative categorized variables (such as the age of the individual, grouped into classes) are used as auxiliary information in a survey. In this context, although we do not have a complete list of individuals, the proposed estimators can still be computed because the necessary population information can be found in the databases published by national statistical agencies and in business registers and trade association lists. This is the case in our application of the estimators to data from the survey on opinions and attitudes, as discussed in Section 8.

ACKNOWLEDGMENTS

This study was partially supported by Ministerio de Economía y Competitividad (grant MTM2015-63609-R and FPU grant programme, Spain) and by Consejería de Economía, Innovación, Ciencia y Empleo (grants SEJ2954 and HUM1413 Junta de Andalucía, Spain).

REFERENCES

[1] Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, second ed. Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

[2] Berger, Y.G. and Skinner, C.J. (2005). A jackknife variance estimator for unequal probability sampling, *Journal of the Royal Statistical Society B*, **67**, 79–89.

[3] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review/Revue Internationale de Statistique*, **51**, 279–292.

[4] Campbell, C. (1980). A different view of finite population estimation, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

[5] Cassel, C.M.; Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations, *Biometrika*, **63**, 615–620.

[6] Coelho, P.S. and Casimiro, F. (2008). Post Enumeration Survey of the 2001 portuguese population and housing censuses, *REVSTAT – Statistical Journal*, **6**(3), 231–252.

[7] Christensen, R.H.B. ordinal - Regression Models for Ordinal Data. R package version 2015, 6–28.

[8] Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376–382.

[9] Diana, G. and Perri, P.F. (2013). A class of estimators in two-phase sampling with subsampling the non-respondents, *Applied Mathematics and Computation*, **219**(19), 10033–10043.

[10] Duchesne, P. (2003). Estimation of a proportion with survey data, *Journal of Statistics Education*, **11**(3), 1–24.

[11] Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities, *Journal of the Royal Statistical Society*, **15**, 262–269.

[12] Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *The Annals of Statistics*, **7**, 1–26.

[13] Escobar, E.L. samplingestimates: Sampling estimates. R package version 0.1-3.

[14] Escobar, E.L. and Barrios, E. Samplingvarest: Sampling variance estimation. R package version 0.9-9.

[15] Escobar, E.L. and Berger, Y.G. (2013). A jackknife variance estimator for self-weighted two-stage samples, *Statistica Sinica*, **23**, 595–613.

[16] Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion), *Statistical Science*, **22**(2), 153–164.

[17] Godambe, V.P. and Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation, *International Statistical Review/Revue Internationale de Statistique*, **54**(2), 127–138.

[18] Hardin, J.W. and Hilbe, J. (2001). *Generalized Linear Models and Extensions*, College Station, Texas: Stata Press.

[19] Henri, C. and Gavin, P. (2016). The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems, *Department of Civil and Environmental Engineering. Duke University*, `http://people.duke.edu/~hpgavin/ce281/lm.pdf`

[20] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**(377), 663–685.

[21] Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, **77**(377), 89–96.

[22] Kennel, T.L. and Valliant, R. (2010). Logistic generalized regression (LGREG) estimator in cluster samples, *Proceedings of the Section on Survey Research Methods*, 4756–4770.

[23] Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators, *Survey Methodology*, **24**(1), 51–55.

[24] Liu, X. and Koirala, H. (2013). Fitting proportional odds models to educational data with complex sampling designs in ordinal logistic regression, *Journal of Modern Applied Statistical Methods*, **12**(1), 235–248.

[25] Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variable*, Advanced Quantitative Techniques in the Social Sciences, Sage Publications, Thousand Oaks, CA.

[26] Mert, Y. (2015). Generalized least squares and weighted least squares estimation methods for distributional parameters, *REVSTAT – Statistical Journal*, **13**(3), 263–282.

[27] Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization, *The Computer Journal*, **7**, 308–313.

[28] Nordberg, L. (1989). Generalized linear modeling of sample survey data, *Journal of Official Statistics*, **5**, 223–239.

[29] Pavía, J.M. and Larraz, B. (2012). Nonresponse bias and superpopulation models in electoral polls, *Reis*, **137**, 121–150.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015, `https://www.R-project.org/`.

[31] Särndal, C.E. (2007). The calibration approach in survey theory and practice, *Survey Methodology*, **33**, 99–119.

[32] Särndal, C.E.; Swenson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

[33] Tillé, Y. and Matei, A. sampling: Survey sampling. r package version 2.6.

[34] Tukey, J.W. (1958). Bias and confidence in not-quite large samples, *The Annals of Mathematical Statistics*, **29**(2), 614.

[35] Valliant, R.; Dorfman, A. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley, New York.

[36] Winship, C. and Mare, R.D. (1984). Regression models with ordinal variables, *American Sociological Review [Internet]*.

[37] Wolter, K.M. (2007). *Introduction to Variance Estimation*, second ed., Springer.

[38] Wu, C. (2003). Optimal calibration estimators in survey sampling, *Biometrika*, **90**(4), 937–951.

[39] Wu, C. and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data, *Journal of the American Statistical Association*, **96**(453), 185–193.

REVSTAT – Statistical Journal
Volume 16, Number 3, July 2018, 349–363

RESIDUAL ANALYSIS WITH BIVARIATE INAR(1) MODELS

Authors: Predrag M. Popović
– Faculty of Civil Engineering and Architecture, University of Niš, Serbia
popovicpredrag@yahoo.com

Aleksandar S. Nastić
– Faculty of Sciences and Mathematics, University of Niš, Serbia
anastic78@gmail.com

Miroslav M. Ristić
– Faculty of Sciences and Mathematics, University of Niš, Serbia
miristic72@gmail.com

Received: April 2016 Revised: September 2016 Accepted: October 2016

Abstract:

- In this paper we analyze forecasting errors made by random coefficients bivariate integer-valued autoregressive models of order one. These models are based on the thinning operator to support discreteness of data. In order to achieve a comprehensive analysis, we introduce a model that implements a binomial as well as a negative binomial thinning operator. There are two components of the model: survival and innovation. Forecasting errors made by each of these two sources of uncertainty are unobservable in the classic way. Thus, we derive predictive distributions from which we obtain the expected value of each component of the model. We provide an example of residual analysis on real data.

Key-Words:

- *bivariate INAR(1) model; residual analysis; predictive distribution; binomial thinning; negative binomial thinning; geometric marginal distribution.*

AMS Subject Classification:

- 62M10.

1. INTRODUCTION

An integer-valued time series is a sequence of integer data points measured at uniform time intervals. Such series are present in many fields of sciences. For example, in medicine the number of infected persons represents such a series, in finance the number of defaults, in criminology the number of committed crimes, in biology the size of the population of a species, etc. Thus, modelling integer-valued time series and better understanding their nature is a point of interest for many researchers. Some of the first models for non-negative integer-valued time series were introduced by [7], [1] and [2]. These models have autoregressive structure, where the autoregression is achieved through the thinning operator. Models with the full autoregressive-moving average structure were investigated in [8]. Following these ideas, many models have been developed. A survey on integer-valued autoregressive processes can be found in [15].

While these are models with constant coefficients, [17] defined an integer-valued random coefficient model. Using the approach proposed by [3], [12] introduced a bivariate integer-valued random coefficient model. The dependence between processes that this model consists of is achieved through their autoregressive components, which are based on the negative binomial thinning operator. Some modifications of this model regarding the thinning operator and the marginal distribution are discussed in [9]. In this paper we focus on analyzing prediction errors made by these types of models. Since these models are composed of two components, survival and innovation, there are two sources of uncertainty. We try to estimate the portion of prediction error made by the survival and by the innovation component separately. Since these residuals are unobservable, we derive predictive distribution and calculate expected values of these components. Some aspects of predictive distributions for univariate models were presented in [13] and [14]. Residual analysis for univariate models was discussed in [5] and [16]. We extend the research on the bivariate models with random coefficients.

In addition, to cover two types of thinning operators, we introduce a bivariate model whose survival components are generated by different thinning operators, namely, binomial and negative binomial. This mix of thinning operators makes it possible to model two dependent processes whose survival parts have different properties. While the survival component generated by the negative binomial thinning operator does produce new members of the series, the other one generated by the binomial thinning does not and new members depends only on the innovation component. To motivate the model we consider two data series: monthly count of motor vehicle thefts and monthly count of drug dealing activities. The first series is characterized by the fact that offended persons are not provoked to commit the same criminal act, but the second series is to a large extent generated by itself since some amount of drugs has been resold many times.

The paper is organized as follows. In Section 2 we discuss the general form of bivariate integer-valued autoregressive models of order one with random coefficients. Section 3 introduces a bivariate model with both, binomial and negative binomial, thinning operators. We discuss residual analysis in Section 4. Real data modelling is considered in Section 5.

2. BIVARIATE INAR MODELS

In this section we state a general form of the random coefficient bivariate integer-valued autoregressive models of order one (BINAR(1)). In order to define a model suitable for the time series of count, we use the thinning operators. The binomial thinning operator is defined as $\alpha_1 \circ X = \sum_{i=1}^{X} B_i$, where X is a non-negative integer-valued random variable and $\{B_i\}$ is a sequence of i.i.d. Bernoulli random variables with mean parameter α_1. The negative binomial thinning operator is defined as $\alpha_1 * X = \sum_{i=1}^{X} G_i$, where $\{G_i\}$ is a sequence of i.i.d. geometric random variables with mean parameter α_1. For the time being we will not specify the thinning operator in the definition of BINAR(1) model. Let us denote a nonnegative bivariate time series of counts by $\boldsymbol{Z}_n$ and introduce a random matrix $\boldsymbol{A}_n = \begin{bmatrix} U_{1n} & U_{2n} \\ V_{1n} & V_{2n} \end{bmatrix}$, whose elements have the joint probability mass function defined as $P(U_{1n} = \alpha_1, U_{2n} = 0) = p = 1 - P(U_{1n} = 0, U_{2n} = \alpha_1)$ and $P(V_{1n} = \alpha_2, V_{2n} = 0) = q = 1 - P(V_{1n} = 0, V_{2n} = \alpha_2)$, where $\alpha_1, \alpha_2 \in (0, 1)$ and $p, q \in [0, 1]$. Then, the structure of BINAR(1) model is given by

$$\boldsymbol{Z}_n = \boldsymbol{A}_n \star \boldsymbol{Z}_{n-1} + \boldsymbol{e}_n, \quad n \geq 1, \tag{2.1}$$

where $\{\boldsymbol{e}_n\}$ represents the innovation process, which is composed of two independent series. The thinning operator is denoted with $\star$ and it acts as the matrix multiplication. The two processes that figure in $\boldsymbol{Z}_n$ are mutually dependent and their dependence is achieved trough autoregressive components, which are named survival processes. Coefficients that figure in (2.1) are random variables, which make this model significantly different from the similar multivariate INAR models (such as the one presented in [4] and [6]). Notice that

$$E(\boldsymbol{A}_n) = \boldsymbol{A} = \begin{bmatrix} \alpha_1 p & \alpha_1(1-p) \\ \alpha_2 q & \alpha_2(1-q) \end{bmatrix}.$$

It is easy to show that $E(\boldsymbol{A}_n \star \boldsymbol{Z}_n) = \boldsymbol{A}E(\boldsymbol{Z}_n)$. Following the discussion from [6], $\boldsymbol{I} - \boldsymbol{A}$ is a non singular matrix if all eigenvalues of $\boldsymbol{A}$ are inside the unit circle, which is proved for matrix $\boldsymbol{A}$ in [12]. All this implies that $E(\boldsymbol{Z}_n) = (\boldsymbol{I} - \boldsymbol{A})^{-1}E(\boldsymbol{e}_n)$. Since $(\boldsymbol{I} - \boldsymbol{A})^{-1}$ is a matrix of finite values, $E(\boldsymbol{Z}_0) < \infty$ iff $E(\boldsymbol{e}_1) < \infty$. The conditional expectation for process (2.1) is

$E(\boldsymbol{Z}_{n+k}|\boldsymbol{Z}_n) = \boldsymbol{A}^k\boldsymbol{Z}_n + (\boldsymbol{I}-\boldsymbol{A})^{-1}(\boldsymbol{I}-\boldsymbol{A}^k)(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{\mu}$, where $\boldsymbol{\mu} = E(\boldsymbol{Z}_n)$. The correlation structure of BINAR(1) model is given as $Cov(\boldsymbol{Z}_{n+k}, \boldsymbol{Z}_n) = \boldsymbol{A}^k Var(\boldsymbol{Z}_n)$, $k \geq 0$. Since the eigenvalues of matrix $\boldsymbol{A}$ are inside the unit circle, covariance tends to zero and conditional expectation tends to the unconditional one, as k tends to infinity. More details on the correlation structure can be found in [9].

3. MODEL WITH MIXED THINNING OPERATORS

In this section we introduce a new bivariate time series model $\{(X_{1,n},X_{2,n})\}$, $n \in \mathbb{N}_0$, where the two time series are dependent but evolve under different thinning operators. Let $\{X_{1,n}\}$ and $\{X_{2,n}\}$ be the two nonnegative integer-valued time series with probability mass function $P(X_{i,n} = k) = \mu^k/(1+\mu)^{k+1}$, $k \geq 0$, $\mu > 0$ and $i \in \{1, 2\}$. A mixed geometric bivariate autoregressive process of order one (BVMIXGINAR(1)) is given by the following equations

$$X_{1,n} = \begin{cases} \alpha_1 \circ X_{1,n-1} + \varepsilon_{1,n}, & \text{w.p. } p, \\ \alpha_1 \circ X_{2,n-1} + \varepsilon_{1,n}, & \text{w.p. } 1-p, \end{cases} \tag{3.1}$$

$$X_{2,n} = \begin{cases} \alpha_2 * X_{1,n-1} + \varepsilon_{2,n}, & \text{w.p. } q, \\ \alpha_2 * X_{2,n-1} + \varepsilon_{2,n}, & \text{w.p. } 1-q, \end{cases} \tag{3.2}$$

where $\{\varepsilon_{1,n}\}$ and $\{\varepsilon_{2,n}\}$ are i.i.d. sequences. The random vectors $(\varepsilon_{1,n}, \varepsilon_{2,n})$ and $(X_{1,m}, X_{2,m})$ are independent for all $m < n$. The thinning operators are defined in previous section and the counting series in $\alpha_1 \circ X_{1,n}$, $\alpha_1 \circ X_{2,n}$, $\alpha_2 * X_{1,n}$ and $\alpha_2 * X_{2,n}$ are mutually independent for all $n \in \mathbb{N}_0$ and are also independent of innovation processes $\{\varepsilon_{1,n}\}$ and $\{\varepsilon_{2,n}\}$. The distributions of the innovation processes are given by the following theorem.

Theorem 3.1. *Let $X_{1,0}$ and $X_{2,0}$ have the $Geom(\frac{\mu}{1+\mu})$ distribution, where $\mu > 0$. The stationary bivariate time series $\{(X_{1,n}, X_{2,n})\}_{n\in\mathbb{N}_0}$ given by equations* (3.1) *and* (3.2) *has $Geom(\frac{\mu}{1+\mu})$ marginal distributions if and only if the processes $\{\varepsilon_{1,n}\}$ and $\{\varepsilon_{2,n}\}$ are distributed as*

$$\varepsilon_{1,n} \overset{d}{=} \begin{cases} Geom(\frac{\mu}{1+\mu}), & \text{w.p. } 1-\alpha_1, \\ 0, & \text{w.p. } \alpha_1, \end{cases} \tag{3.3}$$

$$\varepsilon_{2,n} \overset{d}{=} \begin{cases} Geom(\frac{\mu}{1+\mu}), & \text{w.p. } \frac{\mu(1-\alpha_2)-\alpha_2}{\mu-\alpha_2}, \\ Geom(\frac{\alpha_2}{1+\alpha_2}), & \text{w.p. } \frac{\alpha_2\mu}{\mu-\alpha_2}, \end{cases} \tag{3.4}$$

where $\alpha_1 \in (0, 1)$, $\alpha_2 \in (0, \frac{\mu}{1+\mu}]$ and $p, q \in [0, 1]$.

Proof: Let us assume that the stationary time series $\{(X_{1,n}, X_{2,n})\}$ has the geometric marginal distribution $Geom(\frac{\mu}{1+\mu})$, $\mu > 0$. Since the random variables $X_{1,n}$ and $X_{2,n}$ are equal in distribution, considering probability generating

functions we obtain $\Phi_{X_{1,n}}(s) = \Phi_{\varepsilon_{1,n}}(s)\Phi_{X_{1,n-1}}(\Phi_{B_i}(s))$, which follows from (3.1). From their geometric distribution we obtain

$$\Phi_{\varepsilon_{1,n}}(s) = \frac{1+\mu\alpha_1(1-s)}{1+\mu(1-s)} = \alpha_1 + (1-\alpha_1)\frac{1}{1+\mu-\mu s},$$

which proves equation (3.3). In a similar manner we derive the probability generating function for $\varepsilon_{2,n}$ and obtain

$$\begin{aligned}\Phi_{\varepsilon_{2,n}}(s) &= \frac{(1+\mu)(1+\alpha_2-\alpha_2 s)-\mu}{(1+\mu-\mu s)(1+\alpha_2-\alpha_2 s)} \\ &= \frac{\mu(1-\alpha_2)-\alpha_2}{\mu-\alpha_2}\frac{1}{a+\mu-\mu s} + \frac{\alpha_2\mu}{\mu-\alpha_2}\frac{1}{1+\alpha_2-\alpha_2 s}.\end{aligned}$$

Now, equation (3.4) follows under the constraints given in [11] for NGINAR(1) model.

Conversely, let us assume that the distributions of the random variables $\varepsilon_{1,n}$ and $\varepsilon_{2,n}$ are given by equations (3.3) and (3.4), respectively. Since $X_{1,0} \stackrel{d}{=} X_{2,0} \stackrel{d}{=} Geom(\mu/(1+\mu))$, we obtain

$$\begin{aligned}\Phi_{X_{1,1}}(s) &= \Phi_{X_{1,0}}(1-\alpha_1+\alpha_1 s)\Phi_{\varepsilon_{1,1}}(s) \\ &= \frac{1}{1+\mu-\mu(1-\alpha_1+\alpha_1 s)}\frac{1+\mu\alpha_1(1-s)}{1+\mu(1-s)} = \frac{1}{1+\mu-\mu s}\end{aligned}$$

and

$$\begin{aligned}\Phi_{X_{2,1}}(s) &= \Phi_{X_{2,0}}\left(\frac{1}{1+\alpha_2-\alpha_2 s}\right)\Phi_{\varepsilon_{2,1}}(s) \\ &= \frac{1}{1+\mu(1-\frac{1}{1+\alpha_2-\alpha_2 s})}\cdot\frac{(1+\mu)(1+\alpha_2-\alpha_2 s)-\mu}{(1+\mu-\mu s)(1+\alpha_2-\alpha_2 s)} = \frac{1}{1+\mu-\mu s}.\end{aligned}$$

Thus, $X_{1,1}$ and $X_{2,1}$ have geometric distribution with parameter $\mu/(1+\mu)$. Using mathematical induction we can prove that $X_{1,n} \stackrel{d}{=} X_{2,n} \stackrel{d}{=} Geom(\mu/(1+\mu))$ for any $n \in \mathbb{N}_0$. □

Even if we assume that $X_{1,0}$ and $X_{2,0}$ have the same arbitrary distribution, $X_{1,n}$ as well as $X_{2,n}$ converges to geometric distribution $Geom(\mu/(1+\mu))$, as $n \to \infty$, if random variables $\varepsilon_{1,n}$ and $\varepsilon_{2,n}$ have the distribution given by Theorem 3.1. This can be proved with the following two equations. The first equation is

$$\begin{aligned}\Phi_{X_{1,n}}(s) &= \Phi_{X_{1,n-1}}(1-\alpha_1+\alpha_1 s)\Phi_{\varepsilon_{1,n}}(s) \\ &= \Phi_{X_{1,0}}(1-\alpha_1^n+\alpha_1^n s)\prod_{k=0}^{n-1}\frac{1+\mu\alpha_1^{k+1}(1-s)}{1+\mu\alpha_1^k(1-s)} \\ &= \Phi_{X_{1,0}}(1-\alpha_1^n+\alpha_1^n s)\frac{1+\mu\alpha_1^n(1-s)}{1+\mu(1-s)} \xrightarrow{n\to\infty} \frac{1}{1+\mu-\mu s}.\end{aligned}$$

Also,

$$\begin{aligned}
\Phi_{X_{2,n}}(s) &= \Phi_{X_{2,n-1}}\left(\frac{1}{1+\alpha_2-\alpha_2 s}\right)\Phi_{\varepsilon_{2,n}}(s)\\
&= \Phi_{X_{2,0}}\left(\frac{1-\alpha_2+\alpha_2(1-s)(1-\alpha_2^{n-1})}{1-\alpha_2+\alpha_2(1-s)(1-\alpha_2^{n})}\right)\prod_{k=0}^{n-1}\Phi_{\varepsilon_{2,k}}\left(\frac{1-\alpha_2+\alpha_2(1-s)(1-\alpha_2^{k-1})}{1-\alpha_2+\alpha_2(1-s)(1-\alpha_2^{k})}\right)\\
&= \Phi_{X_{2,0}}\left(\frac{1-\alpha_2+\alpha_2(1-s)(1-\alpha_2^{n-1})}{1-\alpha_2+\alpha_2(1-s)(1-\alpha_2^{n})}\right)\\
&\quad\times\frac{(1-\alpha_2^2(1-s)-\alpha_2 s)(1+\alpha_2^n\mu(1-s)-\alpha_2^{n+1}(1+\mu)(1-s)-\alpha_2 s)}{(1-\alpha_2^{n+1}(1-s)-\alpha_2 s)(1+\alpha_2\mu(1-s)-\alpha_2^2(1+\mu)(1-s)-\alpha_2 s)}\\
&\xrightarrow{n\to\infty}\frac{1}{1+\mu-\mu s}.
\end{aligned}$$

Random variables $X_{1,n}$ and $X_{2,n}$ are independent for known $X_{1,n-1}$ and $X_{2,n-1}$. Thus, the conditional distribution of $(X_{1,n}, X_{2,n})$, given $(X_{1,n-1}, X_{2,n-1})$, is defined as

$$\begin{aligned}
&P(X_{1,n}=x, X_{2,n}=y|X_{1,n-1}=u, X_{2,n-1}=v)\\
&=P(X_{1,n}=x|X_{1,n-1}=u, X_{2,n-1}=v)P(X_{2,n}=y|X_{1,n-1}=u, X_{2,n-1}=v).
\end{aligned}$$

The conditional probability mass function of the random variable $X_{1,n}$ for given $X_{1,n-1}$ and $X_{2,n-1}$ has the form

$$\begin{aligned}
&P(X_{1,n}=x|X_{1,n-1}=u, X_{2,n-1}=v)\\
&=p\sum_{k=0}^{\min(x,u)}P(\varepsilon_{1,n}=x-k)P(\alpha_1\circ X_{1,n-1}=k|X_{1,n-1}=u)\\
&+(1-p)\sum_{k=0}^{\min(x,v)}P(\varepsilon_{1,n}=x-k)P(\alpha_1\circ X_{2,n-1}=k|X_{2,n-1}=v). \qquad (3.5)
\end{aligned}$$

Similarly, for $X_{2,n}$ the form is

$$\begin{aligned}
&P(X_{2,n}=y|X_{1,n-1}=u, X_{2,n-1}=v)\\
&=q\sum_{k=0}^{y}P(\varepsilon_{2,n}=y-k)P(\alpha_2*X_{1,n-1}=k|X_{1,n-1}=u)\\
&+(1-q)\sum_{k=0}^{y}P(\varepsilon_{2,n}=y-k)P(\alpha_2*X_{2,n-1}=k|X_{2,n-1}=v). \qquad (3.6)
\end{aligned}$$

The random variables $\alpha_1\circ X$ and $\alpha_2 * X$ under the condition $X=u$ have binomial and negative binomial distribution with parameters (u,α_1) and $(u, \frac{\alpha_2}{1+\alpha_2})$, respectively (where the probability mass function for negative binomial distribution is taken as $P(\alpha_2 * X = k|X=u) = \binom{u+k-1}{k}\frac{\alpha_2^k}{(1+\alpha_2)^{k+u}}$). Notice that the probability mass functions for the innovation processes are, respectively,

$$P(\varepsilon_{1,n}=x-k)=1_{\{x=k\}}\alpha_1+(1-\alpha_1)\frac{\mu^{x-k}}{(1+\mu)^{x-k+1}},$$

$$P(\varepsilon_{2,n} = y - k) = \frac{\mu(1-\alpha_2) - \alpha_2}{\mu - \alpha_2} \frac{\mu^{y-k}}{(1+\mu)^{y-k+1}} + \frac{\alpha_2 \mu}{\mu - \alpha_2} \frac{\alpha_2^{y-k}}{(1+\alpha_2)^{y-k+1}},$$

where 1_A is the indicator function of a random event A.

