Revista Colombiana de Estadística

Volumen 35. Número 3 - diciembre - 2012 ISSN 0120 - 1751

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ
FACULTAD DE CIENCIAS
DEPARTAMENTO DE ESTADÍSTICA

Revista Colombiana de Estadística
http://www.estadistica.unal.edu.co/revista
http://es.wikipedia.org/wiki/Revista_Colombiana_de_Estadistica
http://www.emis.de/journals/RCE/
revcoles_fcbog@unal.edu.co

Indexada en: Ulrichsweb, Scopus, Science Citation Index Expanded (SCIE), Web of Science (WoS), SciELO Colombia, Current Index to Statistics, Mathematical Reviews (MathSci), Zentralblatt Für Mathematik, Redalyc, Latindex, Publindex (A_1)

La *Revista Colombiana de Estadística* es una publicación semestral del Departamento de Estadística de la Universidad Nacional de Colombia, sede Bogotá, orientada a difundir conocimientos, resultados, aplicaciones e historia de la estadística. La Revista contempla también la publicación de trabajos sobre la enseñanza de la estadística.

Se invita a los editores de publicaciones periódicas similares a establecer convenios de canje o intercambio.

Dirección Postal:
Revista Colombiana de Estadística
© Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Carrera 30 No. 45-03
Bogotá – Colombia
Tel: 57-1-3165000 ext. 13231
Fax: 57-1-3165327

Adquisiciones:
Punto de venta, Facultad de Ciencias, Bogotá.
Suscripciones:
revcoles_fcbog@unal.edu.co
Solicitud de artículos:
Se pueden solicitar al Editor por correo físico o electrónico; los más recientes se pueden obtener en formato PDF desde la página Web.

Edición en LaTeX: Patricia Chávez R. E-mail: apchavezr@gmail.com

Impresión: Editorial Universidad Nacional de Colombia, Tel. 57-1-3165000 Ext. 19645, Bogotá.

Revista Colombiana de Estadística	Bogotá	Vol. 35	N° 3
ISSN 0120 - 1751	COLOMBIA	diciembre-2012	Págs. 331-521

Contenido

Editorial

Leonardo Trujillo[a]

Department of Statistics, Universidad Nacional de Colombia, Bogotá, Colombia

Welcome to the third issue of the 35th volume of the Revista Colombiana de Estadistica (Colombian Journal of Statistics). The first issue was published last June and the second one was a past Special Issue about Biostatistics with Professors Liliana Lopez-Kleine and Piedad Urdinola as Guest Editors. We will keep also, as the first issue, the characteristic of being an issue entirely published in English language as part of the requirements of being the winners of an Internal Grant for a second year in a row at the National University of Colombia (Universidad Nacional de Colombia) among many Journals (see editorial of December 2011). We are also very proud to announce that the Colombian Journal of Statistics have maintained its categorization as an A1 Journal by Publindex (Colciencias) which ranges the journals in the country, being A1 the maximum category. Thanks to all the Editorial and Scientific Committees and Patricia Chávez, our assistant in the Journal, as this is a result of the continuous help obtained from all of them. More information available at `http://201.234.78.173:8084/publindex/EnIbnPublindex/resultados.do`.

The topics in this current issue range over diverse areas of statistics: two papers in Survey Sampling by Mahdizadeh and Arghami and another one by Nayak and Sahoo; one paper in Biostatistics by Leiva, Ponce, Merchant and Bustos; one paper in Censored Data by Salinas, Pérez, Gonzalez and Vaquera; one paper in Expectation Maximization by Pereira, Marques and da Costa; one paper in Geostatistics by Giraldo, Mateu and Delicado; one paper in Medical Statistics by Tovar and Achcar; one paper in Nonparametric Statistics by Marozzi; one paper in Textual Statistics by Guerrero and one paper in Time Series Analysis by Ocampo and Rodriguez.