The estimation of unknown parameters of the bivariate INAR(1) models with random coefficients is discussed in details in [9]. We consider the conditional maximum likelihood method for parameters estimation of the model presented in this paper.

For the given values $\{(X_{1,k}, X_{2,k})\}_{k=\overline{0,n}}$, we set the conditional likelihood function as

$$L_1(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln P(X_{1,i} = x_{1,i}, X_{2,i} = x_{2,i} | X_{1,i-1} = x_{1,i-1}, X_{2,i-1} = x_{2,i-1}, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\alpha_1, \alpha_2, p, q, \mu)$ is a vector of unknown parameters. The probability mass function is defined as a product of functions (3.5) and (3.6). The maximization of the log-likelihood function is obtained by numerical procedure, which, in our case, is conducted through the programming language `R` and the function `nlm`.

4. RESIDUALS

The standard statistic used for determining a goodness of fit is obtained by summing squared residuals. The residuals are obtained as a difference between a value at time n and an expected value of the process in time n for the given value at $n-1$, i.e.,

$$r_{X_1,n} = X_{1,n} - \alpha_1 p X_{1,n-1} - \alpha_1(1-p) X_{2,n-1} - \mu_{\varepsilon_1},$$
$$r_{X_2,n} = X_{2,n} - \alpha_2 q X_{1,n-1} - \alpha_2(1-q) X_{2,n-1} - \mu_{\varepsilon_2},$$

where μ_{ε_i} are the expected values of the random variables ε_i, $i \in \{1, 2\}$. Since our process is composed of two sources of uncertainty (survival process and innovation process) it would be useful to track residuals of each source separately. This idea for one-dimensional INAR process is presented in [5] and [16], while we extend it for the bivariate case where the coefficients of the model are random variables. The residual analysis for a bivariate model is also investigated in [10], but the model has constant coefficients, independent survival processes and dependent innovation processes, which makes it significantly different from BVMIXGINAR(1) model.

If we introduce the two pairs of random variables (U_{1n}, U_{2n}) and (V_{1n}, V_{2n})

defined in Section 2, we can present BVMIXGINAR(1) model as

$$X_{1,n} = U_{1n} \circ X_{1,n-1} + U_{2n} \circ X_{2,n-1} + \varepsilon_{1,n}, \tag{4.1}$$

$$X_{2,n} = V_{1n} * X_{1,n-1} + V_{2n} * X_{2,n-1} + \varepsilon_{2,n}. \tag{4.2}$$

The process in this form is more tractable in terms of survival and innovation components. Therefore, we get two sets of residuals: $r^{sur}_{X_{1,n}} = U_{1n} \circ X_{1,n-1} + U_{2n} \circ X_{2,n-1} - \alpha_1 p X_{1,n-1} - \alpha_1(1-p)X_{2,n-1}$ and $r^{in}_{X_{1,n}} = \varepsilon_{1,n} - \mu_{\varepsilon_1}$ (analogous for the process $\{X_{2,n}\}$). The problem that arises here is that the binomial thinning component and the innovation component are not observable. Thus, we have to consider their conditional expectation with respect to the σ-algebra generated by vectors $(X_{1,n}, X_{2,n}), (X_{1,n-1}, X_{2,n-1}), ..., (X_{1,0}, X_{2,0})$. Since the process $\{(X_{1,n}, X_{2,n})\}$ is lag-one dependent, we investigate conditional expectations with respect to the σ-algebra generated only with random vectors at moments n and $n-1$, $\mathcal{F}_n(X_{1,n}, X_{2,n}, X_{1,n-1}, X_{2,n-1})$.

Let us introduce the notations $P_{u,v}(A) = P(A|X_{1,n-1} = u, X_{2,n-1} = v)$ and $P_{x_1,x_2,u,v}(A) = P(A|X_{1,n} = x_1, X_{2,n} = x_2, X_{1,n-1} = u, X_{2,n-1} = v)$. The conditional probability mass function of the first addend in equation (4.1) with respect to the σ-algebra $\mathcal{F}_n$, for $m, x, y, u, v \in \mathbb{N}_0$, is

$$\begin{aligned}
&P_{x_1,x_2,u,v}(U_{1n} \circ X_{1,n-1} = m) \\
&= \frac{P_{u,v}(U_{1n} \circ X_{1,n-1} = m, U_{1n} \circ X_{1,n-1} + U_{2n} \circ X_{2,n-1} + \varepsilon_{1,n} = x_1)}{P_{u,v}(X_{1,n} = x_1)} \\
&= \frac{1}{P_{u,v}(X_{1,n} = x_1)}[pP_{u,v}(\alpha_1 \circ X_{1,n-1} = m, 0 \circ X_{2,n-1} + \varepsilon_{1,n} = x_1 - m) \\
&\qquad + (1-p)P_{u,v}(0 \circ X_{1,n-1} = m, \alpha_1 \circ X_{2,n-1} + \varepsilon_{1,n} = x_1 - m)] \\
&= \frac{1}{P_{u,v}(X_{1,n} = x_1)}[pP(Bin(u, \alpha_1) = m)P(\varepsilon_{1,n} = x_1 - m) \\
&\qquad + (1-p)1_{\{m=0\}}P(Bin(v, \alpha_1) + \varepsilon_{1,n} = x_1 - m)],
\end{aligned}$$

where $Bin(u, \alpha)$ denotes a random variable with binomial distribution and parameters u i α. In a similar manner, for $r, k, s \in \mathbb{N}_0$, we obtain the following three equations,

$$\begin{aligned}
P_{x_1,x_2,u,v}(U_{2n} \circ X_{2,n-1} = r) = &\frac{1}{P_{u,v}(X_{1,n} = x_1)}[pI_{\{r=0\}}P(Bin(u, \alpha_1) + \varepsilon_{1,n} = x_1) \\
&+ (1-p)P(Bin(v, \alpha_2) = r)P(\varepsilon_{1,n} = x_1 - r)],
\end{aligned}$$

$$\begin{aligned}
P_{x_1,x_2,u,v}(V_{1n} * X_{1,n-1} = k) = &\frac{1}{P_{u,v}(X_{2,n} = x_2)}[qP(NB(u, \alpha_2) = k)P(\varepsilon_{2,n} = x_2 - k) \\
&+ (1-q)I_{\{k=0\}}P(NB(v, \alpha_2) + \varepsilon_{2,n} = x_2)],
\end{aligned}$$

$$\begin{aligned}
P_{x_1,x_2,u,v}(V_{2n} * X_{2,n-1} = s) = &\frac{1}{P_{u,v}(X_{2,n} = x_2)}[qI_{\{s=0\}}P(NB(u, \alpha_2) + \varepsilon_{2,n} = x_2) \\
&+ (1-q)P(NB(v, \alpha_2) = s)P(\varepsilon_{2,n} = y - s)].
\end{aligned}$$

In the last two equations the notation $NB(u,\alpha)$ stands for a random variable with negative binomial distribution with parameters u and $\frac{\alpha}{1+\alpha}$.

With these results in mind and applying some algebra, we obtain the following equations

$$\begin{aligned}
&E(U_{in}\circ X_{i,n-1}|\mathcal{F}_n)=\sum_{j=1}^{u_i} jP_{x_1,x_2,u,v}(U_{in}\circ X_{i,n-1}=j)\\
&=\frac{p_i}{P_{u,v}(X_{1,n}=x_1)}\sum_{j=1}^{\min(u_i,x_1)} j\binom{u_i}{j}\alpha_1^j(1-\alpha_1)^{u_i-j}P(\varepsilon_{1,n}=x_1-j)\\
&=\frac{\alpha_1 p_i u_i}{P_{u,v}(X_{1,n}=x_1)}\sum_{j=0}^{\min(u_i-1,x_1-1)}\binom{u_i-1}{j}\alpha_1^j(1-\alpha_1)^{u_i-1-j}P(\varepsilon_{1,n}=x_1-1-j)\\
&=\frac{\alpha_1 p_i u_i}{P_{u,v}(X_{1,n}=x_1)}P'_{i,u_i-1}(\alpha_1\circ X_{i,n-1}+\varepsilon_{1,n}=x_1-1)
\end{aligned}$$

and

$$\begin{aligned}
&E(V_{in}* X_{i,n-1}|\mathcal{F}_n)=\sum_{j=1}^{x_2} jP_{x_1,x_2,u,v}(V_{in}* X_{i,n-1}=j)\\
&=\frac{q_i}{P_{u,v}(X_{2,n}=x_2)}\sum_{j=1}^{x_2} j\binom{u_i+j-1}{j}\frac{\alpha_2^j}{(1+\alpha_2)^{u_i+j}}P(\varepsilon_{2,n}=x_2-j)\\
&=\frac{\alpha_2 q_i u_i}{P_{u,v}(X_{2,n}=x_2)}\sum_{j=0}^{x_2-1}\binom{u_i+1+j-1}{j}\frac{\alpha_2^j}{(1+\alpha_2)^{u_i+1+j}}P(\varepsilon_{2,n}=x_2-1-j)\\
&=\frac{\alpha_2 q_i u_i}{P_{u,v}(X_{2,n}=x_2)}P'_{i,u_i+1}(\alpha_2* X_{i,n-1}+\varepsilon_{2,n}=x_2-1),
\end{aligned}$$

where we introduced the notations $P'_{i,x}(A)=P(A|X_{i,n-1}=x)$, $i=1,2$, $p_1=p$, $p_2=1-p$, $q_1=q$, $q_2=1-q$, $u_1=u$ and $u_2=v$. Thus, we can conclude that the conditional expectation of the survival part of the process (4.1) is calculated as

$$\begin{aligned}
&E(U_{1n}\circ X_{1,n-1}+U_{2n}\circ X_{2,n-1}|\mathcal{F}_n)=pE(\alpha_1\circ X_{1,n-1}|\mathcal{F}_n)+(1-p)E(\alpha_1\circ X_{2,n-1}|\mathcal{F}_n)\\
&=\frac{1}{P_{u,v}(X_{1,n}=x_1)}\cdot[\alpha_1 puP_{1,u-1}(\alpha_1\circ X_{1,n-1}+\varepsilon_{1,n}=x_1-1)\\
&\quad+\alpha_1(1-p)vP_{2,v-1}(\alpha_1\circ X_{2,n-1}+\varepsilon_{1,n}=x_1-1)]
\end{aligned}$$

and for the process (4.2) as

$$\begin{aligned}
&E(V_{1n}* X_{1,n-1}+V_{2n}* X_{2,n-1}|\mathcal{F}_n)=qE(\alpha_2* X_{1,n-1}|\mathcal{F}_n)+(1-q)E(\alpha_2* X_{2,n-1}|\mathcal{F}_n)\\
&=\frac{1}{P_{u,v}(X_{2,n}=x_2)}\cdot[qu\alpha_2P_{1,u+1}(\alpha_2* X_{1,n-1}+\varepsilon_{2,n}=x_2-1)\\
&\quad+(1-q)v\alpha_2P_{2,v+1}(\alpha_2* X_{2,n-1}+\varepsilon_{2,n}=x_2-1)]\,.
\end{aligned}$$

We have defined the innovation processes such that $(\varepsilon_{1,n}, \varepsilon_{2,n})$ is independent of $(X_{1,m}, X_{2,m})$ for $m < n$. Since we observed the conditional expectation at time n with respect to the σ-algebra $\mathcal{F}_n$, we need to pay special attention here. The conditional probability mass functions are

$$\begin{aligned} P_{x_1,x_2,u,v}(\varepsilon_{1,n} = x_1 - k) = & \frac{P(\varepsilon_{1,n} = x_1 - k)}{P_{u,v}(X_{1,n} = x_1)} \\ & \times [pP_{u,v}(\alpha_1 \circ X_{1,n-1} = k) + (1-p)P_{u,v}(\alpha_1 \circ X_{2,n-1} = k)] \end{aligned}$$

and

$$\begin{aligned} P_{x_1,x_2,u,v}(\varepsilon_{2,n} = x_2 - k) = & \frac{P(\varepsilon_{2,n} = x_2 - k)}{P_{u,v}(X_{2,n} = x_2)} \\ & \times [qP_{u,v}(\alpha_2 * X_{1,n-1} = k) + (1-q)P_{u,v}(\alpha_2 * X_{2,n-1} = k)] . \end{aligned}$$

Hance, the corresponding conditional expectations are

$$\begin{aligned} E(\varepsilon_{1,n}|\mathcal{F}_n) = & \frac{1}{P_{u,v}(X_{1,n} = x_1)} \\ & \times [p\,(x_1 P_{1,u}(\alpha_1 \circ X_{1,n-1} + \varepsilon_{1,n} = x_1) - \alpha_1 u P_{1,u-1}(\alpha_1 \circ X_{1,n-1} + \varepsilon_{1,n} = x_1 - 1)) \\ & + (1-p)\,(x_1 P_{2,v}(\alpha_1 \circ X_{2,n-1} + \varepsilon_{1,n} = x_1) - \alpha_1 v P_{2,v-1}(\alpha_1 \circ X_{2,n-1} + \varepsilon_{1,n} = x_1 - 1))] \end{aligned}$$

and

$$\begin{aligned} E(\varepsilon_{2,n}|\mathcal{F}_n) = & \frac{1}{P_{u,v}(X_{2,n} = x_2)} \\ & \times [q\,(x_2 P_{1,u}(\alpha_2 * X_{1,n-1} + \varepsilon_{2,n} = x_2) - \alpha_2 u P_{1,u+1}(\alpha_2 * X_{1,n-1} + \varepsilon_{2,n} = x_2 - 1)) \\ & + (1-q)\,(x_2 P_{2,v}(\alpha_2 * X_{2,n-1} + \varepsilon_{2,n} = x_2) - \alpha_2 v P_{2,v+1}(\alpha_2 * X_{2,n-1} + \varepsilon_{2,n} = x_2 - 1))] . \end{aligned}$$

Now we can distinguish between the error from the survival and the error from the innovation process. If we sum these two values we obtain the following results

$$\begin{aligned} r^{surr}_{X_1,n} + r^{in}_{X_1,n} = & E(p\alpha_1 \circ X_{1,n-1} + (1-p)\alpha_1 \circ X_{2,n-1}|X_{1,n}, X_{1,n-1}) \\ & - \alpha_1 p X_{1,n-1} - \alpha_1(1-p)X_{2,n-1} + E(\varepsilon_{1,n}|X_{1,n}, X_{1,n-1}) - \mu_\varepsilon \\ = & E(p\alpha_1 \circ X_{1,n-1} + (1-p)\alpha_1 \circ X_{2,n-1} + \varepsilon_{1,n}|X_{1,n}, X_{1,n-1}) \\ & - \alpha_1 p X_{1,n-1} - \alpha_1(1-p)X_{2,n-1} - \mu_{\varepsilon_{1,n}} \\ = & X_{1,n} - \alpha_1 p X_{1,n-1} - \alpha_1(1-p)X_{2,n-1} - \mu_{\varepsilon_1} = r_{X_1,n}. \end{aligned} \tag{4.3}$$

We can conclude that the sum of these two error terms is equal to the error term obtained by using conditional expectation for the process $\{X_{1,n}\}$ with respect to σ-algebra $\mathcal{F}_{n-1}$. The conclusion is analogous for the process $\{X_{2,n}\}$.

5. APPLICATION

In this section, we discuss the characteristics of data for which BVMIXGINAR(1) model is the most adequate. We compare results of BVMIXGINAR(1)

model with results of some other bivariate models. At the end of the section we analyze prediction errors of BVMIXGINAR(1) model and suggest how the model can be improved.

We analyze data from the Pittsburgh police department number 407, which can be found on the website www.forecastingprinciples, where we focus on the number of stolen vehicles (MVTHEFT) and the number of reported drug activities (C_DRUG) per month from January 1990 to December 2001. The average values of these two series are 1.74 and 1.5, and the variances are 2.98 and 5.01, respectively. The correlation between the series is 0.22. The bar plots and correlograms are given in Figure 1. Both corelograms show the presence of lag 1 autocorrelation. Although there are some autocorrelations on higher lags for series C_DRUG, the value on the first lag is dominant. High positive correlation between the series, overdispersion and the first lag autocorrelation imply that BVMIXGINAR(1) might be adequate.

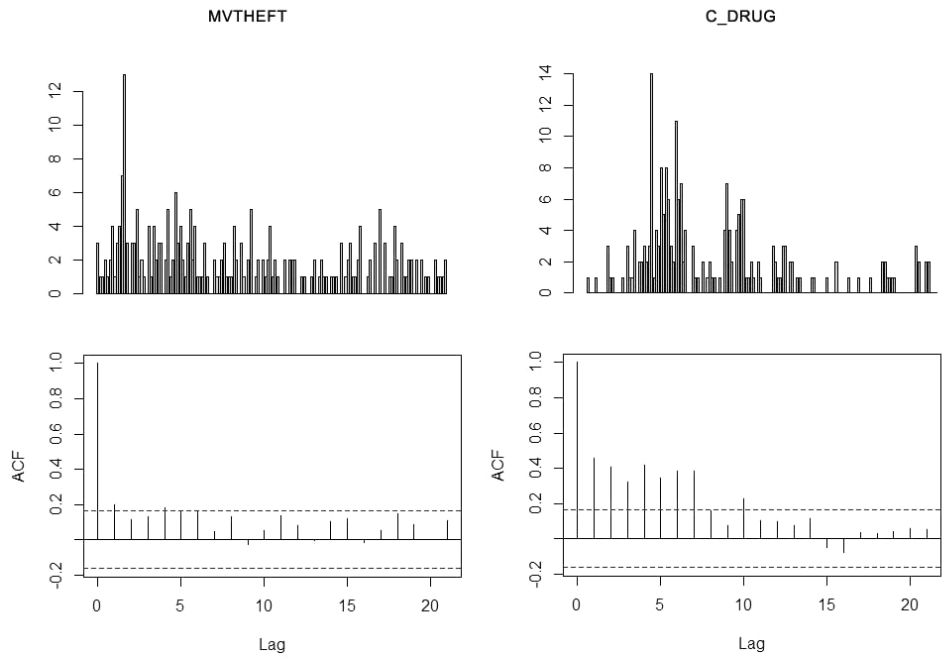

Figure 1: Data series and autocorrelation function for MVTHEFT and C_DRUG series.

We compare BVMIXGINAR(1) model with models BVNGINAR(1) introduced in [12], and BVPOINAR(1) introduced in [9], since both of these models have random coefficients and a similar structure. For the three models, we compare their values of the log-likelihood functions and the root mean square errors (RMS) made by one step ahead prediction. The results are presented in Table 1.

Table 1: Parameter estimates of INAR models, root mean square errors and the log-likelihood function for MVTHEFT and C_DRUG data series.

Model	CML estimates	log-likelihood	RMS MVTHEFT	RMS C_DRUG
BVMIXGINAR(1)	$\hat{\alpha_1} = 0.148(0.1)$, $\hat{\alpha_2} = 0.472(0.095)$, $\hat{p} = 0.469(0.325)$, $\hat{q} = 0.02(0.08)$, $\hat{\mu} = 1.598(0.164)$	-477.51	1.695	1.990
BVNGINAR(1)	$\hat{\alpha_1} = 0.242(0.166)$, $\hat{\alpha_2} = 0.473(0.095)$, $\hat{p} = 0.39(0.244)$, $\hat{q} = 0.02(0.078)$, $\hat{\mu} = 1.604(0.171)$	-478.5	1.703	1.990
BVPOINAR(1)	$\hat{\alpha_1} = 0.217(0.066)$, $\hat{\alpha_2} = 0.41(0.055)$, $\hat{p} = 0.328(0.172)$, $\hat{q} = 0.068(0.066)$, $\hat{\lambda} = 1.687(0.11)$	-516.62	1.701	1.998

According to the test results, BVMIXGINAR(1) is the most adequate for these data. Notice that models with geometric distribution obtain higher values of the likelihood function. BVMIXGINAR(1) achieves slightly higher log-likelihood values than BVNGINAR(1) but much lower RMS for MVTHEFT series. Modelling C_DRUG series with geometric distribution where survival processes evolve under negative binomial thinning provides the best results. RMS for C_DRUG are the same for BVMIXGINAR(1) and BVNGINAR(1). The improvement with BVMIXGINAR(1) is with RMS for MVTHEFT. The assumption that one survival process evolves under binomial and the other survival process under negative binomial thinning improves prediction performance. We need this mix of thinning operators when we model two series with different behavior, as the case here. Since once sold drugs are often resold, but once stolen vehicle cannot be stolen again, we have here one process that is self-generated and one that is not.

The estimated parameters of BVMIXGINAR(1) indicate that drug activities influence the number of stolen vehicles in this area, while vice versa does not hold since the value of parameter q is statistically equal to zero.

We continue with the prediction performance analysis by focusing on the prediction errors made by the survival and innovation components separately. We calculate these residuals and plot them to assess the adequacy of each component. As given by equation (4.3), the sum of these two residuals is equal to the residuals obtained by the usual definition. The residuals are presented in Figure 2. It can be noticed that the residuals of the innovation processes are much higher than the residuals of the survival processes, apart from the few cases of C_DRUG series. Further, the correlation between the two type of residuals is 0.425 for MVTHEFT and 0.506 for C_DRUG series. The correlation is positive but not as high as one might expect. These results also add value to the model since an imprecise prediction of one component can be absorbed by the prediction of the other component. Another interesting point is a low correlation of only 0.11 between the innovation processes of the two series, which supports the structural assumption that innovation processes are independent. Higher residuals generated by the

innovation processes indicate that future work should focus on improving the innovation processes.

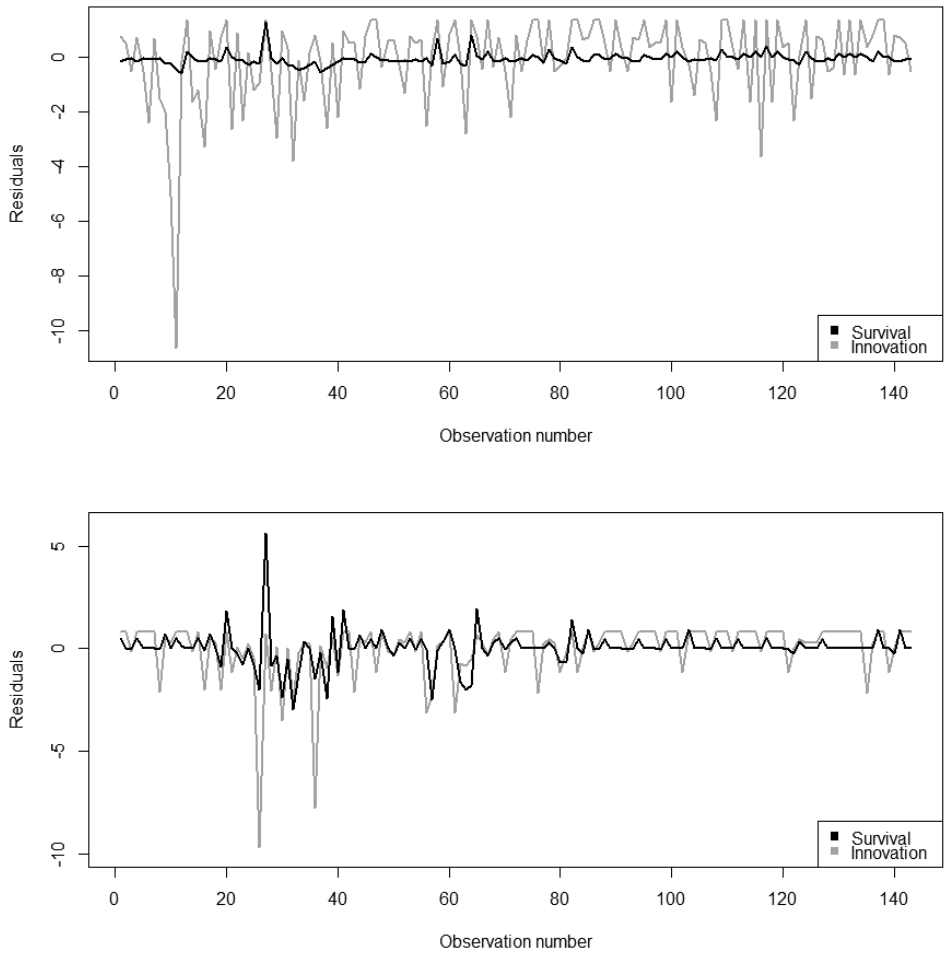

Figure 2: Upper figure shows the residuals for MVTHEFT and lower for C_DRUG series.

REFERENCES

[1] Al-Osh, M.A. and Alzaid, A.A. (1987). First-order Integer-valued Autoregressive (INAR(1)) Process, *Journal of Time Series Analysis*, **8**, 261–275.

[2] Al-Osh, M.A. and Aly, A.A.E. (1992). First order autoregressive time series with negative binomial and geometric marginals, *Communications in Statistics – Theory and Methods*, **21**, 2483–2492.

[3] Dewald, S.L.; Lewis, P.A.W. and McKenzie, E. (1989). A bivariate First-Order Autoregressive Time Series Modiel in Exponential Variables (BEAR(1)), *Management Science*, **35**, 1236–1246.

[4] Franke, J. and Rao, T.S. (1993). *Multivariate first-order integer-valued autoregressions*, Technical Report, Mathematics Department, UMIST.

[5] Freeland, R.K. and McCabe, B.P.M. (2004). Analysis of low count time series data by poisson autoregression, *Journal of Time Series Analysis*, **25**, 701–722.

[6] Latour, A. (1997). The Multivariate GINAR(p) Process, *Advances in Applied Probability*, **29**, 228–248.

[7] McKenzie, E. (1985). Some simple models for discrete variate time series, *Journal of the American Water Resources Association*, **21**, 645–650.

[8] McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts, *Advances in Applied Probability*, **20**, 822–835.

[9] Nastić, A.S.; Ristić, M.M. and Popović, P.M. (2016). Estimation in a Bivariate Integer-Valued Autoregressive Process, *Communications in Statistics – Theory and Methods*, **45**, 5660–5678.

[10] Pedeli, X. and Karlis, D. (2011). A Bivariate INAR(1) Process with Application, *Statistical Modelling*, **11**, 325–349.

[11] Ristić, M.M.; Nastić, A.S. and Bakouch, H.S. (2009). A new geometric first-order integer-valued autoregressive (NGINAR(1)) process, *Journal of Statistical Planning and Inference*, **139**, 2218–2226.

[12] Ristić, M.M.; Nastić, A.S.; Jayakumar, K. and Bakouch, H.S. (2012). A bivariate INAR(1) time series model with geometric marginals, *Applied Mathematics Letters*, **25**, 481–485.

[13] Silva, N.; Pereira, I. and Silva, E.M. (2008). Estimation and forecasting in SUINAR (1) model, *REVSTAT*, **6**, 253–277.

[14] Silva, N.; Pereira, I. and Silva, E.M. (2009). Forecasting in INAR (1) model, *REVSTAT*, **7**, 119–134.

[15] Weiss, H.C. (2008). Thinning operations for modeling time series of counts–a survey, *Advances in Statistical Analysis*, **92**, 319–341.

[16] Park, Y. and Kim, H.Y. (2012). Diagnostic checks for integer-valued autoregressive models using expected residuals, *Statistical Papers*, **53**, 951–971.

[17] Zheng, H.; Basawa, I.V. and Datta, S. (2005). Inference for p^{th}-order random coefficient integer-valued autoregressive processes, *Journal Of Time Series Analysis*, **27**, 411–440.

REVSTAT – Statistical Journal
Volume 16, Number 3, July 2018, 365–403

LIKELIHOOD RATIO TEST FOR THE HYPER-BLOCK MATRIX SPHERICITY COVARIANCE STRUCTURE — CHARACTERIZATION OF THE EXACT DISTRIBUTION AND DEVELOPMENT OF NEAR-EXACT DISTRIBUTIONS FOR THE TEST STATISTIC

Authors: Bárbara R. Correia
– CMA/UNL – Centro de Matemática e Aplicações,
Universidade Nova de Lisboa, Portugal
regadas.correia@gmail.com

Carlos A. Coelho
– Departamento de Matemática,
CMA/UNL – Centro de Matemática e Aplicações,
Universidade Nova de Lisboa, Portugal
cmac@fct.unl.pt

Filipe J. Marques
– Departamento de Matemática,
CMA/UNL – Centro de Matemática e Aplicações,
Universidade Nova de Lisboa, Portugal
fjm@fct.unl.pt

Received: March 2016 Revised: July 2016 Accepted: September 2016

Abstract:

- In this paper the authors introduce the hyper-block matrix sphericity test which is a generalization of both the block-matrix and the block-scalar sphericity tests and as such also of the common sphericity test. This test is a tool of crucial importance to verify elaborate assumptions on covariance matrix structures, namely on meta-analysis and error covariance structures in mixed models and models for longitudinal data. The authors show how by adequately decomposing the null hypothesis of the hyper-block matrix sphericity test it is possible to easily obtain the expression for the likelihood ratio test statistic as well as the expression for its moments. From the factorization of the exact characteristic function of the logarithm of the statistic, induced by the decomposition of the null hypothesis, and by adequately replacing some of the factors with an asymptotic result, it is possible to obtain near-exact distributions that lie very close to the exact distribution. The performance of these near-exact distributions is assessed through the use of a measure of proximity between distributions, based on the corresponding characteristic functions.

Key-Words:

- *equality of matrices test; Generalized Integer Gamma distribution; Generalized Near-Integer Gamma distribution; independence test; near-exact distributions; mixtures of distributions.*

AMS Subject Classification:

- 62H15, 62H10, 62E10, 62E15, 62E20.

1. INTRODUCTION

Likelihood ratio tests (l.r.t.'s) have a large scope of application in different fields of research such as for example engineering, economics, medicine and ecology [27, 20, 5, 21]. However, in most cases, the exact distribution of the l.r.t. statistics has a very complicated expression which makes difficult the practical use of the testing procedure. On the other hand the commonly used asymptotic approximations [2, 26] display lack of precision mainly in extreme situations such as for high number of variables and/or small sample sizes [14, 9] and situations where the parameters of interest and/or nuisance parameters are on the boundary of the parameter space [11]. This is a well known and recognized problem in standard likelihood ratio testing procedures which becomes even more serious when one wants to perform tests for more elaborate covariance structures. These elaborate structures have recently become very important in different statistical techniques for the validation of assumptions required in different models such as in hierarchical or mixed linear univariate and multivariate models.

In this paper the authors introduce the hyper-block matrix (HBM) sphericity test. This test is a useful generalization of the block-matrix and of the block-scalar sphericity tests and is of crucial importance to validate elaborate assumptions on covariance matrix structures, for example on meta-analysis and error covariance structures in mixed models and models for longitudinal data.

We will say that a covariance matrix Σ has a HBM spherical structure if we can write

$$\Sigma = \begin{bmatrix} I_{k_1} \otimes \Delta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & I_{k_m} \otimes \Delta_m \end{bmatrix}, \quad (\Delta_\ell \text{ unspecified}, \ell = 1, ..., m)\,. \tag{1.1}$$

where $\otimes$ denotes the Kronecker product, and for $\ell = 1, \ldots, m$, I_{k_ℓ} denotes the identity matrix of order k_ℓ and Δ_ℓ is a positive-definite matrix.

The HBM spherical structure may arise in many situations and has as particular cases many interesting and important structures which may be of interest not only as covariance structures in multivariate analysis as well as covariance structures for the error in linear mixed and repeated measures models.

Let us consider a situation in which the same p^* random variables (r.v.'s), $X_1, \ldots, X_{p^*}$, are measured in m "locals", in the ℓ-th of which ($\ell = 1, \ldots, m$) we take k_ℓ measurements, that is, a sample of size k_ℓ, and let us suppose we organize such a meta-sample in a matrix X of dimensions $p^* \times n$, with $n = \sum_{\ell=1}^{m} k_\ell$, as in Figure 1.

$$
\underset{(p^*\times n)}{X} = \left[\begin{array}{ccccccc}
\overbrace{X_{11}\cdots X_{1k_1}}^{k_1} & \cdots & \overbrace{X_{1,k_1+\cdots+k_{\ell-1}+1}\cdots X_{1,k_1+\cdots+k_\ell}}^{k_\ell} & \cdots & & \overbrace{X_{1,k_1+\cdots+k_{m-1}+1}\cdots X_{1,k_1+\cdots+k_m}}^{k_m} & \\
X_{21}\cdots X_{2k_1} & \cdots & X_{2,k_1+\cdots+k_{\ell-1}+1}\cdots X_{2,k_1+\cdots+k_\ell} & \cdots & & X_{2,k_1+\cdots+k_{m-1}+1}\cdots X_{2,k_1+\cdots+k_m} & \\
\vdots \quad\quad \vdots & \vdots & \vdots & & & \vdots \quad\quad \vdots & \\
X_{p^*1}\cdots X_{p^*k_1} & \cdots & X_{p^*,k_1+\cdots+k_{\ell-1}+1}\cdots X_{p^*,k_1+\cdots+k_\ell} & \cdots & & X_{p^*,k_1+\cdots+k_{m-1}+1}\cdots X_{p^*,k_1+\cdots+k_m} &
\end{array}\right]
$$

Figure 1: Data matrix illustrating a situation of a meta-sample from m "locals", with sample size k_ℓ for the ℓ-th local and $p^*_\ell = p^*$ $(\ell = 1, \ldots, m)$.

Here "locals" is a general designation for example for different locals, factories, companies, hospitals, etc., and if we consider the p^* r.v.'s $X_1, \ldots, X_{p^*}$ organized in the random vector $\underline{X}^* = [X_1, \ldots, X_{p^*}]'$, with $Cov(\underline{X}^*) = \Delta$, then we have

$$
Cov(X) = Cov(vec(X)) = \begin{bmatrix} I_{k_1} \otimes \Delta & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & I_{k_m} \otimes \Delta \end{bmatrix}.
$$

But, the HBM sphericity setup allows for more general situations as for example those in which we may want to study or model possible differences in strength break in a set of p^* components manufactured by m different companies, or the measurements of p^* variables thought to be possible important indicators of some disorder or disease, measured across m hospitals, or measurements of p^* pollutants in m different locals, or measurements of p^* atmospheric variables and indicators in m different cities, by taking a sample of size k_ℓ in the ℓ-th "local", but that, furthermore, not in every "local", city, hospital or company, it was possible to obtain measurements of all p^* variables, although we still want to consider as many of these in each "local" as possible. Then we may want to consider a meta-sample as the one illustrated in Figure 2, for $p^* = 5$.

In this case, since in different "locals" we may have different subsets of the p^* variables being analyzed, we may end-up with a covariance setup for the matrix X as the one in (1.1), with different covariance matrices for each "local".

Once we assume the HBM spherical structure for the covariance structure in our model, we may then be interested in testing if that is indeed a plausible

model for our covariances. The issues are thus: (i) how can we carry out a test for such an elaborate structure, and (ii) in case we find a way of doing so, how will we then be able to obtain p-values and/or quantiles for our test statistic, since this may have a quite elaborate exact distribution.

$$
\underset{(p^*\times n)}{X} = \begin{bmatrix}
\overbrace{X_{11}\dots X_{1k_1}}^{k_1} & \dots & \overbrace{X_{1,k_1+\dots+k_{\ell-1}+1}\dots X_{1,k_1+\dots+k_\ell}}^{k_\ell} & \dots & \overbrace{X_{1,k_1+\dots+k_{m-1}+1}\dots X_{1,k_1+\dots+k_m}}^{k_m} \\
X_{21}\dots X_{2k_1} & \dots & \text{———} \quad \text{—} \quad \text{———} & \dots & X_{2,k_1+\dots+k_{m-1}+1}\dots X_{2,k_1+\dots+k_m} \\
X_{31}\dots X_{3k_1} & \dots & X_{3,k_1+\dots+k_{\ell-1}+1}\dots X_{3,k_1+\dots+k_\ell} & \dots & \dots \\
\text{—} \quad \text{—} \quad \text{—} & \dots & X_{4,k_1+\dots+k_{\ell-1}+1}\dots X_{4,k_1+\dots+k_\ell} & \dots & X_{4,k_1+\dots+k_{m-1}+1}\dots X_{4,k_1+\dots+k_m} \\
X_{51}\dots X_{5k_1} & \dots & \text{———} \quad \text{—} \quad \text{———} & \dots & X_{5,k_1+\dots+k_{m-1}+1}\dots X_{5,k_1+\dots+k_m}
\end{bmatrix}
$$

Figure 2: Data matrix illustrating a situation of a meta-sample from m "locals", with sample size k_ℓ for the ℓ-th local ($\ell = 1, \dots, m$), $p_1^* = 4$, $p_\ell^* = 3$ and $p_m^* = 4$, with different covariance matrices Δ_ℓ ($\ell = 1, \dots, m$).

These are indeed the issues we propose to address in this paper, namely showing how one can quite easily build the l.r.t. statistic for the test of the HBM Spherical structure and how we can then obtain the expression for the moments of the statistic and even for the characteristic function (c.f.) of its logarithm, from which factorization we will then be able to obtain very sharp approximations for the exact distribution of the statistic.

The HBM sphericity test is thus a test where the null hypothesis is written as

$$
(1.2) \qquad H_0 : \Sigma = \begin{bmatrix} I_{k_1} \otimes \Delta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & I_{k_m} \otimes \Delta_m \end{bmatrix}, \qquad (\Delta_\ell \text{ unspecified}, \ell = 1, ..., m)
$$

where Σ is the covariance matrix of the random vector $\underline{X}$ and the matrices Δ_ℓ are $p_\ell^* \times p_\ell^*$, ($\ell = 1, ..., m$), with $p_\ell = k_\ell \times p_\ell^*$ and $p = \sum_{\ell=1}^m p_\ell$.

This test is a generalization of the standard sphericity test and it has as particular cases a number of interesting and important tests:

(i) the block-matrix sphericity (BM-Sph) test, for $m = 1$ [4, 3, 15],

(ii) the block-scalar sphericity (BS-Sph) test, for $p_\ell^* = 1, (\ell = 1, ..., m)$ [19, 18, 13],

(iii) the block independence (BI) test, for $k_\ell = 1, (\ell = 1, ..., m)$ [24, 25], [1, Chap. 9], [17, Sec. 11.2], [6, 7],

(iv) the common independence (Ind) test, for $p_\ell^* = k_\ell = 1, (\ell = 1, ..., m)$ [22, Sec. 7.4.3], [9, Secs. 1,2], and

(v) the sphericity (Sph) test, for $m = p_1^* = 1$ [16], [1, Sec. 10.7], [17, Sec. 8.3], [9].

These particular cases as well as their relations may be analyzed in Figure 3.

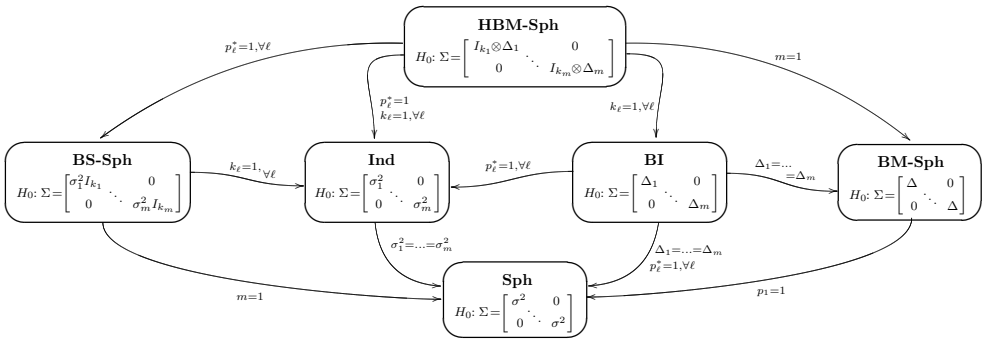

Figure 3: Particular cases of the HBM (Hyper-Block Matrix) test and their inter-relations:

HBM-Sph: Hyper-block matrix sphericity test;
BM-Sph: Block-matrix sphericity test;
BS-Sph: Block-scalar sphericity test;
Sph: Sphericity test; **Ind**: Independence test; **BI**: Block independence test.

The exact distribution of the HBM sphericity test statistic is almost intractable in practical terms, thus our propose is to develop near-exact distributions for the test statistic and its logarithm, based on an adequate factorization of the c.f. of the logarithm of the test statistic.

In Section 2 we will show how we may decompose the overall null hypothesis in (1.2) into a set of three conditionally independent hypotheses and then how from this decomposition (see [8]) we may derive expressions for the l.r.t. statistic and its h-th moment. In Section 3 we will show how we may easily obtain the expression for the c.f. of the logarithm of the test statistic and how we may use the decomposition of the null hypothesis in Section 2, to induce an adequate factorization of this c.f. in two factors, one that is the c.f. a Generalized Integer Gamma (GIG) distribution [6], and the other the c.f. of a sum of independent r.v.'s whose exponentials have Beta distributions. Then, in Section 4 we will use this factorization to build very sharp near-exact distributions both for the test statistic and its logarithm.

Near-exact distributions are asymptotic distributions built using a different approach. Usually working from an adequate factorization of the c.f. of the logarithm of the l.r.t. statistic, we leave unchanged the set of factors that correspond to a manageable distribution and approximate asymptotically the remaining set of factors, in such a way that the resulting c.f., which we will call a near-exact c.f., corresponds to a known manageable distribution, from which p-values and quantiles may be easily computed. These near-exact distributions lie much closer to the exact distribution than any common asymptotic distribution and, when correctly built for statistics used in Multivariate Analysis, will show a marked asymptotic behavior not only for increasing sample sizes but also for increasing number of variables involved.

In Section 5 we will use a measure of proximity between the exact distribution and the near-exact distributions, based on the corresponding characteristic functions in order to assess the quality and the asymptotic properties of the near-exact distributions developed.