The International Year of Statistics (Statistics2013) is now promoted around the globe by the International Statistical Institute (ISI). It is important that Colombian and world organizations including government institutions, research institutions and universities join this big event in order to promote our discipline throughout all over the world and its impact on all aspects of society. You can find more information available at `http://www.statistics2013.org/`. From Colombia, the following colleges and universities will be participants in Statistics2013: Corporacion Universitaria Empresarial Alexander von Humboldt, Universidad La Gran Colombia, Universidad Nacional de Colombia (Bogota and Medellin branches), Universidad Industrial de Santander. Also, the Colombian National Department of Statistics - DANE (`http://www.statistics2013.org/`

[a] Editor in chief.
E-mail: ltrujilloo@unal.edu.co

participants.cfm). More information about how to get involved in these activities can be found at http://www.isi-web.org/recent-pages/621-statistics 2013-update-november-2012. Additionally, more than 100 scientific societies, universities, research institutes, and organizations all over the world have banded together to dedicate 2013 as a special year for the Mathematics of Planet Earth (http://mpe2013.org/).

For 2014, the XIII CLAPEM (Latin American Congress of Probability and Mathematical Statistics) will be held for the first time in Colombia at the city of Cartagena. It is held with the Latin American Chapter of the Bernoulli Society. CLAPEM is the largest conference gathering scientists in the particular areas of Probability and Mathematical Statistics in the region and takes place every two/three years. It has already been organized in Argentina, Brazil, Chile, Cuba, Mexico, Peru, Uruguay and Venezuela. The CLAPEM activities include lectures held by invited researchers, satellite meetings, sessions of oral and poster contributions, short courses, and thematic sessions. The XIII CLAPEM is organized by the Bernoulli Society, Universidad Nacional de Colombia, Universidad del Rosario, Universidad de los Andes and Universidad de Cartagena. The Scientific Committee is as follows: Alejandro Jara (Chile), Antonio Galves (Brazil), Graciela Boente (Argentina), José Rafael Leon (Venezuela), Karine Bertin (Chile), Leonardo Trujillo (Colombia), Pablo Ferrari (Argentina), Paola Belmolen (Uruguay), Ramón Giraldo (Colombia), Serguei Popov (Brazil), Victor Perez Abreu (Mexico). The Local and Scientific Committees have started to work and further information will be available soon. If you are interested you can also get more details with Ricardo Fraiman (president of the XIII CLAPEM, fraimanricardo@gmail.com) or Leonardo Trujillo (ltrujilloo@unal.edu.co).

This year, four eminent statisticians have passed away: David Binder, George Casella, James Durbin and Gad Nathan. We want to make recognition of their work in this Editorial. David Binder got his PhD in 1977 at the Imperial College, London, UK. He worked as a survey methodologist in Statistics Canada and developed methods in order to make inference from data obtained from complex survey designs. These methods are now available in commercial statistical packages such as SAS, SPSS, STATA and SUDAAN. He also published many papers at Biometrika, the Journal of the American Statistical Association, Survey Methodology and The Canadian Journal of Statistics, among other journals.

George Casella (1951-2012) was born in New York, USA and he got a PhD in Statistics from Purdue University. He was a Professor at the University of Florida. His research interests were on decision theory, environmental statistics, genetic statistics, objective and empirical Bayes and statistical confidence.

James Durbin (1923-2012) was a British econometrican and statistician, very well-known from his contributions on balanced incomplete experimental designs (Durbin test), Brownian motion, econometrics, estimating equations, goodness of fit tests, linear algebra (Levinson-Durbin recursion), serial correlation in regression (Durbin and Watson test) and time series analysis. He was a Professor at the London School of Economics. He was president of the International Statistical Institute and the Royal Statistical Society.

Gad Nathan was a distinguished professor in survey sampling and got his PhD at the Case Institute of Technology in Cleveland, USA. His contributions ranged from analyses from complex survey designs for longitudinal analysis, regression analysis and tests of independence in contingency tables; non-response adjustments and treatment of sensitive questions in surveys. A memorial session for him will be held in Hong Kong at the next ISI Statistical Congress.

Editorial

Leonardo Trujillo[a]

Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia

Bienvenidos al tercer número del volumen 35 de la Revista Colombiana de Estadística. El primer número fue publicado en Junio pasado y el segundo número corresponde al Numero Especial en Bioestadística que conto con las Profesoras Liliana López - Kleine y Piedad Urdinola como Editoras Invitadas. Hemos mantenido, como en el primer numero anterior, la característica de contar con artículos publicados únicamente en idioma ingles como parte de los requerimientos de ser ganadores por segundo año consecutivo de una Convocatoria Interna en la Universidad Nacional de Colombia entre otras revistas (ver editorial de Diciembre 2011). Estamos también muy gratos de anunciar que la Revista Colombiana de Estadística ha mantenido su categoría A1 ante Publindex (Colciencias) que categoriza las revistas a nivel nacional y siendo esta la máxima categoría de calidad para revistas nacionales. Gracias a todos los Comités Científico y Editorial y a Patricia Chávez, la asistente de la Revista, pues este es el resultado de la continua ayuda obtenida por parte de todos ellos. Mas información disponible en la página web `http://201.234.78.173:8084/publindex/EnIbnPublindex/resultados.do`.