Section 6 is dedicated to power studies, where the very good behavior of the test is revealed, through studies based on 1 000 000 pseudo-random samples and carried out on several scenarios of violation of the null hypothesis of HBM sphericity.

2. THE TEST STATISTICS AND ITS MOMENTS

In general terms, a null hypothesis H_0 may be decomposed into a sequence of three conditionally independent null hypotheses, if H_0 admits the decomposition

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

where '$\circ$' is to be read as 'after', as long as

$$\Omega_{H_0} \equiv \Omega_{H_{03|1,2}} \subset \Omega_{H_{13|1,2}} \equiv \Omega_{H_{02|1}} \subset \Omega_{H_{12|1}} \equiv \Omega_{H_{01}} \subset \Omega_{H_{11}} \equiv \Omega_{H_1}$$

where Ω_{H_0} is the parameter space under H_0 and Ω_{H_1} the union of the parameter spaces under H_0 and H_1, and where H_{1*} represents the alternative hypothesis to H_{0*} (where '$*$' is used as a wildcard).

The null hypothesis

$$H_{01} : \Sigma = bdiag\,(\Sigma_{\ell\ell}; \ell = 1, ..., m) \tag{2.1}$$

corresponds to the test of independence of m groups of variables, the ℓ-th group having $p_\ell = p_\ell^* \times k_\ell$ variables ($\ell = 1, ..., m$).

If we consider that the random vector $\underline{X}$ has a p-variate Normal distribution with expected value vector $\underline{\mu}$ and covariance matrix Σ, that is, if we consider the vector $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma\right)$ and suppose that we have a sample of size $N\,(>p)$ from $\underline{X}$, then the l.r.t. statistic used to test H_{01} and its h-th moment are respectively given by (see secs. 9.2 and 9.3.2 in [1])

$$\Lambda_1 = \frac{|A|^{\frac{N}{2}}}{\prod\limits_{\ell=1}^{m} |A_{\ell\ell}|^{\frac{N}{2}}} \tag{2.2}$$

and

$$E\left[(\Lambda_1)^h\right] = \prod_{\ell=1}^{m-1}\prod_{k=1}^{p_\ell} \frac{\Gamma\left(\frac{N-q_\ell-k}{2}+\frac{N}{2}h\right)}{\Gamma\left(\frac{N-q_\ell-k}{2}\right)} \frac{\Gamma\left(\frac{N-k}{2}\right)}{\Gamma\left(\frac{N-k}{2}+\frac{N}{2}h\right)}, \quad \left(h>\tfrac{p-p_m}{N}-1\right) \tag{2.3}$$

where the matrix A is the maximum likelihood estimator (m.l.e.) of Σ, $A_{\ell\ell}$ its ℓ-th diagonal block of order p_ℓ $(\ell = 1, ..., m)$ and $q_\ell = p_{\ell+1} + ... + p_m$.

The null hypothesis

$$H_{02|1} = \bigwedge_{\ell=1}^{m} H_{02|1}^{\ell} \tag{2.4}$$

where for $\ell = 1, ..., m$

$$\begin{aligned} H_{02|1}^{\ell}:\ & \Sigma_{\ell\ell} = bdiag\left(\Sigma_{vv}^{\ell}, v = 1, ..., k_\ell\right) \\ & \text{assuming } \Sigma = bdiag\left(\Sigma_{\ell\ell}, \ell = 1, ..., m\right) \\ & \text{that is, assuming } H_{01} \end{aligned} \tag{2.5}$$

is the null hypothesis of a test of independence of k_ℓ groups of variables, with p_ℓ^* variables each. The l.r.t. statistic to test $H_{02|1}^{\ell}$ in (2.5) and its h-th moment are respectively given by (see Secs. 9.2 and 9.3.2 in [1])

$$\Lambda_{2|1}^{\ell} = \frac{|A_{\ell\ell}|^{\frac{N}{2}}}{\prod\limits_{v=1}^{k_\ell} \left|A_{\ell\ell}^{v}\right|^{\frac{N}{2}}}$$

and

$$E\left[\left(\Lambda_{2|1}^{\ell}\right)^h\right] = \prod_{v=1}^{k_\ell-1}\prod_{k=1}^{p_\ell^*} \frac{\Gamma\left(\frac{N-q_v^\ell-k}{2}+\frac{N}{2}h\right)}{\Gamma\left(\frac{N-q_v^\ell-k}{2}\right)} \frac{\Gamma\left(\frac{N-k}{2}\right)}{\Gamma\left(\frac{N-k}{2}+\frac{N}{2}h\right)}, \quad \left(h>\tfrac{p_\ell}{N}-1\right)$$

where the matrix $A_{\ell\ell}$ is the maximum likelihood estimator of $\Sigma_{\ell\ell}$, $A_{\ell\ell}^{v}$ its v-th $(v = 1, ..., k_\ell)$ diagonal block of order p_ℓ^* and $q_v^\ell = (k_\ell - v)\, p_\ell^*$, $(v = 1, ..., k_\ell)$.

The l.r.t. statistic to test the null hypothesis in (2.4) is thus

$$\Lambda_{2|1} = \prod_{\ell=1}^{m} \Lambda_{2|1}^{\ell} = \prod_{\ell=1}^{m} \frac{|A_{\ell\ell}|^{\frac{N}{2}}}{\prod\limits_{v=1}^{k_\ell} \left|A_{\ell\ell}^{v}\right|^{\frac{N}{2}}} \tag{2.6}$$

and, given the fact that the statistics $\Lambda_{2|1}^{\ell}$, $\ell = 1, \ldots, m$, form a set of m independent statistics, since under H_{01} in (2.1) the m matrices $A_{\ell\ell}$ are independent and each statistic $\Lambda_{2|1}^{\ell}$ is built only from $A_{\ell\ell}$, the h-th moment of $\Lambda_{2|1}$ is

$$
\begin{aligned}
E\left[\left(\Lambda_{2|1}\right)^{h}\right] &= \prod_{\ell=1}^{m} E\left[\left(\Lambda_{2|1}^{\ell}\right)^{h}\right] \\
&= \prod_{\ell=1}^{m}\prod_{v=1}^{k_\ell-1}\prod_{k=1}^{p_\ell^*} \frac{\Gamma\left(\frac{N-q_v^\ell-k}{2}+\frac{N}{2}h\right)}{\Gamma\left(\frac{N-q_v^\ell-k}{2}\right)} \frac{\Gamma\left(\frac{N-k}{2}\right)}{\Gamma\left(\frac{N-k}{2}+\frac{N}{2}h\right)}, \\
&\qquad \left(h > \max\left\{\tfrac{p_\ell}{N}-1, \ell = 1, ..., m\right\}\right).
\end{aligned} \tag{2.7}
$$

Finally, the null hypothesis $H_{03|1,2}$ may be written as

$$
H_{03|1,2} = \bigwedge_{\ell=1}^{m} H_{03|1,2}^{\ell} \tag{2.8}
$$

where, for $\ell = 1, ..., m$,

$$
\begin{aligned}
H_{03|1,2}^{\ell} :\ & \Sigma_{11}^{\ell} = \cdots = \Sigma_{k_\ell k_\ell}^{\ell} = \Delta_\ell \ \ (\Delta_\ell \text{ unspecified}) \\
& \text{assuming } \Sigma = bdiag\left(\Sigma_{\ell\ell} = \left(bdiag\left(\Sigma_{vv}^{\ell}, v = 1, ..., k_\ell\right)\right)\right) \\
& \text{that is, assuming } H_{02|1} \text{ and } H_{01}
\end{aligned} \tag{2.9}
$$

is the null hypothesis corresponding to the test of equality of k_ℓ covariance matrices each with dimensions $p_\ell^* \times p_\ell^*$.

Since under $H_{02|1}$, for each $\ell = 1, \ldots, m$, the k_ℓ matrices $A_{\ell\ell}^{v}$ $(v = 1, \ldots, k_\ell)$ are independent, The l.r.t. statistic to test each null hypothesis $H_{03|1,2}^{\ell}$ in (2.9) and its h-th moment are respectively, (see Secs. 10.2 and 10.4.2 in [1])

$$
\Lambda_{3|1,2}^{\ell} = \frac{\prod_{v=1}^{k_\ell} \left|A_{\ell\ell}^{v}\right|^{\frac{N}{2}}}{\left|A_\ell^*\right|^{k_\ell \frac{N}{2}}} k_\ell^{\frac{N p_\ell}{2}}
$$

and

$$
E\left[\left(\Lambda_{3|1,2}^{\ell}\right)^{h}\right] = \prod_{k=1}^{p_\ell^*}\prod_{v=1}^{k_\ell} \frac{\Gamma\left(\frac{N-1}{2}-\frac{k-1}{2k_\ell}+\frac{v-1}{k_\ell}\right)}{\Gamma\left(\frac{N-1}{2}-\frac{k-1}{2k_\ell}+\frac{v-1}{k_\ell}+\frac{N}{2}h\right)} \frac{\Gamma\left(\frac{N-k}{2}+\frac{N}{2}h\right)}{\Gamma\left(\frac{N-k}{2}\right)}
$$

where the matrix $A_{\ell\ell}$ is the maximum likelihood estimator of $\Sigma_{\ell\ell}$, $A_{\ell\ell}^{v}$ its v-th diagonal block of order p_ℓ^* $(v = 1, ..., k_\ell)$ and $A_\ell^* = A_{\ell\ell}^{1} + \cdots + A_{\ell\ell}^{k_\ell}$.

The l.r.t. statistic to test (2.8) is thus

$$
\Lambda_{3|1,2} = \prod_{\ell=1}^{m} \Lambda_{3|1,2}^{\ell} = \prod_{\ell=1}^{m} \frac{\prod_{v=1}^{k_\ell} \left|A_{\ell\ell}^{v}\right|^{\frac{N}{2}}}{\left|A_\ell^*\right|^{k_\ell \frac{N}{2}}} k_\ell^{\frac{N p_\ell}{2}} \tag{2.10}
$$

and, since under H_{01} in (2.1), the m statistics $\Lambda^{\ell}_{3|1,2}$ are independent, given that each statistic $\Lambda^{\ell}_{3|1,2}$ is built only from $A_{\ell\ell}$ and under H_{01} the m matrices $A_{\ell\ell}$ $(\ell = 1, \ldots, m)$ are independent, the h-th moment of $\Lambda_{3|1,2}$ is given by

$$\begin{aligned} E\left[\left(\Lambda_{3|1,2}\right)^h\right] &= \prod_{\ell=1}^{m} E\left[\left(\Lambda^{\ell}_{3|1,2}\right)^h\right] \\ &= \prod_{\ell=1}^{m}\prod_{k=1}^{p^*_\ell}\prod_{v=1}^{k_\ell} \frac{\Gamma\left(\frac{N-1}{2} - \frac{k-1}{2k_\ell} + \frac{v-1}{k_\ell}\right)\Gamma\left(\frac{N-k}{2} + \frac{N}{2}h\right)}{\Gamma\left(\frac{N-1}{2} - \frac{k-1}{2k_\ell} + \frac{v-1}{k_\ell} + \frac{N}{2}h\right)\Gamma\left(\frac{N-k}{2}\right)}, \\ &\qquad \left(h > \max\left\{\frac{p^*_\ell}{N} - 1, \ell = 1, ..., m\right\}\right). \end{aligned}$$

From Lemma 10.3.1 in [1], the l.r.t. statistic to test (1.2) is thus the product of the l.r.t. statistics used to test the null hypotheses in (2.1), (2.4) and (2.8), that is, the product of the l.r.t. statistics in (2.2), (2.6) and (2.10), thus with

$$\begin{aligned} \Lambda &= \Lambda_1 \times \Lambda_{2|1} \times \Lambda_{3|1,2} \\ &= \frac{|A|^{\frac{N}{2}}}{\prod_{\ell=1}^{m}|A_{\ell\ell}|^{\frac{N}{2}}} \times \left\{\prod_{\ell=1}^{m}\frac{|A_{\ell\ell}|^{\frac{N}{2}}}{\prod_{v=1}^{k_\ell}\left|A^v_{\ell\ell}\right|^{\frac{N}{2}}}\right\} \times \left\{\prod_{\ell=1}^{m}\frac{\prod_{v=1}^{k_\ell}|A^v_{\ell\ell}|^{\frac{N}{2}}}{\left|A^*_\ell\right|^{k_\ell\frac{N}{2}}}\, k_\ell^{\frac{Np_\ell}{2}}\right\} \\ &= \left\{\prod_{\ell=1}^{m} k_\ell^{\frac{Np_\ell}{2}}\right\} \times \frac{|A|^{\frac{N}{2}}}{\prod_{\ell=1}^{m}\left|A^*_\ell\right|^{k_\ell\frac{N}{2}}} \end{aligned} \tag{2.11}$$

where the matrix A is the m.l.e. of Σ, $A_{\ell\ell}$ is the ℓ-th diagonal block of order p_ℓ of A $(\ell = 1, ..., m)$, with $p = \sum_{\ell=1}^{m} p_\ell$ and $A^*_\ell = A^1_{\ell\ell} + \cdots + A^{k_\ell}_{\ell\ell}$, where $A^v_{\ell\ell}$ is the v-th $(v = 1, ..., k_\ell)$ diagonal block of order p^*_ℓ of $A_{\ell\ell}$ and $p_\ell = k_\ell \times p^*_\ell$. We should note that the expression in (2.11) is identical to the one we obtain when we use the usual method of derivation of the l.r.t. statistic, through its definition (see Appendix A).

The hypotheses $H_{03|1,2}$, $H_{02|1}$ and H_{01} are independent in the sense that under the overall null hypothesis H_0 it is possible to prove that the l.r.t. statistics used to test these hypotheses are independent. Indeed, by Lemma 10.4.1 in [1] or Theorem 5 in [10], the l.r.t. statistic Λ_1 in (2.2) is independent of the m matrices $A_{\ell\ell}$, $(\ell = 1, ..., m)$, so that since $\Lambda_{2|1}$ and $\Lambda_{3|1,2}$ are built only from the m matrices $A_{\ell\ell}$ $(\ell = 1, ..., m)$, these l.r.t. statistics are independent of Λ_1. But then, since we may use the same two results to argue that each statistic $\Lambda^{\ell}_{2|1}$ is independent of the k_ℓ matrices $A^v_{\ell\ell}$ $(v = 1, \ldots, k_\ell)$, the statistic $\Lambda_{2|1}$ is independent of all $\sum_{\ell=1}^{m} k_\ell$ matrices $A^v_{\ell\ell}$ $(\ell = 1, \ldots, k_\ell; v = 1, \ldots, m)$ and as such independent of $\Lambda_{3|1,2}$ which is built only on these matrices.

As such, we may obtain the expression of the h-th moment of Λ from the expressions for the h-th moment of each of the statistics $\Lambda_{3|1,2}$, $\Lambda_{2|1}$ and Λ_1 writing it as

$$\begin{aligned}
E\left[(\Lambda)^h\right] &= E\left[(\Lambda_1)^h\right] \times E\left[\left(\Lambda_{2|1}\right)^h\right] \times E\left[\left(\Lambda_{3|1,2}\right)^h\right] \\
&= \prod_{\ell=1}^{m}\prod_{k=1}^{p_\ell^*}\prod_{v=1}^{k_\ell-1} \frac{\Gamma\left(\frac{N-1}{2}-\frac{k-1}{2k_\ell}+\frac{v-1}{k_\ell}\right)}{\Gamma\left(\frac{N-1}{2}-\frac{k-1}{2k_\ell}+\frac{v-1}{k_\ell}+\frac{N}{2}h\right)} \frac{\Gamma\left(\frac{N-q_v^\ell-k}{2}+\frac{N}{2}h\right)}{\Gamma\left(\frac{N-q_v^\ell-k}{2}\right)} \\
&\quad \times \prod_{\ell=1}^{m}\prod_{k=1}^{p_\ell^*} \frac{\Gamma\left(\frac{N-1}{2}-\frac{k-1}{2k_\ell}+\frac{k_\ell-1}{k_\ell}\right)}{\Gamma\left(\frac{N-1}{2}-\frac{k-1}{2k_\ell}+\frac{k_\ell-1}{k_\ell}+\frac{N}{2}h\right)} \frac{\Gamma\left(\frac{N-k}{2}+\frac{N}{2}h\right)}{\Gamma\left(\frac{N-k}{2}\right)} \\
&\quad \times \prod_{\ell=1}^{m-1}\prod_{k=1}^{p_\ell} \frac{\Gamma\left(\frac{N-q_\ell-k}{2}+\frac{N}{2}h\right)}{\Gamma\left(\frac{N-q_\ell-k}{2}\right)} \frac{\Gamma\left(\frac{N-k}{2}\right)}{\Gamma\left(\frac{N-k}{2}+\frac{N}{2}h\right)},
\end{aligned}$$

for $h > \max\{(p-p_m)/N-1, p_\ell/N-1 (\ell=1,...,m)\}$ and where $p=\sum_{\ell=1}^m p_\ell$, $q_\ell = p_{\ell+1}+...+p_m$, $p_\ell = k_\ell \times p_\ell^*$ and $q_v^\ell = (k_\ell - v)\, p_\ell^*$ $(\ell=1,...,m; v=1,...,k_\ell)$.

Actually we may note that the two first null hypotheses may be condensed into a single null hypothesis to test the independence of $q^{**}=\sum_{\ell=1}^m k_\ell$ groups of variables, the ν-th group with p_ν^{**} variables for

$$p^{**} = \Big[\underbrace{p_1^*,\ldots,p_1^*}_{k_1},\underbrace{p_2^*,\ldots,p_2^*}_{k_2},\ \ldots\ ,\underbrace{p_\ell^*,\ldots,p_\ell^*}_{k_\ell},\ \ldots\ ,\underbrace{p_m^*,\ldots,p_m^*}_{k_m}\Big], \tag{2.12}$$

with $\nu=1,\ldots,\sum_{\ell=1}^m k_\ell$, that is, with

$$p_\nu^{**} = p_\ell^*, \quad \text{for} \quad 1+\sum_{i=1}^{\ell-1} k_i \le \nu \le \sum_{i=1}^{\ell} k_i\,.$$

This null hypothesis may be written as

$$H_{0,12}: \Sigma = bdiag\Big(\underbrace{\Delta_1,\ldots,\Delta_{k_1}}_{\text{of order } p_1^*},\ \ldots\ ,\underbrace{\Delta_{k_1+\cdots+k_{\ell-1}+1},\ldots,\Delta_{k_1+\cdots+k_\ell}}_{\text{of order } p_\ell^*},\ \ldots \\ \ldots\ ,\underbrace{\Delta_{k_1+\cdots+k_{m-1}+1},\ldots,\Delta_{k_1+\cdots+k_m}}_{\text{of order } p_m^*}\Big).$$

The l.r.t. statistic to test $H_{0,12}$ is given by

$$\Lambda_{1,2} = \frac{|A|^{\frac{N}{2}}}{\prod_{\nu=1}^{q^{**}} |A_{\nu\nu}^*|^{\frac{N}{2}}} \tag{2.13}$$

where $A^*_{\nu\nu}$ is the ν-th diagonal block of order p^{**}_ν ($\nu = 1, ..., q^{**}$), and the expression of its h-th moment is given by (see secs. 9.2 and 9.3.2 in [1])

$$E\left[(\Lambda_{1,2})^h\right] = \prod_{\nu=1}^{q^{**}-1}\prod_{k=1}^{p^{**}_\nu} \frac{\Gamma\left(\frac{N-q^*_\nu-k}{2}+\frac{N}{2}h\right)}{\Gamma\left(\frac{N-q^*_\nu-k}{2}\right)} \frac{\Gamma\left(\frac{N-k}{2}\right)}{\Gamma\left(\frac{N-k}{2}+\frac{N}{2}h\right)}$$

for

$$q^*_\nu = p^{**}_{\nu+1} + \cdots + p^{**}_{q^{**}} \quad \text{where} \quad q^{**} = \sum_{\ell=1}^{m} k_\ell\,.$$

Note that the l.r.t. statistic in (2.13) may be also given by the product of the l.r.t.'s in (2.2) and (2.6), used to test the null hypotheses in (2.1) and (2.4) respectively, and the expression of its h-th moment may also be given by the product of the expressions of the h-th moments in (2.3) and (2.7) of the l.r.t.'s in (2.2) and (2.6) respectively.

Finally, the expression of the h-th moment of Λ may be re-written as

$$\begin{aligned} E\left[(\Lambda)^h\right] &= E\left[(\Lambda_{1,2})^h\right] \times E\left[\left(\Lambda_{3|1,2}\right)^h\right] \\ &= \prod_{\nu=1}^{q^{**}-1}\prod_{k=1}^{p^{**}_\nu} \frac{\Gamma\left(\frac{N-q^*_\nu-k}{2}+\frac{N}{2}h\right)}{\Gamma\left(\frac{N-q^*_\nu-k}{2}\right)} \frac{\Gamma\left(\frac{N-k}{2}\right)}{\Gamma\left(\frac{N-k}{2}+\frac{N}{2}h\right)} \\ &\quad \times \prod_{\ell=1}^{m}\prod_{k=1}^{p^*_\ell}\prod_{v=1}^{k_\ell} \frac{\Gamma\left(\frac{N-1}{2}-\frac{k-1}{2k_\ell}+\frac{v-1}{k_\ell}\right)}{\Gamma\left(\frac{N-1}{2}-\frac{k-1}{2k_\ell}+\frac{v-1}{k_\ell}+\frac{N}{2}h\right)} \frac{\Gamma\left(\frac{N-k}{2}+\frac{N}{2}h\right)}{\Gamma\left(\frac{N-k}{2}\right)}\,. \end{aligned} \tag{2.14}$$

The factorization of the c.f. of $W = -\log\Lambda$ developed in the next section will have as a starting base this last expression.

3. THE CHARACTERISTIC FUNCTION OF $W = -\log(\Lambda)$

Since in (2.14) the Gamma functions remain valid for any strictly complex h, if we take $W_{1,2} = -\log\Lambda_{1,2}$ and $W_3 = -\log\Lambda_{3|1,2}$, we may write the c.f. of $W = -\log\Lambda$ as

$$
\begin{aligned}
\Phi_W(t) &= E\left(e^{-it\log\Lambda}\right) = E\left[\Lambda^{-it}\right] = E\left[\Lambda_{1,2}^{-it}\right] E\left[\Lambda_{3|1,2}^{-it}\right] \\
&= \Phi_{W_{1,2}}(t)\,\Phi_{W_3}(t) \\
&= \underbrace{\prod_{\nu=1}^{q^{**}-1}\prod_{k=1}^{p_\nu^{**}} \frac{\Gamma\left(\frac{N-q_\nu^*-k}{2}-\frac{N}{2}it\right)}{\Gamma\left(\frac{N-q_\nu^*-k}{2}\right)} \frac{\Gamma\left(\frac{N-k}{2}\right)}{\Gamma\left(\frac{N-k}{2}-\frac{N}{2}it\right)}}_{\Phi_{W_{1,2}}(t)} \\
&\quad \times \underbrace{\prod_{\ell=1}^{m}\prod_{k=1}^{p_\ell^*}\prod_{v=1}^{k_\ell} \frac{\Gamma\left(\frac{N-1}{2}-\frac{k-1}{2k_\ell}+\frac{v-1}{k_\ell}\right)}{\Gamma\left(\frac{N-1}{2}-\frac{k-1}{2k_\ell}+\frac{v-1}{k_\ell}-\frac{N}{2}it\right)} \frac{\Gamma\left(\frac{N-k}{2}-\frac{N}{2}it\right)}{\Gamma\left(\frac{N-k}{2}\right)}}_{\Phi_{W_3}(t)},
\end{aligned}
\tag{3.1}
$$

where $t \in \mathbb{R}$, $i = \sqrt{-1}$, $q_\nu^* = p_{\nu+1}^{**} + \cdots + p_{q^{**}}^{**}$, $q^{**} = \sum_{\ell=1}^m k_\ell$, and p_ν^{**} are defined in (2.12). From this expression we may state that

$$
\Lambda \overset{d}{\equiv} \left\{\prod_{\nu=1}^{q^{**}-1}\prod_{k=1}^{p_\nu^{**}} (Y_{\nu k})^{N/2}\right\} \times \left\{\prod_{\ell=1}^{m}\prod_{k=1}^{p_\ell^*}\prod_{v=1}^{k_\ell} (Y_{\ell kv})^{N/2}\right\}
$$

where

$$
Y_{\nu k} \sim Beta\left(\frac{N-q_\nu^*-k}{2}, \frac{q_\nu^*}{2}\right) \quad \text{and} \quad Y_{\ell kv} \sim Beta\left(\frac{N-k}{2}, \frac{v-1}{k_\ell}+\frac{k-1}{2}\,\frac{k_\ell-1}{k_\ell}\right)
$$

are independent r.v.'s.

In order to be able to build sharp near-exact distributions for W and Λ, we need to further factorize $\Phi_{W_{1,2}}(t)$ and $\Phi_{W_3}(t)$, by writing each one of these c.f.'s as the product of two factors, one that is the c.f. of a Generalized Integer Gamma (GIG) distribution and the other the c.f. of a sum of independent r.v.'s whose exponentials have Beta distributions.

3.1. The factorization of the characteristic function of $W_{1,2} = -\log\Lambda_{1,2}$

The results in [7, 14] may be used to show that $\Phi_{W_{1,2}}(t)$ in (3.1) may be written as

$$
\Phi_{W_{1,2}}(t) = \Phi_{W_{1,2,a}}(t) \times \Phi_{W_{1,2,b}}(t)\ ,
$$

where

$$
\Phi_{W_{1,2,a}}(t) = \prod_{k=3}^{p}\left(\frac{N-k}{N}\right)^{r_{1,k}}\left(\frac{N-k}{N}-it\right)^{-r_{1,k}} \tag{3.2}
$$

is the c.f. of the sum of $p-2$ independent integer Gamma r.v.'s, that is a Generalized Integer Gamma distribution of depth $p-2$, with integer shape parameters

$r_{1,k}$ given by

$$r_{1,k} = \begin{cases} h_{1,k-2} + (-1)^k k^* & k = 3, 4 \\ r_{1,k-2} + h_{1,k-2}, & k = 5, ..., p \end{cases} \tag{3.3}$$

with $k^* = \left\lfloor \frac{m^*}{2} \right\rfloor$, where m^* is the number of sets of variables with an odd number of variables, among the q^{**} groups of variables, the ν-th of which with p_ν^{**} variables, and

$$h_{1;k} = (\#\text{ of } p_\nu^{**}\,(\nu = 1, .., q^{**}) \geq k) - 1, \quad k = 1, ..., p-2$$

and

$$\Phi_{W_{1,2,b}}(t) = \left(\frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)} \frac{\Gamma\left(\frac{N-2}{2} - \frac{N}{2}it\right)}{\Gamma\left(\frac{N-1}{2} - \frac{N}{2}it\right)} \right)^{k^*} \tag{3.4}$$

is the c.f. of the sum of k^* independent r.v.'s with Logbeta distributions multiplied by $\frac{N}{2}$. We should note that when $k^* = 0$, $\Phi_{W_{1,2,b}}(t) = 1$.

3.2. The factorization of the characteristic function of $W_3 = -\log \Lambda_{3|1,2}$

Based on the results in [14] we may re-write the c.f. of W_3, $\Phi_{W_3}(t)$ in (3.1), as

$$\Phi_{W_3}(t) = \Phi_{W_{3,a}}(t) \times \Phi_{W_{3,b}}(t)$$

where

$$\Phi_{W_{3,a}}(t) = \prod_{\ell=1}^{m} \prod_{k=2}^{p_\ell^*} \left(\frac{N-k}{N} \right)^{r_{3,k-1}^\ell} \left(\frac{N-k}{N} - it \right)^{-r_{3,k-1}^\ell} \tag{3.5}$$

is the c.f. of the sum of $\sum_{\ell=1}^{m} p_\ell^* - m$ independent Gamma r.v.'s, that is, a Generalized Integer Gamma distribution of depth $\sum_{\ell=1}^{m} p_\ell^* - m$ with integer shape parameters $r_{3;k}^\ell$ given by (B.1) in Appendix B, and

$$\begin{aligned} \Phi_{W_{3,2}}(t) = \prod_{\ell=1}^{m} \Bigg\{ & \prod_{k=1}^{\lfloor p_\ell^*/2 \rfloor} \prod_{v=1}^{k_\ell} \frac{\Gamma\left(a_k^\ell + b_{kv}^\ell\right)}{\Gamma\left(a_k^\ell + b_{kv}^{\ell *}\right)} \frac{\Gamma\left(a_k^\ell + b_{vk}^{\ell *} - Nit\right)}{\Gamma\left(a_k^\ell + b_{vk}^{\ell} - Nit\right)} \\ & \times \left(\prod_{v=1}^{k_\ell} \frac{\Gamma\left(a_{p_\ell^*}^\ell + b_{p_\ell^* v}^\ell\right)}{\Gamma\left(a_{p_\ell^*}^\ell + b_{p_\ell^* v}^{\ell *}\right)} \frac{\Gamma\left(a_{p_\ell^*}^{\ell *} + b_{p_\ell^* v}^{\ell *} - \frac{N}{2}it\right)}{\Gamma\left(a_{p_\ell^*}^\ell + b_{p_\ell^* v}^{\ell} - \frac{N}{2}it\right)} \right)^{(p_\ell^* \bmod 2)} \Bigg\} \end{aligned} \tag{3.6}$$

with

$$a_k^\ell = N - 2k,\ b_{kv}^\ell = 2k - 1 + \tfrac{v-2k}{k_\ell},\ b_{vk}^{\ell *} = \left\lfloor b_{vk}^\ell \right\rfloor \tag{3.7}$$

and

$$a^{\ell}_{p^*_\ell} = \tfrac{N-p^*_\ell}{2},\ b^{\ell}_{p^*_\ell v} = \tfrac{p_\ell - k_\ell - p^*_\ell + 2v - 1}{2k_\ell},\ b^{\ell *}_{p^*_\ell v} = \left\lfloor b^{\ell}_{p^*_\ell v} \right\rfloor, \tag{3.8}$$

is the c.f. of the sum of $\sum_{\ell=1}^{m} \left\lfloor \frac{p^*_\ell}{2} \right\rfloor k_\ell$ independent Logbeta r.v.'s multiplied by N and $\sum_{\ell=1}^{m} k_\ell \,(p^*_\ell \bmod 2)$ independent Logbeta r.v.'s multiplied by $\frac{N}{2}$.

As such, the c.f. of W may be written as

$$\Phi_W(t) = \Phi_1(t) \times \Phi_2(t) \tag{3.9}$$

where

$$\Phi_1(t) = \Phi_{W_{1,2,a}}(t) \times \Phi_{W_{3,a}}(t), \tag{3.10}$$

with $\Phi_{W_{1,2,a}}(t)$ and $\Phi_{W_{3,a}}(t)$ given by (3.2) and (3.5), respectively, and

$$\Phi_2(t) = \Phi_{W_{1,2,b}}(t) \times \Phi_{W_{3,b}}(t) \tag{3.11}$$

with $\Phi_{W_{1,2,b}}(t)$ and $\Phi_{W_{3,b}}(t)$ in (3.4) and (3.6), respectively.

The c.f. $\Phi_1(t)$ in (3.10) can be seen as the c.f. of a GIG distribution of depth $p-1$, and it may be written as

$$\Phi_1(t) = \prod_{k=2}^{p} \left(\frac{N-k}{N}\right)^{r^+_k} \left(\frac{N-k}{N} - it\right)^{-r^+_k} \tag{3.12}$$

where

$$r^+_k = r^*_{1,k} + \sum_{\ell=1}^{m} r^{\ell *}_{3,k}, \tag{3.13}$$

with

$$r^*_{1,k} = \begin{cases} 0 & k = 2 \\ r_{1,k} & k = 3, \ldots, p \end{cases} \quad \text{and} \quad r^{\ell *}_{3,k} = \begin{cases} r^{\ell}_{3,k} & k = 2, \ldots, p^*_\ell \\ 0 & k = p^*_\ell, \ldots, p \end{cases} \tag{3.14}$$

with $r_{1,k}$ given by (3.3) and $r^{\ell}_{3,k}$ given by (B.1) in Appendix B, while $\Phi_2(t)$ is the c.f. of a sum of $k^* + \sum_{\ell=1}^{m} \left\lfloor \frac{p^*_\ell}{2} \right\rfloor k_\ell$ independent Logbeta r.v.'s multiplied by N and $\sum_{\ell=1}^{m} k_\ell \,(p^*_\ell \bmod 2)$ independent Logbeta r.v.'s multiplied by $\frac{N}{2}$.

From this alternative expression for the c.f. of $W = -\log \Lambda$ given by (3.9), we may see that the exact distribution of Λ in (2.11) may be written as

$$\Lambda \overset{d}{\equiv} \left\{ \prod_{k=2}^{p} e^{Z_k} \right\} \left\{ \prod_{k=1}^{k^*} (Y_{1,k})^{\frac{N}{2}} \right\} \left\{ \prod_{\ell=1}^{m} \prod_{k=1}^{\lfloor p^*_\ell/2 \rfloor} \prod_{v=1}^{k_\ell} \left(Y^{\ell}_{3,kv}\right)^N \right\} \times \left\{ \prod_{\ell=1}^{m} \left\{ \prod_{v=1}^{k\ell} (Y^{\ell}_{3,v})^{\frac{N}{2}} \right\}^{p^*_\ell \bmod 2} \right\}$$

where $\stackrel{d}{\equiv}$ means "equivalent in distribution" and all the r.v.'s involved are independent, with

$$
\begin{aligned}
& Z_k \sim \Gamma\left(r_k^+, \frac{N-k}{N}\right), && k = 2, \ldots, p \\
& Y_{1,k} \sim Beta\left(\frac{N-2}{2}, \frac{1}{2}\right), && k = 1, \ldots, k^* \\
& Y_{3,kv}^{\ell} \sim Beta\Big(a_k^{\ell} + b_{kv}^{\ell *}, b_{kv}^{\ell} - b_{kv}^{\ell *}\Big), && k = 1, \ldots, \left\lfloor \tfrac{p_\ell^*}{2} \right\rfloor;\ v = 1, \ldots, k_\ell; \\
& & & \qquad \ell = 1, \ldots, m \\
& Y_{3,v}^{\ell} \sim Beta\left(a_{p_\ell^*}^{\ell} + b_{p_\ell^* v}^{\ell *}, b_{p_\ell^* v}^{\ell} - b_{p_\ell^* v}^{\ell *}\right),
\end{aligned}
\tag{3.15}
$$

with r_k^+ given by (3.13) and (3.14), a_k^ℓ, b_{kv}^ℓ and $b_{kv}^{\ell *}$ given by (3.7) and $a_{p_\ell^*}^\ell$, $b_{p_\ell^* v}^\ell$ and $b_{p_\ell^* v}^{\ell *}$ given by (3.8).

This representation will enable us to develop well-fitting near-exact distributions, which bear an extreme closeness to the exact distribution of Λ.

4. NEAR-EXACT DISTRIBUTIONS FOR W AND Λ

To build the near-exact distributions of $W = -\log \Lambda$ and Λ we will leave $\Phi_1(t)$ in (3.9) and (3.12) unchanged and we will replace $\Phi_2(t)$ in (3.9) and (3.11) by a sharp asymptotic approximation in such a way that the resulting c.f. corresponds to a known manageable distribution.

From the results in Section 5 of [23], which show that we may asymptotically replace a $Logbeta(a, b)$ distribution by an infinite mixture of $\Gamma(b + j, a)$ distributions, with $j = 0, 1, \ldots$, using a somewhat heuristic approach, we will replace $\Phi_2(t)$ by

$$
\Phi_2^*(t) = \sum_{j=0}^{m^+} \pi_j \theta^{r+j} (\theta - it)^{-(r+j)}\,, \tag{4.1}
$$

which is the c.f. of a finite mixture of $\Gamma(r + j, \theta)$ distributions, where

$$
\begin{aligned}
r \;&=\; \tfrac{k^*}{2} + \sum_{\ell=1}^{m} \sum_{k=1}^{\lfloor p_\ell^*/2 \rfloor} \sum_{v=1}^{k_\ell} \tfrac{v-2k}{k_\ell} - \left\lfloor \tfrac{v-2k}{k_\ell} \right\rfloor \\
&\qquad + \sum_{\ell=1}^{m} \left(\sum_{v=1}^{k_\ell} \tfrac{2v-p_\ell^*-1}{2k_\ell} - \left\lfloor \tfrac{2v-p_\ell^*-1}{2k_\ell} \right\rfloor \right)^{p_\ell^* \bmod 2} \\
&=\; \tfrac{k^*}{2} + \sum_{\ell=1}^{m} \left\lfloor \tfrac{p_\ell^*+1}{2} \right\rfloor \tfrac{k_\ell - 1}{2}
\end{aligned}
\tag{4.2}
$$

is the sum of all the second parameters of the Beta r.v.'s in (3.15) and θ is obtained, together with s_1, s_2 and π^*, as the numerical solution of the system of equations

$$\frac{d^h}{dt^h}\Phi_2(t)\Big|_{t=0} = \frac{d^h}{dt^h}\Big(\pi^*\theta^{s_1}(\theta-it)^{-s_1} + (1-\pi^*)\theta^{s_2}(\theta-it)^{-s_2}\Big)\Big|_{t=0}, \quad h=1,\ldots,4, \tag{4.3}$$

that is, as the rate parameter of a mixture of two Gamma distributions with a common rate, which matches the first 4 derivatives of $\Phi_2(t)$, at $t=0$, so that $\Phi_1(t)\times\Big(\pi^*\theta^{s_1}(\theta-it)^{-s_1} + (1-\pi^*)\theta^{s_2}(\theta-it)^{-s_2}\Big)$ corresponds to a distribution that matches the first 4 exact moments of W. Then the weights π_j, $j=0,\ldots,m^+-1$ are determined in such a way that

$$\frac{d^h}{dt^h}\Phi_2(t)\Big|_{t=0} = \frac{d^h}{dt^h}\Phi_2^*(t)\Big|_{t=0}, \quad h=1,\ldots,m^+, \tag{4.4}$$

with $\pi_{m^+} = 1-\sum_{j=0}^{m^+-1}\pi_j$.

We will thus take as near-exact c.f. of W the c.f.

$$\begin{aligned}\Phi_W^*(t) &= \Phi_1(t)\times\Phi_2^*(t)\\ &= \left\{\prod_{k=2}^{p}\left(\frac{N-k}{N}\right)^{r_k^+}\left(\frac{N-k}{N}-it\right)^{-r_k^+}\right\}\\ &\qquad\times\left\{\sum_{j=0}^{m^+}\pi_j\,\theta^{r+j}\,(\theta-it)^{-(r+j)}\right\}\\ &= \sum_{j=0}^{m^+}\pi_j\left\{\theta^{r+j}\,(\theta-it)^{-(r+j)}\prod_{k=2}^{p}\left(\frac{N-k}{N}\right)^{r_k^+}\left(\frac{N-k}{N}-it\right)^{-r_k^+}\right\}\end{aligned} \tag{4.5}$$

with r and r_k^+ respectively given by (4.2) and (3.13), which is the c.f. of a mixture of m^++1 GNIG distributions of depth p that matches the first m^+ exact moments of W. This c.f. yields near-exact distributions for W with p.d.f.

$$f_W(w) = \sum_{j=0}^{m^+}\pi_j f^{GNIG}\Big(w\,\Big|\,r_2^+,r_3^+,...,r_p^+,r;\tfrac{N-2}{N},\tfrac{N-3}{N},...,\tfrac{N-p}{N},\theta;p\Big), \tag{4.6}$$

and c.d.f.

$$F_W(w) = \sum_{j=0}^{m^+}\pi_j F^{GNIG}\Big(w\,\Big|\,r_2^+,r_3^+,...,r_p^+,r;\tfrac{N-2}{N},\tfrac{N-3}{N},...,\tfrac{N-p}{N},\theta;p\Big), \tag{4.7}$$

for $w>0$, and near-exact distributions for Λ with p.d.f.

$$f_\Lambda(z) = \sum_{j=0}^{m^+}\pi_j f^{GNIG}\Big(-\log z\,\Big|\,r_2^+,r_3^+,...,r_p^+,r;\tfrac{N-2}{N},\tfrac{N-3}{N},...,\tfrac{N-p}{N},\theta;p\Big)\tfrac{1}{z}, \tag{4.8}$$

and c.d.f.

$$F_\Lambda(z)=\sum_{j=0}^{m^+}\pi_j\Big(1-F^{GNIG}\Big(-\log z\,\Big|\,r_2^+,r_3^+,...,r_p^+,r;\tfrac{N-2}{N},\tfrac{N-3}{N},...,\tfrac{N-p}{N},\theta;p\Big)\Big), \tag{4.9}$$

for $0<z<1$.

The modules for the GNIG c.d.f. and p.d.f. are available in [12] and on the web-page https://sites.google.com/site/nearexactdistributions/. Using these modules, the computation of the p.d.f.'s and c.d.f.'s of the near-exact distributions becomes easy and very manageable, once the system of equations in (4.4) is linear and as such very simple to solve, as it is also the case with the system of equations in (4.3). The authors make available a set of Mathematica® modules to implement the computation of p.d.f, c.d.f., p-values and quantiles for the near-exact distributions developed in the paper, as well as a module to compute the value of the l.r.t. statistic from a sample, on the web-page https://sites.google.com/site/nearexactdistributions/hyper-block-matrix-sphericity. In Appendix C the authors present a short manual for the use of these modules, along with some examples.

5. NUMERICAL STUDIES

In order to assess the performance of the near-exact distributions obtained in the previous section we will use the measure

$$\Delta^*=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\left|\frac{\Phi_W(t)-\Phi_W^*(t)}{t}\right|dt\,, \tag{5.1}$$

with

$$\max_w|F_W(w)-F_W^*(w)|=\max_z|F_\Lambda(z)-F_\Lambda^*(z)|\;\le\;\Delta^*\,,$$

where $\Phi_W(t)$ and $\Phi_W^*(t)$ represent respectively the exact and near-exact c.f.'s of W; $F_W(\,\cdot\,)$, $F_W^*(\,\cdot\,)$, the exact and near-exact c.d.f.'s of W and $F_\Lambda(\,\cdot\,)$ and $F_\Lambda^*(\,\cdot\,)$ those of Λ. Values for this measure Δ^*, which is therefore an upper bound on the difference between the exact and near-exact c.d.f.'s of both W and Λ for the near-exact distributions developed in the previous section, for different values of p_ℓ^* and k_ℓ, may be analyzed in Tables 1 and 2, with smaller values of the measure indicating even better agreements between the near-exact and the exact distributions.

From Tables 1 and 2 we may see the clear asymptotic behavior of the near-exact distributions not only for increasing sample sizes but also for increasing values of p_ℓ^* and k_ℓ, as well as the very good performance of the near-exact distributions for very small sample sizes, which barely exceed the number of variables in use.

Table 1: Values of the measure Δ^* in (5.1) for different values of m^+ (the number of exact moments matched by the near-exact distributions) and for $m = 4$ and increasing values of k_ℓ, p_ℓ^* and N, for base values of $k_\ell = \{3, 2, 3, 4\}$, $p_\ell^* = \{3, 5, 6, 4\}$ and sample sizes $N = p + 2, 200, 500$, with $p = \sum_{\ell=1}^{m} k_\ell \times p_\ell^*$.