Los temas del presente numero varían a través de diversas áreas de la estadística: dos artículos en Muestreo por Mahdizadeh y Arghami y otro por Nayak y Sahoo; un artículo en Análisis de Series de Tiempo por Ocampo y Rodríguez; un artículo en Bioestadística por Leiva, Ponce, Merchant y Bustos; un artículo en Datos Censurados por Salinas, Pérez, González y Vaquera; un artículo en Estadística Medica por Tovar y Achcar; un artículo en Estadística no Paramétrica por Marozzi; un artículo en Estadística Textual por Guerrero; un artículo en Expectation Maximization por Pereira, Marques y da Costa y un articulo en Geoestadística por Giraldo, Mateu y Delicado.

Se ha iniciado la promoción del Año Internacional de la Estadística (Statistics2013) alrededor del mundo por intermedio del International Statistical Institute (ISI). Es importante que organizaciones nacionales e internacionales incluyendo instituciones del gobierno, institutos de investigación y universidades se unan a este gran evento con el fin de promover la disciplina estadística y su impacto en muchos de los aspectos de la sociedad. Se puede encontrar mas información en la pagina web http://www.statistics2013.org/. Por Colombia, las siguientes instituciones educativas son participantes de Statistics2013: Corporación Universitaria Empresarial Alexander von Humboldt, Universidad La Gran Colombia, Universidad Nacional de Colombia (sedes Bogotá y Medellín), Universidad Industrial de Santander. También, el Departamento Administrativo Nacional de Estadística - DANE

[a] Editor
E-mail: ltrujilloo@unal.edu.co

(http://www.statistics2013.org/participants.cfm). Para obtener mas información acerca de como participar en estas actividades se puede encontrar en `http://www.isi-web.org/recent-pages/621-statistics2013-update-november-2012`. Adicionalmente, más de 100 sociedades científicas, universidades, institutos de investigación y organizaciones alrededor del mundo se han unido con el fin de dedicar el año 2013 como año especial para las Matemáticas en el Planeta Tierra (`http://mpe2013.org/`).

Para 2014, la XIII CLAPEM (Conferencia Latinoamericana de Probabilidad y Estadística Matemática) será organizada por primera vez en Colombia en la ciudad de Cartagena de Indias. Esta será organizada por el Capitulo Latinoamericano de la Sociedad Bernoulli. CLAPEM es la principal y mayor conferencia que reúne científicos en las áreas de Probabilidad y Estadística Matemática en la región y toma lugar cada dos o tres años. Se ha llevado a cabo anteriormente en Argentina, Brasil, Chile, Cuba, México, Perú, Uruguay y Venezuela. La conferencia incluye actividades como charlas a cargo de investigadores internacionales invitados, cursos cortos, reuniones satélites, sesiones de contribuciones orales y posters y sesiones temáticas. La XIII CLAPEM será organizada por la Sociedad Bernoulli, la Universidad Nacional de Colombia, Universidad del Rosario, Universidad de los Andes y Universidad de Cartagena. El Comité Científico esta conformado por: Alejandro Jara (Chile), Antonio Galves (Brasil), Graciela Boente (Argentina), José Rafael León (Venezuela), Karine Bertín (Chile), Leonardo Trujillo (Colombia), Pablo Ferrari (Argentina), Paola Belmolen (Uruguay), Ramón Giraldo (Colombia), Serguei Popov (Brasil), Víctor Pérez Abreu (México). Los Comités Científico y Local han comenzado a trabajar y más información acerca del evento estará disponible pronto. Si esta interesado puede obtener mas detalles con Ricardo Fraiman (presidente del XIII CLAPEM, fraimanricardo@gmail.com) o con Leonardo Trujillo (`ltrujilloo@unal.edu.co`).

Este año, cuatro estadísticos eminentes han fallecido: David Binder, George Casella, James Durbin y Gad Nathan. En esta Editorial queremos hacer un reconocimiento a su trabajo en Estadística. David Binder obtuvo su Doctorado en el Imperial College de Londres en 1977. Se desempeño como metodólogo de encuestas en Statistics Canada y desarrollo métodos para hacer inferencia en datos provenientes de diseños muestrales complejos. Estos métodos se encuentran ahora disponibles en paquetes estadísticos tales como SAS, SPSS, STATA y SUDAAN. Binder publico muchos artículos en revistas como Biometrika, la Journal of the American Statistical Association, Survey Methodology y The Canadian Journal of Statistics, entre otras.

George Casella (1951-2012) nació en New York, USA y obtuvo su Doctorado en Estadística en Purdue University. Fue Profesor de la Universidad de Florida. Sus áreas de investigación fueron principalmente confiabilidad estadística, estadística ambiental, estadística Bayesiana, estadística genética y teoría de la decisión.

James Durbin (1923-2012) fue un econometrista y estadístico británico, muy conocido por sus contribuciones en algebra lineal (recursión de Levinson - Durbin), análisis de series de tiempo, autocorrelación en regresión (prueba de Durbin y Watson), diseños experimentales incompletos balanceados (test de Durbin), econometría, ecuaciones de estimación, movimiento Browniano y pruebas de bondad de ajuste. He was a Professor at the London School of Economics. He was president of the International Statistical Institute and the Royal Statistical Society.

Gad Nathan fue un profesor israelí muy distinguido en muestreo y obtuvo su Doctorado en el Case Institute of Technology en Cleveland, USA. Sus contribuciones se extendieron desde ajustes por no respuesta; análisis para datos provenientes de muestras complejas en análisis longitudinales, análisis de regresión y pruebas de independencia para tablas de contingencia; y, tratamiento de preguntas sensibles en encuestas. Una sesión en su memoria se organizara en Hong Kong en el próximo Congreso Estadístico del ISI (International Statistical Institute).

Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 331 a 347

Two Dependent Diagnostic Tests: Use of Copula Functions in the Estimation of the Prevalence and Performance Test Parameters

Dos pruebas para diagnóstico clínico: uso de funciones copula en la estimación de la prevalencia y los parámetros de desempeño de las pruebas

José Rafael Tovar[1,a], Jorge Alberto Achcar[2,b]

[1] Centro de Investigaciones en Ciencias de la Salud (CISC), Escuela de Medicina y Ciencias de la Salud, Universidad del Rosario, Bogotá, Colombia

[2] Departamento de Medicina Social FMRP, Faculdade de Saúde, Universidade de São Paulo, Riberão Preto, Brasil

Abstract

In this paper, we introduce a Bayesian analysis to estimate the prevalence and performance test parameters of two diagnostic tests. We concentrated our interest in studies where the individuals with negative outcomes in both tests are not verified by a gold standard. Given that the screening tests are applied in the same individual we assume dependence between test results. Generally, to capture the possible existing dependence between test outcomes, it is assumed a binary covariance structure, but in this paper, as an alternative for this modeling, we consider the use of copula function structures. The posterior summaries of interest are obtained using standard MCMC (Markov Chain Monte Carlo) methods. We compare the results obtained with our approach with those obtained using binary covariance and assuming independence. We considerate two published medical data sets to illustrate the approach.

***Key words*:** Bayes analysis, Copula, Dependence, Monte Carlo Simulation, Public health.

Resumen

En este articulo introducimos un análisis Bayesiano para estimar la prevalencia y los parámetros de desempeño de pruebas para diagnóstico clínico, con datos obtenidos bajo estudios de tamizaje que incluyen el uso de dos pruebas diagnósticas en los cuales, los individuos con resultado negativo en

[a] Lecturer. E-mail: rtovar34@hotmail.com

[b] Associate Professor. E-mail: achcar@fmrp.usp.br

las dos pruebas no son confirmados con una prueba patrón de oro. Dado que las pruebas de tamizaje son aplicadas al mismo indivíduo, nosotros asumimos dependencia entre los resultados de las pruebas. Generalmente, para capturar la posible dependencia existente entre los resultados de las pruebas diagnósticas, se asume una estrutura de covarianza binaria, pero en este artículo, nosotros consideramos el uso de estructuras que pueden ser modaladas usando funciones cópula, como una alternativa al modelamiento de la dependencia. Las estadísticas a posteriori de interés son obtenidas usando métodos MCMC. Los resultados obtenidos usando nuestra aproximación son comparados con los obtenidos usando modelos que asumen estructura binária y con los obtenidos usando modelos bajo el supuesto de independencia entre resultados de las pruebas para diagnóstico clínico. Para ilustrar la aplicación del método y para hacer las comparaciones se usaron los datos de dos estudios publicados en la literatura.