N	m^+					
	0	1	2	4	6	10
$k_\ell + 0,\ p_\ell^* + 0,\ p = 53$						
$p+2$	7.06×10^{-6}	4.36×10^{-8}	1.33×10^{-12}	2.31×10^{-19}	1.30×10^{-26}	9.24×10^{-37}
$p+200$	7.40×10^{-7}	1.04×10^{-8}	3.21×10^{-14}	4.24×10^{-21}	9.44×10^{-28}	8.24×10^{-41}
$p+500$	1.60×10^{-7}	2.35×10^{-9}	1.52×10^{-15}	1.39×10^{-22}	5.96×10^{-30}	3.50×10^{-44}
$k_\ell + 2,\ p_\ell^* + 0,\ p = 89$						
$p+2$	2.24×10^{-6}	7.56×10^{-9}	1.67×10^{-13}	1.49×10^{-20}	1.51×10^{-27}	2.76×10^{-41}
$p+200$	5.28×10^{-7}	4.22×10^{-9}	4.99×10^{-14}	2.71×10^{-21}	1.49×10^{-28}	7.42×10^{-43}
$p+500$	1.35×10^{-7}	1.15×10^{-9}	6.45×10^{-15}	9.96×10^{-23}	1.50×10^{-30}	5.65×10^{-46}
$k_\ell + 5,\ p_\ell^* + 0,\ p = 143$						
$p+2$	8.66×10^{-7}	1.70×10^{-9}	2.33×10^{-14}	4.31×10^{-22}	8.30×10^{-30}	4.34×10^{-45}
$p+200$	3.75×10^{-7}	1.77×10^{-9}	2.13×10^{-14}	4.19×10^{-22}	8.01×10^{-30}	3.79×10^{-45}
$p+500$	1.16×10^{-7}	5.99×10^{-10}	3.99×10^{-15}	2.71×10^{-23}	1.77×10^{-31}	9.58×10^{-48}
$k_\ell + 2,\ p_\ell^* + 2,\ p = 129$						
$p+2$	1.14×10^{-6}	2.50×10^{-9}	1.58×10^{-14}	3.24×10^{-22}	7.86×10^{-30}	6.94×10^{-45}
$p+200$	4.34×10^{-7}	2.30×10^{-9}	8.28×10^{-15}	1.72×10^{-22}	3.85×10^{-30}	2.66×10^{-45}
$p+500$	1.28×10^{-7}	7.39×10^{-10}	1.19×10^{-15}	8.62×10^{-24}	6.44×10^{-32}	4.83×10^{-48}
$k_\ell + 2,\ p_\ell^* + 5,\ p = 189$						
$p+2$	6.62×10^{-7}	9.46×10^{-10}	2.22×10^{-15}	1.41×10^{-23}	1.27×10^{-31}	1.88×10^{-47}
$p+200$	3.84×10^{-7}	1.32×10^{-9}	2.92×10^{-15}	1.69×10^{-23}	2.33×10^{-31}	7.10×10^{-47}
$p+500$	1.35×10^{-7}	5.18×10^{-10}	2.47×10^{-16}	1.05×10^{-24}	6.62×10^{-33}	3.62×10^{-49}
$k_\ell + 5,\ p_\ell^* + 2,\ p = 207$						
$p+2$	3.75×10^{-7}	4.84×10^{-10}	1.88×10^{-15}	9.07×10^{-24}	4.51×10^{-32}	1.36×10^{-48}
$p+200$	2.38×10^{-7}	7.36×10^{-10}	2.92×10^{-15}	2.24×10^{-23}	1.66×10^{-31}	1.01×10^{-47}
$p+500$	8.99×10^{-8}	3.03×10^{-10}	7.05×10^{-16}	2.33×10^{-24}	7.17×10^{-33}	7.46×10^{-50}
$k_\ell + 5,\ p_\ell^* + 5,\ p = 303$						
$p+2$	1.75×10^{-7}	1.47×10^{-10}	1.53×10^{-16}	1.85×10^{-25}	2.43×10^{-34}	5.01×10^{-52}
$p+200$	1.52×10^{-7}	3.02×10^{-10}	3.43×10^{-16}	9.02×10^{-25}	2.43×10^{-33}	1.90×10^{-50}
$p+500$	6.79×10^{-8}	1.54×10^{-10}	1.04×10^{-16}	1.50×10^{-25}	2.12×10^{-34}	4.44×10^{-52}
$p+1000$	2.75×10^{-8}	6.65×10^{-11}	2.57×10^{-17}	1.70×10^{-26}	1.07×10^{-35}	4.24×10^{-54}

This asymptotic behavior is more marked for the near-exact distributions that match more exact moments. This may be seen from the more accentuated decrease in the values of the measure Δ^* for these near-exact distributions, that is, e.g. increases either in sample size or in the number of variables make the values of the measure Δ^* to decrease more for the near-exact distributions that match 10 exact moments than for those that match 6 exact moments. It is interesting to note that even near-exact distributions that match a very small number of exact moments or even no exact moment, and that, as such, are much simpler in their

structure, and faster to compute, exhibit these asymptotic behaviors, with the behavior of the near-exact distribution that matches no exact moment being absolutely remarkable. This latter one is a very simple near-exact distribution, for the computation of which we do not even need to solve the system of equations in (4.4). In this case we will have $m^+ = 0$, and from (4.5)-(4.9) it is easy to see that the near-exact distribution is just a GIG or a GNIG distribution, according to r in (4.2) being integer or not.

Table 2: Values of the measure Δ^* in (5.1) for different values of m^+ (the number of exact moments matched by the near-exact distributions) and for $m = 5$ and increasing values of k_ℓ, p^*_ℓ and N, for base values of $k_\ell = \{3, 2, 3, 4, 4\}$, $p^*_\ell = \{3, 5, 6, 4, 5\}$ and sample sizes $N = p + 2, 200, 500$, with $p = \sum_{\ell=1}^{m} k_\ell \times p^*_\ell$.

N	m^+					
	0	1	2	4	6	10
$k_\ell + 0,\ p^*_\ell + 0,\ p = 73$						
$p+2$	3.81×10^{-6}	1.61×10^{-8}	1.57×10^{-13}	1.14×10^{-21}	2.24×10^{-27}	4.53×10^{-40}
$p+200$	6.64×10^{-7}	6.61×10^{-9}	5.43×10^{-15}	2.03×10^{-21}	2.09×10^{-28}	9.91×10^{-43}
$p+500$	1.58×10^{-7}	1.66×10^{-9}	1.82×10^{-15}	7.74×10^{-23}	1.86×10^{-30}	3.79×10^{-46}
$k_\ell + 2,\ p^*_\ell + 0,\ p = 119$						
$p+2$	1.39×10^{-6}	3.35×10^{-9}	4.07×10^{-14}	1.39×10^{-21}	5.16×10^{-29}	1.00×10^{-43}
$p+200$	4.80×10^{-7}	2.78×10^{-9}	2.26×10^{-14}	6.90×10^{-22}	2.09×10^{-29}	2.40×10^{-44}
$p+500$	1.37×10^{-7}	8.61×10^{-10}	3.48×10^{-15}	3.46×10^{-23}	3.29×10^{-31}	3.67×10^{-47}
$k_\ell + 5,\ p^*_\ell + 0,\ p = 188$						
$p+2$	5.50×10^{-7}	7.93×10^{-10}	6.19×10^{-15}	4.93×10^{-23}	3.95×10^{-31}	3.17×10^{-47}
$p+200$	3.19×10^{-7}	1.10×10^{-9}	8.81×10^{-15}	9.96×10^{-23}	1.07×10^{-30}	1.42×10^{-46}
$p+500$	1.12×10^{-7}	4.31×10^{-10}	2.02×10^{-15}	9.19×10^{-24}	3.89×10^{-32}	7.89×10^{-49}
$k_\ell + 2,\ p^*_\ell + 2,\ p = 171$						
$p+2$	6.57×10^{-7}	1.05×10^{-9}	3.25×10^{-15}	2.49×10^{-23}	2.25×10^{-31}	2.55×10^{-47}
$p+200$	3.44×10^{-7}	1.33×10^{-9}	2.73×10^{-15}	2.87×10^{-23}	3.31×10^{-31}	5.50×10^{-47}
$p+500$	1.16×10^{-7}	4.92×10^{-10}	4.57×10^{-16}	1.95×10^{-24}	8.74×10^{-33}	2.10×10^{-49}
$k_\ell + 2,\ p^*_\ell + 5,\ p = 249$						
$p+2$	3.50×10^{-7}	3.68×10^{-10}	5.29×10^{-16}	1.36×10^{-24}	4.43×10^{-33}	7.49×10^{-50}
$p+200$	2.61×10^{-7}	6.53×10^{-10}	7.82×10^{-16}	3.81×10^{-24}	2.29×10^{-32}	1.18×10^{-48}
$p+500$	1.06×10^{-7}	2.98×10^{-10}	1.69×10^{-16}	4.17×10^{-25}	1.23×10^{-33}	1.43×10^{-50}
$k_\ell + 5,\ p^*_\ell + 2,\ p = 270$						
$p+2$	2.38×10^{-7}	2.28×10^{-10}	5.34×10^{-16}	1.13×10^{-24}	2.43×10^{-33}	1.28×10^{-50}
$p+200$	1.90×10^{-7}	4.31×10^{-10}	1.78×10^{-15}	4.88×10^{-24}	1.93×10^{-32}	3.20×10^{-49}
$p+500$	8.00×10^{-8}	2.06×10^{-10}	3.53×10^{-16}	7.31×10^{-25}	1.40×10^{-33}	5.37×10^{-51}
$k_\ell + 5,\ p^*_\ell + 5,\ p = 393$						
$p+2$	9.00×10^{-8}	5.69×10^{11}	2.98×10^{-17}	1.40×10^{-26}	7.15×10^{-36}	2.13×10^{-54}
$p+200$	9.26×10^{-8}	1.35×10^{-10}	8.75×10^{-17}	1.03×10^{-25}	1.25×10^{-34}	1.93×10^{-52}
$p+500$	4.72×10^{-8}	7.98×10^{-11}	3.32×10^{-17}	2.48×10^{-26}	1.82×10^{-35}	1.00×10^{-53}
$p+1000$	2.09×10^{-8}	3.81×10^{-11}	9.49×10^{-18}	3.62×10^{-27}	1.31×10^{-36}	1.68×10^{-55}

6. POWER STUDIES

In order to try to assess the behavior of the test under the alternative hypothesis, some power studies, based on simulations, were carried out. These studies focused on two forms of violation of the null hypothesis: (i) the violation of the equality of the Δ_ℓ matrices inside each block of k_ℓ of these matrices and (ii) the violation of the block-independence inside each group of $p_\ell = p_\ell^* \times k_\ell$ variables (see (1.2)).

First of all we should bring to the attention of the reader the fact that we are working with a random vector

$$\underline{X} = \left[\underline{X}_1', \ldots, \underline{X}_m'\right]',$$

where in turn, for $\ell = 1, \ldots, m$,

$$\underline{X}_\ell = \left[\underline{X}_{\ell 1}', \ldots, \underline{X}_{\ell k_\ell}'\right]',$$

with

$$\underline{X}_{\ell j} \sim N_{p_\ell^*}\left(\underline{\mu}_{\ell j}, \Delta_\ell\right) \qquad j = 1, \ldots, k_\ell$$

for some positive-definite matrix Δ_ℓ and some real $p_\ell^* \times 1$ vector $\underline{\mu}_{\ell j}$, and with

$$Cov\left(\underline{X}_{\ell j}, \underline{X}_{\ell' j'}\right) = 0_{p_\ell^* \times p_{\ell'}^*}$$

for either $\ell = \ell'$ or $\ell \neq \ell'$, with $j \neq j'$ if $\ell = \ell'$.

To keep things not too much involved, mainly in terms of easiness of exposition and to restrain the number of possible scenarios, while at the same time being able to give a view of a quite wide variety of situations under the alternative hypothesis, we considered a case with $m = 2$ and $k_1 = 2$ and $k_2 = 3$ with $p_1^* = 5$ and $p_2^* = 2$, with

$$\Delta_1 = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 \\ 1/2 & 2 & 2/3 & 2/4 & 2/5 \\ 1/3 & 2/3 & 3 & 3/4 & 3/5 \\ 1/4 & 2/4 & 3/4 & 4 & 4/5 \\ 1/5 & 2/5 & 3/5 & 4/5 & 5 \end{bmatrix} \qquad \text{and} \qquad \Delta_2 = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 2 \end{bmatrix},$$

where the choice of Δ_1 and Δ_2 did not obey to any other particular criteria than that of being two positive-definite matrices.

In the next two subsections we will perform power studies for the cases of violation of the hypothesis of equality of the diagonal blocks within each block of k_ℓ matrices and the hypothesis of independence generating for each scenario 1 000 000 pseudo random samples of size 29.

6.1. Violation of the equality hypothesis

In order to implement the violation of the hypothesis of equality of the diagonal blocks inside each $I_{k_\ell} \otimes \Delta_\ell$ block ($\ell = 1, 2$), we considered 40 different scenarios, with Σ covariance matrices of the form

$$\begin{bmatrix} \delta_{11}\Delta_1 & 0 & 0 & 0 & 0 \\ 0 & \delta_{12}\Delta_1 & 0 & 0 & 0 \\ 0 & 0 & \delta_{21}\Delta_2 & 0 & 0 \\ 0 & 0 & 0 & \delta_{22}\Delta_2 & 0 \\ 0 & 0 & 0 & 0 & \delta_{23}\Delta_2 \end{bmatrix}$$

with δ_{11}, δ_{12}, δ_{21}, δ_{22} and δ_{23} assuming the values in Table 3. In this Table are also defined the values for δ_1^* and δ_2^*. These new parameters summarize in a single value, respectively, the variability of the combinations of the values of δ_{11}, δ_{12} and δ_{21}, δ_{22}, δ_{23}. The values of δ_1^* and δ_2^* in Table 3 are defined based on the rank of the values of $\sum_{i=1}^{k_\ell} \left(1/\delta_{\ell i} - 1/\overline{\delta}_\ell\right)^2$, $(\ell = 1, 2; k_\ell = 2, 3)$, since this is for our purpose a more adequate measure of dispersion of the values of δ_{1j} and δ_{2j} $(j = 1, \ldots, k_\ell; \ell = 1, 2)$ than the usual variance. We will see that with this choice

Table 3: Definition of the values for δ_1^* and δ_2^*.

δ_1^*	$\Longleftrightarrow$	δ_{11}	δ_{12}	δ_2^*	$\Longleftrightarrow$	δ_{21}	δ_{22}	δ_{23}
1		1	1	1		1	1	1
2		1	2	2		1	1	2
3		1/2	2	3		1/2	1	2
4		1/3	2	4		1/2	1/2	2
5		1/3	3	5		1/3	1	2
				6		1/3	1	3
				7		1/3	1/3	2
				8		1/3	1/3	3

for the definition of the values of δ_1^* and δ_2^*, the power of the test will be an increasing function of the values of both δ_1^* and δ_2^*. The values δ_{11}, δ_{12} and the values δ_{21}, δ_{22}, δ_{23} were indeed chosen in such a way that they would generate a wide range of values of δ_1^* and δ_2^* that could show how the power of the test behaves for this variety of values. We should remark that for $\delta_1^* = 1$ and $\delta_2^* = 1$ we are under the null hypothesis in (1.2), while for any other combination of

values of δ_1^* and δ_2^* we will be under various forms of the alternative hypothesis due to the fact that for values of δ_1^* different from 1, the null hypothesis $H^1_{03|1,2}$ in (2.5) is violated, since in these cases we have $Var(\underline{X}_{11}) \neq Var(\underline{X}_{12})$, while for $\delta_2^* \neq 1$ it is the hypothesis $H^2_{03|1,2}$ in (2.5) that is violated, since for $\delta_2^* \neq 1$ we have at least two of $Var(\underline{X}_{21})$, $Var(\underline{X}_{22})$ or $Var(\underline{X}_{23})$ different. Increasing values of either δ_1^* or δ_2^* indicate a larger departure from H_0 in (1.2).

In Tables 4 and 5 we may analyze the power values for different values of δ_1^* and δ_2^*, respectively for $\alpha = 0.05$ and $\alpha = 0.01$. We may note how for $\delta_1^* = \delta_2^* = 1$, situation in which we are under the null hypothesis H_0 in (1.2), we obtain a value for power which coincides with the α value, showing the unbiasedness characteristic of the test. We may also see how power has a good rate of convergence towards 1 for increasing values of δ_1^* and δ_2^*.

Table 4: Power values, rounded to three decimal places, for $\alpha = 0.05$ and different values of δ_1^* and δ_2^*.

δ_1^*	δ_2^*							
	1	2	3	4	5	6	7	8
1	0.050	0.113	0.309	0.512	0.545	0.833	0.845	0.983
2	0.170	0.293	0.556	0.736	0.769	0.939	0.942	0.996
3	0.805	0.888	0.967	0.988	0.992	0.999	0.999	1.000
4	0.988	0.995	0.999	1.000	1.000	1.000	1.000	1.000
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 5: Power values, rounded to three decimal places, for $\alpha = 0.01$ and different values of δ_1^* and δ_2^*.

δ_1^*	δ_2^*							
	1	2	3	4	5	6	7	8
1	0.010	0.030	0.118	0.259	0.278	0.605	0.634	0.927
2	0.050	0.109	0.290	0.482	0.513	0.802	0.816	0.975
3	0.559	0.696	0.871	0.941	0.954	0.992	0.992	1.000
4	0.940	0.970	0.993	0.998	0.999	1.000	1.000	1.000
5	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000

In Figure 4 we present smoothed surface and line plots of the power values for the cases considered in Tables 4 and 5.

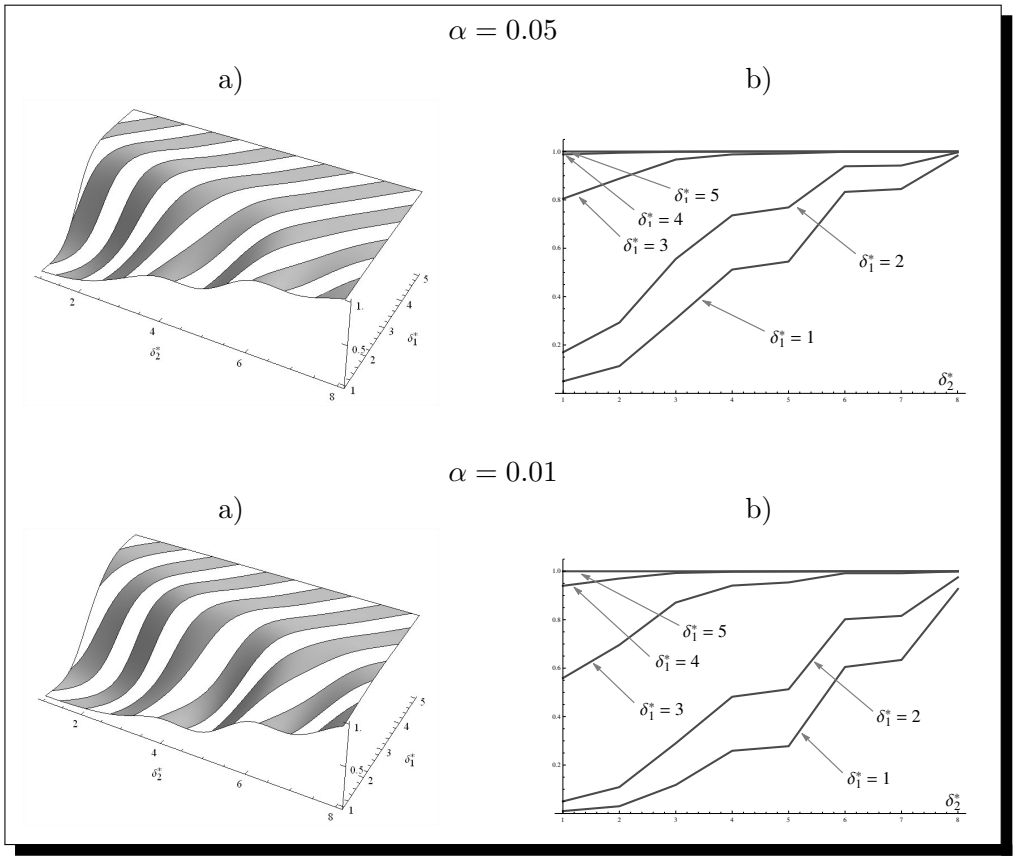

Figure 4: a) Smoothed surface plots and b) non-smoothed profile plots for different values of δ_1^* and running value of δ_2^*, for the values of power for the violation of the hypothesis of equality of diagonal blocks within each $I_{k_\ell} \otimes \Delta_\ell$ block ($\ell = 1, 2$).

6.2. Violation of the independence hypothesis

To implement the violation of the independence hypothesis, we consider 65 different scenarios with covariance matrices of the form

$$\begin{bmatrix} \Delta_1 & \gamma_1 C_1 & 0 & 0 & 0 \\ \gamma_1 C_1 & \Delta_1 & 0 & 0 & 0 \\ 0 & 0 & \Delta_2 & \gamma_{21} C_2 & \gamma_{22} C_2 \\ 0 & 0 & \gamma_{21} C_2 & \Delta_2 & \gamma_{23} C_2 \\ 0 & 0 & \gamma_{22} C_2 & \gamma_{23} C_2 & \Delta_2 \end{bmatrix}$$

where

$$C_1 = \frac{1}{10}\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 6 & 7 \\ 4 & 5 & 6 & 7 & 8 \\ 5 & 6 & 7 & 8 & 9 \end{bmatrix} \qquad \text{and} \qquad C_2 = \frac{1}{10}\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix},$$

and where γ_1 assumes the values $0.0, 1.0, 1.5, 1.75$ and 1.95 and γ_{21}, γ_{22} and γ_{23} assume the values in Table 6.

While for $\gamma_1 = 0$ we have the hypothesis $H^1_{02|1}$, of independence between $\underline{X}_{11}$ and $\underline{X}_{12}$, confirmed, for values of γ_1 different from zero we will be under the alternative hypothesis, since then the independence between these two sets of variables will be violated, with increasing values of γ_1 indicating an "increasing non-independence" of these two sets of variables, or equivalently, decreasing values of the determinant of the matrix

$$\Sigma_1 = \begin{bmatrix} \Delta_1 & \gamma_1 C_1 \\ \gamma_1 C_1 & \Delta_1 \end{bmatrix}.$$

In what concerns the values of γ_{21}, γ_{22} and γ_{23}, we will have the hypothesis of independence among $\underline{X}_{21}$, $\underline{X}_{22}$ and $\underline{X}_{23}$ confirmed when all these three parameters are equal to zero, and we will be under the alternative hypothesis if at least one of them is different from zero, with

$$\begin{aligned} \gamma_{21} \neq 0 &\implies Cov(\underline{X}_{21}, \underline{X}_{22}) = \gamma_{21} C_2 \neq 0\,, \\ \gamma_{22} \neq 0 &\implies Cov(\underline{X}_{21}, \underline{X}_{23}) = \gamma_{22} C_2 \neq 0\,, \\ \gamma_{23} \neq 0 &\implies Cov(\underline{X}_{22}, \underline{X}_{22}) = \gamma_{23} C_2 \neq 0\,. \end{aligned}$$

In order to define a hierarchy of the triplets of values of these three parameters, we compute the determinant of the matrix

$$\Sigma_2 = \begin{bmatrix} \Delta_2 & \gamma_{21} C_2 & \gamma_{22} C_2 \\ \gamma_{21} C_2 & \Delta_2 & \gamma_{23} C_2 \\ \gamma_{22} C_2 & \gamma_{23} C_2 & \Delta_2 \end{bmatrix}.$$

Values for $|\Sigma_2|$, as a function of the values of γ_{21}, γ_{22} and γ_{23} are shown in Table 6. These values are listed in decreasing order of $|\Sigma_2|$ and they are used to define the values of the new parameter γ^*, with increasing values of γ^* corresponding to decreasing values of $|\Sigma_2|$. The parameter γ^* is then used ahead in Tables 7 and 8 and Figure 5.