***Palabras clave*:** análisis bayesiano, copula, dependencia, simulación Monte Carlo, salud pública.

1. Introduction

In literature, there area designs to evaluate new screening tests it which more than one diagnostic test is applied to the same individual and where in some cases all patients cannot be verified by a test free of error to classify individuals or Gold Standard. This situation implies in the presence of verification bias. When the design considers the use of two continuous scale diagnostic tests transformed to a binary scale using a cut-off point to classify an individual as positive or negative to a given disease, these tests could have dependent outcomes within a continuous dependence structure but as we have the final binary results to do the data analysis, we could model the dependence considering a bivariate Bernoulli distribution with the covariance as a dependence parameter. This approach has been studied by different authors such as Thibodeau (1981), Vacek (1985) and Walter & Irwig (1988), amongst others. Assuming binary structure, Böhning & Patilea (2008), developed two indexes to study the dependence between two diagnostic tests: a first is derived using the λ reparametrization introduced by Georgiadis, Johnson & Gardner (2003) and a second index derived by applying the OR (odds ratio) concept on 2×2 probability tables associated with the two diagnostic test results. Some approaches such as those of Brenner (1996), Qu & Hadgu (1998) and Torrance-Rynard & Walter (1997), have considered the continuous structure in the data to study the dependence between test outcomes using models of latent variable.

In this paper, we introduce a Bayesian model to estimate the prevalence, performance test parameters and the dependence between them, using two copula functions, the FGM (Farlie-Gumbel-Morgenstern) copula and the Gumbel copula. The FGM is a copula function that allows modeling very weak linear dependencies usually not easily observed using traditional bivariate plots.

If the continuous traits that make up the diagnostic tests have a dependence like FGM structure, usually the data analyst assumes independence in the statistical model used to obtain the parameter estimates. The form of the Gumbel copula used in this work, models relatively weak negative linear dependencies but the copula parameter of dependence belongs in the space (0,1). In agreement with some simulation results not showed in this paper, the bivariate plots obtained under different levels of Gumbel copula dependence show a dispersion similar with that observed when the data are obtained under independence assumption, then, it is not easy to observe the presence of a negative correlation between test outcomes. The use of this copula, also allows us to study dependencies with not necessarily linear structures which is possible in diagnostic situations whose results are obtained after dichotomization.

We compare the estimates obtained using copula models with those obtained assuming binary covariance structure and independence assumption. In our approach, we assume that the diagnostic procedure includes two (observable or not) variables measured on a continuous scale with some type of positive dependence between them that can be modeled using copula functions. Copula functions have been widely used for modeling the dependence between continuous scale variables regardless the type of distribution underlying in the margins, in many other subject or topic areas as hydrology and finance.

To illustrate our proposed models, we use two data sets introduced in the literature. The first one, was obtained from Smith, Bullock & Catalona (1997), who screened 19,476 men for prostate cancer using the Digital Rectal Exam (DRE) and the Prostate Specific Antigen (PSA) in serum. With that same data set, Böhning & Patilea (2008) and Martinez, Achcar & Louzada (2005) studied the association between diagnostic test results. The second data set was introduced by Ali, Moodambail, Hamrah, Bin-Nakhi & Sadeq (2007), where they evaluated a fast method to detect urinary tract infection in 132 children of both genders with ages ranging from three days to 11 years.

This paper is organized as follows: In Section 2 we introduce the model formulation for two associated diagnostic tests; in Section 3, we present our Bayesian estimation procedure; in Section 4, we introduce two examples; finally in section 5, we present some discussion on the obtained results.

2. Model Formulation for Two Dependent Diagnostic Tests

We consider four different models that can be used, the first model assumes conditionally independent tests results and the other three models assume that the tests are dependent conditionally on the disease status.