Then, while for $\gamma^* = 1$ we will be under the null hypothesis $H^2_{02|1}$ of independence among the sets of variables $\underline{X}_{21}$, $\underline{X}_{22}$ and $\underline{X}_{23}$, for increasing values of γ^* we will be increasingly further away from this null hypothesis.

We may see by looking at Tables 7 and 8 how the values for power give the value of α for $\gamma_1 = 0$ and $\gamma^* = 1$, situation in which we are under the null

hypothesis H_0 in (1.2), while for all other combinations of values of these two parameters the value of power increases as the values of γ^* and/or γ_1 increase, easily reaching 1.

Table 6: Definition of the values of γ^* for the different values of the parameters γ_{21}^*, γ_{22}^* and γ_{23}^*.

γ_{21}	0.0	3.0	3.5	4.0	3.0	3.0	3.5	4.5	4.0	4.5	3.5	4.95	4.95
γ_{22}	0.0	0.0	0.0	0.0	3.0	3.0	3.5	0.0	4.0	4.5	3.5	0.0	4.95
γ_{23}	0.0	0.0	0.0	0.0	3.0	0.0	3.5	0.0	4.0	4.5	0.0	0.0	4.95
$\lvert\Sigma_2\rvert$	5.359	3.405	2.706	1.904	1.843	1.479	1.121	1.001	0.534	0.142	0.105	0.105	0.001
γ^*	1	2	3	4	5	6	7	8	9	10	11	12	13

Table 7: Power values, rounded to three decimal places, for $\alpha = 0.05$ and different values of γ_1 and γ^*.

γ_1	γ^*												
	1	2	3	4	5	6	7	8	9	10	11	12	13
0.0	0.050	0.153	0.217	0.362	0.379	0.483	0.605	0.672	0.864	0.994	1.000	1.000	1.000
1.0	0.108	0.249	0.346	0.511	0.523	0.632	0.732	0.793	0.923	0.998	1.000	1.000	1.000
1.5	0.311	0.523	0.629	0.773	0.773	0.854	0.901	0.938	0.981	1.000	1.000	1.000	1.000
1.75	0.666	0.833	0.892	0.951	0.947	0.975	0.984	0.993	0.998	1.000	1.000	1.000	1.000
1.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 8: Power values, rounded to three decimal places, for $\alpha = 0.01$ and different values of γ_1 and γ^*.

γ_1	γ^*												
	1	2	3	4	5	6	7	8	9	10	11	12	13
0.0	0.010	0.039	0.070	0.144	0.161	0.222	0.338	0.383	0.659	0.967	0.996	0.996	1.000
1.0	0.027	0.086	0.137	0.247	0.265	0.351	0.474	0.531	0.771	0.984	0.999	0.999	1.000
1.5	0.115	0.257	0.350	0.509	0.521	0.626	0.725	0.785	0.916	0.997	1.000	1.000	1.000
1.75	0.378	0.591	0.689	0.818	0.816	0.887	0.923	0.953	0.986	1.000	1.000	1.000	1.000
1.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

In Figure 5 we present smoothed surface and line plots of the power values for the cases considered in Tables 6 and 7. From the plots in this Figure and the values in Tables 7 and 8 we may see how power attains the value 1 for the larger values of γ_1 and γ^*, as expected. We may also note that as in the case of the

previous subsection, for $\gamma_1 = 0$ and $\gamma^* = 1$, in which case we are under the null hypothesis, the value of the power equals the α value considered, showing again the unbiasedness characteristic of the test.

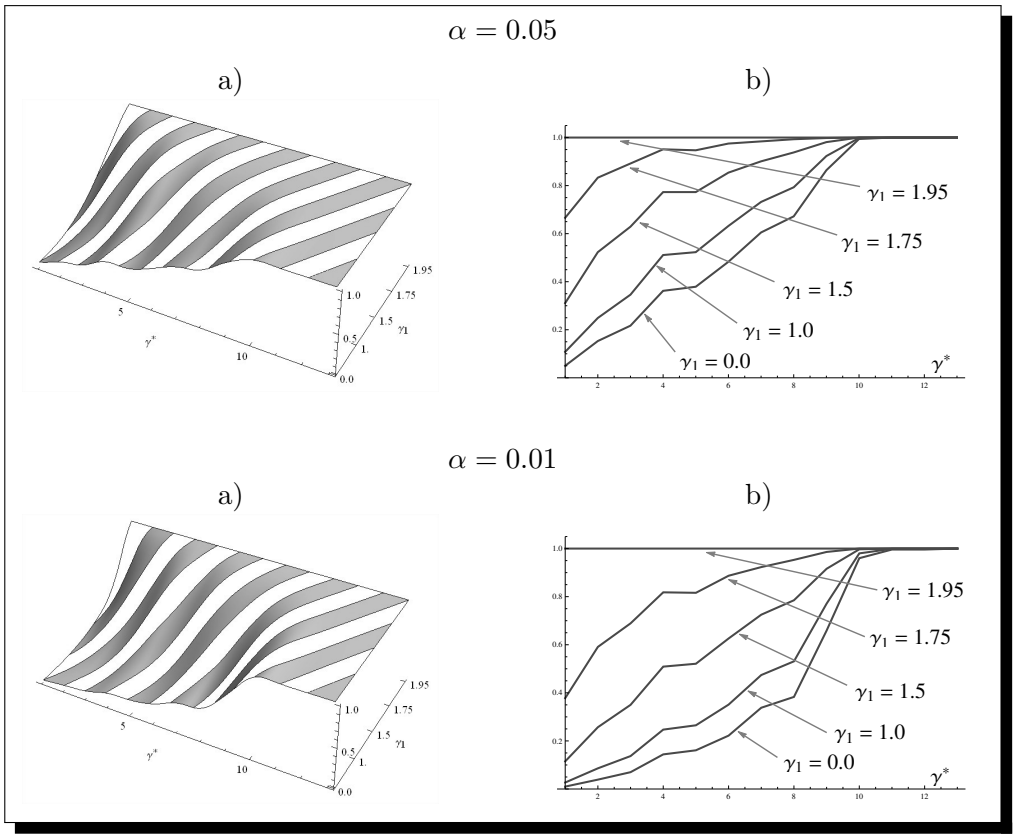

Figure 5: a) Smoothed surface plots; and b) non-smoothed profile plots, for the different values of γ_1 and running value of γ^*, for the values of power in Tables 7 and 8.

7. CONCLUSIONS

The procedure developed in this paper makes it possible to test elaborate covariance structures such as the HBM spherical structure through the use of very precise near-exact approximations. The testing procedure is based on an adequate decomposition of the overall null hypothesis into a sequence of sub-hypotheses, in our case the ones used to test the independence of several groups of variables and the equality of several covariance matrices in different sequences of covariance matrices. This decomposition of the null hypothesis allows us to

obtain the likelihood ratio test statistic, the expression of its h-th moment and the expression of the characteristic function of its logarithm. Furthermore, the suitable decomposition of the null hypothesis also induces a factorization of this characteristic function which is the basis for the development of the near-exact approximations. These approximations can be easily implemented since there are already computational modules available in the internet for the two main distributions involved, which are the GNIG and GIG distributions.

The high precision of the near-exact distributions, which was assessed in the numerical studies section, makes them an efficient tool to obtain p-values and quantiles for the test statistic, even in cases where the sample size is very small and/or the number of variables is large.

Power studies conducted through simulations show the unbiased nature of the test as well as its good power properties, reaching rapidly powers close to 1 in the different scenarios considered.

The procedure developed may be very useful to address other, eventually, more complex structures. A natural extension of this framework is to consider the same global structure but for complex Normal random variables or even for quaternion random variables. Other possible extension is to consider specific structures for the block-diagonal covariance matrices such as the circular or the compound symmetry structures.

ACKNOWLEDGMENTS

This research was supported by National Funds through FCT – Fundação para a Ciência e a Tecnologia, project UID/MAT/00297/2013 (CMA/UNL).

The authors would also like to thank the referee and the Associate Editor for their comments which helped in improving the readability of the paper.

APPENDICES

APPENDIX A – Obtaining the expression of the l.r.t. statistic Λ, associated with the null hypothesis H_0 in (1.2), by using the definition of likelihood ratio statistic

Let us consider the vector $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma\right)$ and let us suppose that we have a sample of size N from $\underline{X}$. The l.r.t statistic Λ associated with the HBM sphericity test is defined by

$$\Lambda = \frac{\sup L_0}{\sup L_1} \tag{A.1}$$

where L_0 is the likelihood function when the parameter space is under H_0 in (1.2) and L_1 is the likelihood function under the alternative hypothesis.

The likelihood function associated with the sample is

$$L\left(x_1, \ldots, x_N; \underline{\mu}; \Sigma\right) = \frac{1}{(2\pi)^{\frac{Np}{2}} \left|\Sigma\right|^{\frac{N}{2}}} e^{-\frac{1}{2}tr\left((\mathcal{X}-E_{N1}\underline{\mu}^T)\Sigma^{-1}(\mathcal{X}-E_{N1}\underline{\mu}^T)^T\right)}, \tag{A.2}$$

where E_{rs} denotes a matrix of 1's of dimension $r \times s$.

Let then $\mathcal{L}_0 = \mathcal{L}_0\left(x_1, \ldots, x_N; \underline{\mu}; \Sigma_{|H_o}\right) = \log\left(L_0\left(x_1, \ldots, x_N; \Sigma_{|H_o}\right)\right)$. From (A.2) we have

$$\begin{aligned}\mathcal{L}_0 = -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\log\left|\Sigma_{|H_o}\right| - \frac{1}{2}tr\left(\left(\mathcal{X}-E_{N1}\overline{\underline{x}}^T\right)\Sigma_{|H_o}^{-1}\left(\mathcal{X}-E_{N1}\overline{\underline{x}}^T\right)^T\right) \\ - \frac{1}{2}tr\left(\left(E_{N1}\overline{\underline{x}}^T-E_{N1}\underline{\mu}^T\right)\Sigma_{|H_o}^{-1}\left(E_{N1}\overline{\underline{x}}^T-E_{N1}\underline{\mu}^T\right)^T\right).\end{aligned}$$

As

$$tr\left(\left(\mathcal{X}-E_{N1}\overline{\underline{x}}^T\right)\Sigma_{|H_o}^{-1}\left(\mathcal{X}-E_{N1}\overline{\underline{x}}^T\right)^T\right) = tr\left(\Sigma_{|H_o}^{-1}\mathcal{A}\right)$$

where $\mathcal{A} = \left(\mathcal{X}-E_{N1}\overline{\underline{x}}^T\right)^T\left(\mathcal{X}-E_{N1}\overline{\underline{x}}^T\right) = \mathcal{X}^T\mathcal{X} - \frac{1}{N}\mathcal{X}^T E_{NN}\mathcal{X}$, and

$$\begin{aligned}tr\left(\left(E_{N1}\overline{\underline{x}}^T-E_{N1}\underline{\mu}^T\right)\Sigma_{|H_o}^{-1}\left(E_{N1}\overline{\underline{x}}^T-E_{N1}\underline{\mu}^T\right)^T\right) &= Ntr\left(\Sigma_{|H_o}^{-1}\left(\overline{\underline{x}}-\underline{\mu}\right)\left(\overline{\underline{x}}-\underline{\mu}\right)^T\right) \\ &= N\left(\overline{\underline{x}}-\underline{\mu}\right)^T\Sigma_{|H_o}^{-1}\left(\overline{\underline{x}}-\underline{\mu}\right)\end{aligned}$$

the function $\mathcal{L}_0$ can be written as

$$\mathcal{L}_0 = -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\log\left|\Sigma_{|H_o}\right| - \frac{1}{2}tr\left[\mathcal{A}\Sigma_{|H_o}^{-1}\right] - \frac{1}{2}N\left(\overline{\underline{x}}-\underline{\mu}\right)^T\Sigma_{|H_o}^{-1}\left(\overline{\underline{x}}-\underline{\mu}\right).$$

Given that

$$\begin{aligned}\left|\Sigma_{|H_o}\right| &= \left|bdiag\left(I_{k_\ell}\otimes\Delta_\ell,\ell=1,...,m\right)\right| = \prod_{\ell=1}^{m}\left|I_{k_\ell}\otimes\Delta_\ell\right| \\ &= \prod_{\ell=1}^{m}\left|I_{k_\ell}\right|^{p_\ell^*}\left|\Delta_\ell\right|^{k_\ell} = \prod_{\ell=1}^{m}\left|\Delta_\ell\right|^{k_\ell}\end{aligned}$$

where Δ_ℓ is a matrix of order p_ℓ^*, and

$$\begin{aligned}\Sigma_{|H_o}^{-1} &= \Big(bdiag\left(I_{k_\ell}\otimes\Delta_\ell,\ell=1,...,m\right)\Big)^{-1} \\ &= bdiag\left(\left(I_{k_\ell}\otimes\Delta_\ell\right)^{-1},\ell=1,...,m\Big)\right. \\ &= bdiag\left(I_{k_\ell}\otimes\Delta_\ell^{-1},\ell=1,...,m\right)\end{aligned}$$

with

$$\begin{aligned}tr\left(\mathcal{A}\Sigma_{|H_o}^{-1}\right) &= tr\Big(\mathcal{A}\,.\,bdiag\left(I_{k_\ell}\otimes\Delta_\ell^{-1},\ell=1,\ldots,m\right)\Big) \\ &= \sum_{\ell=1}^{m} tr\left(\mathcal{A}_{\ell\ell}\left(I_{k_\ell}\otimes\Delta_\ell^{-1}\right)\right) \\ &= \sum_{\ell=1}^{m}\sum_{v=1}^{k_\ell} tr\left(\mathcal{A}_{\ell\ell}^{v}\Delta_\ell^{-1}\right),\end{aligned}$$

we can write $\mathcal{L}_0$ as

$$\mathcal{L}_0 = -\frac{Np}{2}\log\left(2\pi\right) - \frac{N}{2}\log\left(\prod_{\ell=1}^{m}\left|\Delta_\ell\right|^{k_\ell}\right) - \frac{1}{2}\sum_{\ell=1}^{m}\sum_{v=1}^{k_\ell} tr\left(\mathcal{A}_{\ell\ell}^{v}\Delta_\ell^{-1}\right) - \frac{1}{2}N\left(\underline{\overline{x}}-\underline{\mu}\right)^T\Sigma_{|H_o}^{-1}\left(\underline{\overline{x}}-\underline{\mu}\right).$$

By solving the system of likelihood equations

$$\left\{\begin{array}{l}\left.\frac{\partial\mathcal{L}_0}{\partial\underline{\mu}}\right|_{\underline{\mu}=\underline{\widehat{\mu}}}=0 \\ \left.\frac{\partial\mathcal{L}_0}{\partial\Delta_\ell}\right|_{\substack{\Delta_\ell=\widehat{\Delta}_\ell \\ \underline{\mu}=\underline{\widehat{\mu}}}}=0\,,\ (\ell=1,...,m)\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}N\left(\underline{\overline{x}}-\underline{\widehat{\mu}}\right)\Sigma_{|H_o}^{-1}=0 \\ \\ \text{———}\end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{l}\underline{\widehat{\mu}}=\underline{\overline{x}} \\ -\frac{N}{2}k_\ell\widehat{\Delta}_\ell^{-1}+\frac{1}{2}\widehat{\Delta}_\ell^{-1}\sum_{v=1}^{k_\ell}A_{\ell\ell}^{v}\widehat{\Delta}_\ell^{-1}=0\,, \\ \qquad\qquad\qquad\qquad (\ell=1,...,m)\end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{l}\underline{\widehat{\mu}}=\underline{\overline{x}} \\ \\ \widehat{\Delta}_\ell=\frac{1}{Nk_\ell}\sum_{v=1}^{k_\ell}A_{\ell\ell}^{v}\,,\quad(\ell=1,...,m)\end{array}\right.$$

we obtain the maximum likelihood estimators of $\underline{\mu}$ and Σ under H_0, which are, $\widehat{\underline{\mu}} = \overline{\underline{X}}$ and

$$\widehat{\Sigma} = \begin{bmatrix} I_{k_1} \otimes \widehat{\Delta}_1 & & 0 \\ & \ddots & \\ 0 & & I_{k_m} \otimes \widehat{\Delta}_m \end{bmatrix} = \begin{bmatrix} I_{k_1} \otimes \frac{1}{Nk_1} A_1^* & & 0 \\ & \ddots & \\ 0 & & I_{k_m} \otimes \frac{1}{Nk_m} A_m^* \end{bmatrix}$$

where $A_\ell^* = \sum_{v=1}^{k_\ell} A_{\ell\ell}^v$ $(\ell = 1, ..., m)$.

Then we have

$$\begin{aligned}
&\sup L_0\left(X_1, \ldots, X_N; \underline{\mu}; \Sigma_{|H_o}\right) \\
&\quad = (2\pi)^{-\frac{Np}{2}} \left\{ \prod_{\ell=1}^{m} |\widehat{\Delta}_\ell|^{k_\ell} \right\}^{-\frac{N}{2}} e^{-\frac{1}{2} tr\left(A.bdiag\left(I_{k_\ell} \otimes \widehat{\Delta}_\ell^{-1}, \ell=1,\ldots,m\right)\right)} \\
&\quad = (2\pi)^{-\frac{Np}{2}} N^{\frac{pN}{2}} \prod_{\ell=1}^{m} k_\ell^{\frac{p_\ell N}{2}} \prod_{\ell=1}^{m} |A_\ell^*|^{-\frac{k_\ell N}{2}} e^{-\frac{N}{2} \sum_{\ell=1}^{m} k_\ell tr\left(A_{\ell\ell}\left(I_{k_\ell} \otimes \left(A_\ell^*\right)^{-1}\right)\right)} \\
&\quad = (2\pi)^{-\frac{Np}{2}} N^{\frac{pN}{2}} \prod_{\ell=1}^{m} k_\ell^{\frac{p_\ell N}{2}} \prod_{\ell=1}^{m} |A_\ell^*|^{-\frac{k_\ell N}{2}} e^{-\frac{N}{2} \sum_{\ell=1}^{m} k_\ell \sum_{v=1}^{k_\ell} tr\left(A_{\ell\ell}^v \left(A_\ell^*\right)^{-1}\right)} \\
&\quad = (2\pi)^{-\frac{Np}{2}} N^{\frac{pN}{2}} \prod_{\ell=1}^{m} k_\ell^{\frac{p_\ell N}{2}} \prod_{\ell=1}^{m} |A_\ell^*|^{-\frac{k_\ell N}{2}} e^{-\frac{N}{2} \sum_{\ell=1}^{m} k_\ell tr\left(A_\ell^* \left(A_\ell^*\right)^{-1}\right)} \\
&\quad = (2\pi)^{-\frac{Np}{2}} N^{\frac{pN}{2}} \prod_{\ell=1}^{m} k_\ell^{\frac{p_\ell N}{2}} \prod_{\ell=1}^{m} |A_\ell^*|^{-\frac{k_\ell N}{2}} e^{-\frac{N}{2} \sum_{\ell=1}^{m} k_\ell tr\left(I_{p_\ell^*}\right)} \\
&\quad = (2\pi)^{-\frac{Np}{2}} N^{\frac{pN}{2}} \prod_{\ell=1}^{m} k_\ell^{\frac{p_\ell N}{2}} \prod_{\ell=1}^{m} |A_\ell^*|^{-\frac{k_\ell N}{2}} e^{-\frac{N}{2} \sum_{\ell=1}^{m} p_\ell} \\
&\quad = (2\pi)^{-\frac{Np}{2}} N^{\frac{pN}{2}} \prod_{\ell=1}^{m} k_\ell^{\frac{p_\ell N}{2}} \prod_{\ell=1}^{m} |A_\ell^*|^{-\frac{k_\ell N}{2}} e^{-\frac{Np}{2}}.
\end{aligned} \tag{A.3}$$

Under H_1, the likelihood function is given by

$$\begin{aligned}
L_1\left(x_1, \ldots, x_N; \underline{\mu}; \Sigma\right) &= (2\pi)^{-\frac{Np}{2}} |\Sigma|^{-\frac{N}{2}} e^{-\frac{1}{2} tr\left(\left(\mathcal{X} - E_{N1}\underline{\mu}^T\right)^T \left(\mathcal{X} - E_{N1}\underline{\mu}^T\right)\Sigma^{-1}\right)} \\
&= (2\pi)^{-\frac{Np}{2}} |\Sigma|^{-\frac{N}{2}} e^{-\frac{1}{2} tr\left(\mathcal{A}\Sigma^{-1}\right)}
\end{aligned}$$

and $\mathcal{L}_1 = \mathcal{L}_1\left(x_1, \ldots, x_N; \underline{\mu}; \Sigma\right) = \log\left(L_1\left(x_1, \ldots, x_N; \underline{\mu}; \Sigma\right)\right)$ is given by

$$\mathcal{L}_1 = -\frac{Np}{2} \log(2\pi) + \frac{N}{2} \log|\Sigma| - \frac{1}{2} tr\left[\mathcal{A}\Sigma^{-1}\right] - \frac{1}{2} N\, tr\left(\left(\overline{\underline{x}} - \underline{\mu}\right)^T \left(\overline{\underline{x}} - \underline{\mu}\right) \Sigma^{-1}\right).$$

By solving the system of likelihood equations

$$\begin{cases} \left.\frac{\partial \mathcal{L}_1}{\partial \underline{\mu}}\right|_{\underline{\mu} = \widehat{\underline{\mu}}} = 0 \\ \left.\frac{\partial \mathcal{L}_1}{\partial \Sigma}\right|_{\substack{\Sigma = \widehat{\Sigma} \\ \underline{\mu} = \widehat{\underline{\mu}}}} = 0 \end{cases} \Leftrightarrow \begin{cases} N\left(\overline{\underline{x}} - \widehat{\underline{\mu}}\right)\Sigma^{-1} = 0 \\ -\frac{N}{2}\widehat{\Sigma}^{-1} + \frac{1}{2}\widehat{\Sigma}^{-1} A \widehat{\Sigma}^{-1} = 0 \end{cases} \Leftrightarrow \begin{cases} \widehat{\underline{\mu}} = \overline{\underline{x}} \\ \widehat{\Sigma} = \frac{1}{N} A \end{cases}$$

we conclude that

$$\begin{aligned}
\sup L_1\left(X_1,\ldots,X_N;\underline{\mu};\Sigma\right) &= (2\pi)^{-\frac{Np}{2}}\,|\widehat{\Sigma}|^{-\frac{N}{2}}\,e^{-\frac{1}{2}tr\left(A\widehat{\Sigma}^{-1}\right)} \\
&= (2\pi)^{-\frac{Np}{2}}\,\left|\tfrac{1}{N}A\right|^{-\frac{N}{2}}\,e^{-\frac{1}{2}tr\left(NAA^{-1}\right)} \\
&= (2\pi)^{-\frac{Np}{2}}\,N^{\frac{Np}{2}}\,|A|^{-\frac{N}{2}}\,e^{-\frac{1}{2}N\,tr(I_p)} \\
&= (2\pi)^{-\frac{Np}{2}}\,N^{\frac{Np}{2}}\,|A|^{-\frac{N}{2}}\,e^{-\frac{1}{2}Np}
\end{aligned} \tag{A.4}$$

Then, from (A.1), (A.3) and (A.4) we have

$$\begin{aligned}
\Lambda &= \frac{\sup L_0\left(x_N;\underline{\mu};\Sigma_{|H_o}\right)}{\sup L_1\left(x_N;\underline{\mu};\Sigma\right)} \\
&= \frac{(2\pi)^{-\frac{Np}{2}}\,N^{\frac{Np}{2}}\left\{\prod_{\ell=1}^{m}k_\ell^{\frac{Np_\ell}{2}}\right\}\left\{\prod_{\ell=1}^{m}|A_\ell^*|^{-\frac{k_\ell N}{2}}\right\}e^{-\frac{Np}{2}}}{(2\pi)^{-\frac{Np}{2}}\,N^{\frac{Np}{2}}\,|A|^{-\frac{N}{2}}\,e^{-\frac{Np}{2}}} \\
&= \left\{\prod_{\ell=1}^{m}k_\ell^{\frac{Np_\ell}{2}}\right\}\frac{|A|^{\frac{N}{2}}}{\prod_{\ell=1}^{m}\left|A_\ell^*\right|^{\frac{k_\ell N}{2}}}
\end{aligned} \tag{A.5}$$

where the matrix A is the maximum likelihood estimator of Σ, $A_{\ell\ell}$ is the ℓ-th diagonal block of order $p_\ell = k_\ell \times p_\ell^*$ of A $(\ell = 1,...,m)$, with $p = \sum_{\ell=1}^{m} p_\ell$ and $A_\ell^* = A_{\ell\ell}^1 + \cdots + A_{\ell\ell}^{k_\ell}$, where $A_{\ell\ell}^v$ is the v-th $(v = 1,...,k_\ell)$ diagonal block of order p_ℓ^* of $A_{\ell\ell}$. We should note how expression (A.5) is the same as expression (2.11).

Appendix B – Shape parameters

The shape parameters $r_{3;k}^{\ell}$ in (3.5) are given by

$$\text{(B.1)}\quad r_{3,k}^{\ell}=\begin{cases} r_k^{\ell *}, & k=1,...,p_\ell^*-1, \\ & k\neq p_\ell^*-1-2\alpha_1^\ell \\ r_k^{\ell *}+(p_\ell^* \bmod 2)\left(\alpha_2^\ell-\alpha_1^\ell\right)\left(k_\ell-\frac{p_\ell^*-1}{2}+k_\ell\left\lfloor\frac{p_\ell^*}{2k_\ell}\right\rfloor\right), & k=p_\ell^*-1-2\alpha_1^\ell \end{cases}$$

where

$$\alpha^\ell=\left\lfloor\frac{p_\ell^*-1}{k_\ell}\right\rfloor,\ \alpha_1^\ell=\left\lfloor\frac{k_\ell-1}{k_\ell}\frac{p_\ell^*-1}{2}\right\rfloor,\ \alpha_2^\ell=\left\lfloor\frac{k_\ell-1}{k_\ell}\frac{p_\ell^*+1}{2}\right\rfloor,$$

and

$$r_k^{\ell *}=\begin{cases} c_k^\ell, & k=1,...,\alpha^\ell+1 \\ k_\ell\left(\left\lfloor\frac{p_\ell^*}{2}\right\rfloor-\left\lfloor\frac{k}{2}\right\rfloor\right), & k=\alpha^\ell+2,...,\min\left(p_\ell^*-2\alpha_1^\ell,p_\ell^*-1\right) \\ & \text{and } k=2+p_\ell^*-2\alpha_1^\ell,...,2\left\lfloor\frac{p_\ell^*}{2}\right\rfloor-1, step\ 2 \\ k_\ell\left(\left\lfloor\frac{p_\ell^*+1}{2}\right\rfloor-\left\lfloor\frac{k}{2}\right\rfloor\right), & k=1+p_\ell^*-2\alpha_1^\ell,...,p_\ell^*-1, step\ 2 \end{cases}$$

with

$$c_k=\left\lfloor\frac{k_\ell}{2}\right\rfloor\left((k-1)\,k_\ell-2\left((k_\ell+1)\bmod 2\right)\left\lfloor\frac{k}{2}\right\rfloor\right)+\left\lfloor\frac{k_\ell}{2}\right\rfloor\left\lfloor\frac{k_\ell+k \bmod 2}{2}\right\rfloor,$$

for $k=1,...,\alpha^\ell$, and

$$\begin{aligned} c_{\alpha+1}^\ell=&-\left(\left\lfloor\frac{p_\ell^*}{2}\right\rfloor-\alpha^\ell\left\lfloor\frac{k_\ell}{2}\right\rfloor\right)^2+k_\ell\left(\left\lfloor\frac{p_\ell^*}{2}\right\rfloor-\left\lfloor\frac{\alpha^\ell+1}{2}\right\rfloor\right) \\ &+(k_\ell \bmod 2)\left(\alpha^\ell\left\lfloor\frac{p_\ell^*}{2}\right\rfloor+\frac{\left(\alpha^\ell \bmod 2\right)}{4}-\frac{\left(\alpha^\ell\right)^2}{4}-\left(\alpha^\ell\right)^2\left\lfloor\frac{k_\ell}{2}\right\rfloor\right). \end{aligned}$$

For the derivation of the expressions for these parameters see [14] and references therein, making the necessary adjustments.

APPENDIX C – The Mathematica® modules

The modules described in this Appendix are available at the web-page https://sites.google.com/site/nearexactdistributions/hyper-block-matrix-sphericity and may be downloaded from this web-page.

C.1 - Computation of the p.d.f. and c.d.f. of the near-exact distributions

The modules made available for the computation of the p.d.f. and c.d.f. of the near-exact distributions for Λ are called, respectively, `NEpdf` and `NEcdf`. These modules have 4 mandatory arguments, which are:

- the sample size,
- a list with the values of p_ℓ^* $(\ell = 1, \ldots, m)$,
- a list with the values of k_ℓ $(\ell = 1, \ldots, m)$,
- the running value where the p.d.f. or c.d.f. is to be computed,

and which have to be given in this order; and 5 optional arguments, which are:

- `nm`: the number of exact moments to be matched by the near-exact distribution, that is, the value of m^+ in (4.1) and (4.4)–(4.9) (default value: 4),
- `prec1`: the number of digits used to print the value of the p.d.f. or c.d.f. (default value: 10),
- `prec2`: the number of precision digits used in the computation of the p.d.f. or c.d.f. (default value: 200),
- `prec3`: the number of precision digits used to store the m^+ exact moments of $W = -\log \Lambda$ computed (default value: 200),
- `prec4`: number of precision digits used in the computation of the m^+ exact moments of $W = -\log \Lambda$ by the module that does this computation (default value: 1500).

These optional arguments may be given in any order, but they will have to be called by their names, as it is exemplified below. If not used, they will assume their default values.

These modules use a number of other modules available on the same web-page which compute the weights π_j and the rate parameter θ in (4.1), the shape and rate parameters in $\Phi_1(t)$ as well as other shape and rate parameters involved in the expressions of the near-exact p.d.f. and c.d.f.. The module that computes the weights π_j uses another module which computes the exact moments of $W = -\log \Lambda$ by applying a numerical method to the exact c.f. of W in (3.1).

For example to compute the near-exact c.d.f. of Λ, on a value near the 0.05 quantile, for a case with the same parameters as those used for the examples for which we computed power in Section 6, which was a case with $m = 2$, $p_1^* = 5$, $p_2^* = 2$, $k_1 = 2$ and $k_2 = 3$, using the default values for all optional arguments, we would use the first command in Figure 6. The second command in that same figure uses the option `prec1` in order to obtain an output with more digits. The options named `prec2`, `prec3`, and `prec4`, will usually not be necessary, unless one suspects from lack of precision in the result obtained, which may happen in cases where the number of variables or the sample size are very large. This fact is illustrated with the third command in Figure 6, where although 500 precision digits are requested for the internal representation of the exact moments of W, the result obtained is exactly the same as the one obtained with the second command. The fourth command in Figure 6 illustrates, together with the third one that the order in which the optional arguments are given is arbitrary.

```
NEcdf[29, {5, 2}, {2, 3}, 5.914780554*10^-44]
0.05000000000
NEcdf[29, {5, 2}, {2, 3}, 5.914780554*10^-44, prec1→30]
0.049999999999933411511569283887674
NEcdf[29, {5, 2}, {2, 3}, 5.914780554*10^-44, prec1→30, prec3→500]
0.049999999999933411511569283887674
NEcdf[29, {5, 2}, {2, 3}, 5.914780554*10^-44, prec3→500, prec1→30]
0.049999999999933411511569283887674
```

Figure 6: Mathematica® commands to be used with the module `NEcdf`.

We remark that, for a given level α, we should reject the null hypothesis when the computed value of the l.r.t. statistic is lower than the α-quantile of the l.r.t. statistic. As such, the computation of the c.d.f. for the l.r.t. statistic also gives automatically the p-value.

We may note the extremely low values that these quantiles attain. This is due to the fact that we chose to use the 'complete' l.r.t. statistic, that is, the l.r.t. statistic with its exponent $N/2$. This is indeed the case why some authors chose to use l.r.t. statistics without this exponent, to make these values not so close to zero, what in some cases may cause some numerical problems. But indeed this poses absolutely no problems to the computation of the near-exact p.d.f.'s or c.d.f.'s.

The computation of quantiles is done with the module `Quant`. Given the sample size and the values for p_ℓ^* and k_ℓ ($\ell = 1, \ldots, m$), the module generates, by default, 10 pseudo-random samples, under the null hypothesis of hyper-block sphericity in (1.2), using then the empirical α-quantile as a 'starting value' for a Newton-type method, which will find the approximate near-exact quantile using the values of the near-exact p.d.f. and c.d.f. computed on the successive iteration values.

This module has 5 mandatory arguments which first one is the α value for the quantile and which last 4 are exactly the same as the 4 mandatory arguments for the modules `NEpdf` and `NEcdf`, given in the exact same order. This module also has 8 optional arguments, which are:

- `nm`: the number of exact moments to be matched by the near-exact distribution, that is, as for `NEpdf` and `NEcdf`, the value of m^+ in (4.1) and (4.4)–(4.9) (default value: 4),
- `prec1`: the number of digits used to print the value of the quantile (default value: 10),
- `prec2`: the number of precision digits used in the computation of the p.d.f. or c.d.f. for the implementation of a Newton-type method (default value: 400),
- `prec3`: the number of precision digits used to store the m^+ exact moments of $W = -\log \Lambda$ computed (default value: 200),
- `prec4`: number of precision digits used in the computation of the m^+ exact moments of $W = -\log \Lambda$ by the module that does this computation (default value: 1500),
- `eps`: the value of the minimum upper-bound for two consecutive quantile approximations obtained from the Newton-type method; if those two consecutive approximations differ a quantity that is less than `eps`, the process stops, giving as result the last approximation found (default value: 10^{-6} times the 'starting value'),
- `nsamp`: the number of pseudo-random samples generated by the module to obtain the 'starting value' (default value: 10).

In Figure 7 we present a few commands that may be used with the module `Quant` to compute the 0.05 quantile of Λ for the same scenario considered in Figure 6. The first command uses all optional arguments with their default values, which will be adequate for most cases. The second command uses the optional argument `prec1` to request 20 digits, instead of 10, for the approximate 0.05 quantile. We may see that when this second command is repeated, as the third command in Figure 7, the result obtained is different. There is indeed no problem, and for the attentive reader there should be not much of a surprise. What happens is that since we use for `eps` its default value, the precision obtained for the approximation of the quantile should ensure at least 6 decimal digits correct. This is exactly what happens. Indeed it seems that at least 11 digits are correct. Then the fourth and fifth commands give the same result, which should be correct for all digits displayed. They illustrate the fact that the order in which the optional arguments are given is arbitrary and also that by giving the optional argument `eps` a small enough value, in this case a value which would ensure that at least 21 digits of the approximate quantile are correct, we will always get the same result.

```
Quant[5 / 100, 29, {5, 2}, {2, 3}]
5.914780554 × 10^-44
Quant[5 / 100, 29, {5, 2}, {2, 3}, prec1 → 20]
5.9147805544727498310 × 10^-44
Quant[5 / 100, 29, {5, 2}, {2, 3}, prec1 → 20]
5.9147805544731374828 × 10^-44
Quant[5 / 100, 29, {5, 2}, {2, 3}, prec1 → 20, eps → 10^-65]
5.9147805544731417794 × 10^-44
Quant[5 / 100, 29, {5, 2}, {2, 3}, eps → 10^-65, prec1 → 20]
5.9147805544731417794 × 10^-44
```

Figure 7: Mathematica® commands to be used with the module `Quant`.

There is also another module called `Lambda`, which may be used to compute the value of the statistic Λ in (2.11) for a given dataset. This dataset has to be given in a file, with observations defining the rows and variables the columns. This module has 3 mandatory arguments, which are:

- the name of the data file (including the path),
- a list with the values of p_ℓ^* $(\ell = 1, \ldots, m)$,
- a list with the values of k_ℓ $(\ell = 1, \ldots, m)$,

and which have to be given in this order.

Further details on these modules and their use are available at the web-page https://sites.google.com/site/nearexactdistributions/hyper-block-matrix-sphericity.

REFERENCES

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

[2] Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria, *Biometrika*, **36**(3/4), 317–346.

[3] Cardeño, L. and Nagar, D.K. (2001). Testing block sphericity of a covariance matrix, *Divulgaciones Matemáticas*, **9**, 25–34.

[4] Chao, C.C. and Gupta, A.K. (1991). Testing of homogeneity of diagonal blocks with blockwise independence, *Communications in Statistics – Theory and Methods*, **20**, 1957–1969.

[5] Chen, P.; Dong, L.; Chen, W. and Lin, J.-G. (2013). Outlier Detection in Adaptive Functional-Coefficient Autoregressive Models Based on Extreme Value Theory, *Mathematical Problems in Engineering*, **2013**, Article ID 910828, 9 pp..

[6] Coelho, C.A. (1998). The Generalized Integer Gamma distribution – a basis for distributions in Multivariate Statistics, *Journal of Multivariate Analysis*, **64**(1), 86–102.

[7] Coelho, C.A. (2004). The generalized near-integer Gamma distribution: a basis for near-exact approximations to the distribution of statistics which are the product of an odd number of independent Beta random variables, *Journal of Multivariate Analysis*, **89**(2), 191–218.

[8] Coelho, C.A. and Marques, F.J. (2009). The advantage of decomposing elaborate hypotheses on covariance matrices into conditionally independent hypotheses in building near-exact distributions for the test statistics, *Linear Algebra and its Applications*, **430**(10), 2592–2606.

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

[10] Jensen, S.T. (1988). Covariance hypotheses which are linear in both the covariance and the inverse covariance, *Annals of Statistics*, **16**(1), 302–322.

[11] Kopylev, L. (2012). Constrained Parameters in Applications: Review of Issues and Approaches, *ISRN Biomathematics*, **2012**, Article ID 872956, 6 pp..

[12] Marques, F.J. and Coelho, C.A. (2008). Near-exact distributions for the sphericity likelihood ratio test statistic, *Journal of Statistical Planning and Inference*, **138**(3), 726–741.

[13] Marques, F.J. and Coelho, C.A. (2012). The block sphericity test – exact and near-exact distributions for the likelihood ratio statistic, *Mathematical Methods in the Applied Sciences*, **35**(4), 373–383.

[14] Marques, F.J.; Coelho, C.A. and Arnold, B.C. (2011). A general near-exact distribution theory for the most common likelihood ratio test statistics used in Multivariate Analysis, *TEST*, **20**(1), 180–203.

[15] Marques, F.J.; Coelho, C.A. and Marques, P. (2013). *The Block-Matrix Sphericity Test: Exact and Near-exact Distributions for the Test Statistic*. In "Recent Developments in Modeling and Applications in Statistics – Studies in Theoretical and Applied Statistics: Selected Papers of the Statistical Societies" (P.E. Oliveira, M.G. Temido, C. Henriques and M. Vichi, Eds.), Springer, 169–177.

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

[17] Muirhead, R.J. (2005). *Aspects of Multivariate Statistical Theory*, 2nd ed., J. Wiley & Sons, New York.

[18] Nagar, D.K.; Jain, S.K. and Gupta, A.K. (1985). Distribution of LRC for testing sphericity of a complex multivariate Gaussian model, *International Journal of Mathematics and Mathematical Sciences*, **8**(3), 555–562.

[19] Nagarsenker, B.N. and Nagarsenker, P.B. (1981). Distribution of the likelihood ratio statistic for testing sphericity structure for a complex normal covariance matrix, *Sankhya*, Ser. B, **43**(3), 352–359.

[20] Nandakumar, K.; Chen, Y.; Dass, S.C. and Jain, A. (2008). Likelihood Ratio-Based Biometric Score Fusion, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(2), 342–347.

[21] Pirie, A.; Wood, A.; Lush, M.; Tyrer, J. and Pharoah, P.D. (2015). The effect of rare variants on inflation of the test statistics in case-control analyses, *BMC Bioinformatics*, **16**(53), 5 pp.

[22] Rencher, A.C. and Christensen, W.F. (2012). *Methods of Multivariate Analysis*, 3rd ed., J. Wiley & Sons, New York.

[23] Tricomi, F.G. and Erdélyi, A. (1951). The asymptotic expansion of a ratio of gamma functions, *Pacific Journal of Mathematics*, **1**(1), 133–142.

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

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

[26] Wilks, S.S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses, *The Annals of Mathematical Statistics*, **9**(1), 60–62.

[27] Zhang, J.; Zou, C. and Wang, Z. (2010). A control chart based on likelihood ratio test for monitoring process mean and variability, *Quality and Reliability Engineering International*, **26**(1), 63–73.

REVSTAT – STATISTICAL JOURNAL

Background

Statistics Portugal (INE, I.P.), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, publishing three times a year papers containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided to publish papers also in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work.

At the time, the Editorial Board was mainly composed by Portuguese university professors, being now composed by national and international university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal.

In 2001, the *Revista de Estatística* published three volumes special issue containing extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

The name of the Journal has been changed to REVSTAT - STATISTICAL JOURNAL, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

— The editorial policy will focus on publishing research articles at the highest level in the domains of Probability and Statistics with emphasis on the originality and importance of the research.

— All research articles will be refereed by at least two persons, one from the Editorial Board and another external.

— The only working language allowed will be English. — Four volumes are scheduled for publication, one in January, one in April, one in July and the other in October.

Aims and Scope

The aim of REVSTAT is to publish articles of high scientific content, in English, developing innovative statistical scientific methods and introducing original research, grounded in substantive problems.

REVSTAT covers all branches of Probability and Statistics. Surveys of important areas of research in the field are also welcome.

Abstract and Indexing Services

The REVSTAT is covered by the following abstracting/indexing services:

- Current Index to Statistics
- Google Scholar
- Mathematical Reviews
- Science Citation Index Expanded
- Zentralblatt für Mathematic

Instructions to Authors, special-issue editors and publishers

The articles should be written in English and may be submitted in two different ways:

- By sending the paper in PDF format to the Executive Editor (revstat@ine.pt) and to one of the two Editors or Associate Editors, whose opinion the author wants to be taken into account, together to the following e-mail address: revstat@fc.ul.pt

- By sending the paper in PDF format to the Executive Editor (revstat@ine.pt), together with the corresponding PDF or PostScript file to the following e-mail address: revstat@fc.ul.pt.

Submission of a paper means that it contains original work that has not been nor is about to be published elsewhere in any form.

Manuscripts (text, tables and figures) should be typed only in black on one side, in double-spacing, with a left margin of at least 3 cm and with less than 30 pages. The first page should include the name, institution and address of the author(s) and a summary of less than one hundred words, followed by a maximum of six key words and the AMS 2000 subject classification.

Authors are obliged to write the final version of accepted papers using LaTeX, in the REVSTAT style. This style (REVSTAT.sty), and examples file (REVSTAT.tex), which may be download to PC Windows System (Zip format), Macintosh, Linux and Solaris Systems (StuffIt format), and Mackintosh System (BinHex Format), are available in the REVSTAT link of the Statistics Portugal Website: http://www.ine.pt/revstat/inicio.html

Additional information for the authors may be obtained in the above link.

Accepted papers

Authors of accepted papers are requested to provide the LaTeX files and also a postscript (PS) or an acrobat (PDF) file of the paper to the Secretary of REVSTAT: revstat@ine.pt.

Such e-mail message should include the author(s)'s name, mentioning that it has been accepted by REVSTAT.

The authors should also mention if encapsulated postscript figure files were included, and submit electronics figures separately in .tiff, .gif, .eps or .ps format. Figures must be a minimum of 300 dpi.

Editorial Board