2.1. Model Under Independence Assumption

Two diagnostic tests are respectively denoted by T_1 and T_2 where $T_\nu = 1$ is related to a positive result for the test ν, $\nu = 1, 2$ and $T_\nu = 0$ is related to a negative result. In Table 1 we have a generic representation of the tests compared with an ideal reference test. If the study design implies that individuals with negative outcome in both tests are not verified by a test free of error to classify the individuals ("Gold Standard"), the values d, h, n_+ and n_- (showed in brackets), are unknown although the sum $u = n_+ + n_-$ is known.

Table 1: Tests results. Values in brackets are unknown under verification bias.

	Diseased subjects			Non-diseased subjects		
	$T_2 = 1$	$T_2 = 0$	Total	$T_2 = 1$	$T_2 = 0$	Total
$T_1 = 1$	a	b	$a + b$	e	f	$e + f$
$T_1 = 0$	c	$[d]$	$c + [d]$	g	$[h]$	$g + [h]$
Total	$a + c$	$b + [d]$	$[n_+]$	$e + g$	$f + [h]$	$[n_-]$

Let us denote by p the prevalence of a disease and by D the true status, when $D = 1$ denotes a diseased individual and $D = 0$ denotes a non-diseased individual. That is, $p = P(D = 1)$. The sensitivities are given by $S_\nu = P(T_\nu = 1 \mid D = 1)$ and the specificities are given by $E_\nu = P(T_\nu = 0 \mid D = 0)$.

For the independence assumption model, we use the Bayesian estimation procedure developed by Martinez et al. (2005) to obtain the likelihood contributions of the eight possible combinations of results among tests and true disease state as appear in the left column in Table 2.

2.2. Model Under Binary Dependence Structure

For a binary structure model, we assume as dependence parameter, a positive covariance between tests based on the joint Bernoulli distribution. We assumed that the dependence between tests is similar in diseased and non-diseased populations in the same way as considered by Dendukuri & Joseph (2001) to obtain the contributions to likelihood function of the eight combinations of results among the two diagnostic tests and the Gold Standard. The results are showed in Table 2.

Table 2: Likelihood contributions of all possible combinations of outcomes of T_1, T_2 and D. (f_i = number of individuals in the cell i; $i = 1, 2, \dots, 8$. Values in brackets are unknown under verification bias).

					Contribution to likelihood	
i	D	T_1	T_2	f_i	Independence assumption	Binary dependence
1	1	1	1	a	pS_1S_2	$p[S_1S_2 + \psi_D]$
2	1	1	0	b	$pS_1(1 - S_2)$	$p[S_1(1 - S_2) - \psi_D]$
3	1	0	1	c	$p(1 - S_1)S_2$	$p[(1 - S_1)S_2 - \psi_D]$
4	1	0	0	[d]	$p(1 - S_1)(1 - S_2)$	$p[(1 - S_1)(1 - S_2) + \psi_D]$
5	0	1	1	e	$(1 - p)(1 - E_1)(1 - E_2)$	$(1 - p)[(1 - E_1)(1 - E_2) + \psi_{ND}]$
6	0	1	0	f	$(1 - p)(1 - E_1)E_2$	$(1 - p)[(1 - E_1)E_2 - \psi_{ND}]$
7	0	0	1	g	$(1 - p)E_1(1 - E_2)$	$(1 - p)[E_1(1 - E_2) - \psi_{ND}]$
8	0	0	0	[h]	$(1 - p)E_1E_2$	$(1 - p)[E_1E_2 + \psi_{ND}]$

2.3. Model Assuming a Dependence Copula Structure

Let us assume that the test outcomes are realizations of the random variables V_1 and V_2 measured on a positive continuous scale ($V_1 > 0$ and $V_2 > 0$) which represent the expression of two biological traits whose behavior is altered by the presence of disease or infection process. Also, let us assume that two cut-off values ξ_1 and ξ_2 are chosen for each test in order to determine when an individual is classified as positive or negative. In this way we assume that an individual is classified as positive for test ν if $V_\nu > \xi_\nu$ that is, $T_\nu = 1$ if and only if $V_\nu > \xi_\nu$ for $\nu = 1, 2$. To model the dependence structure between the random variables V_1 and V_2, let us consider the use of copula functions, which has been studied by many authors ((Nelsen 1999) is a classical book on this topic). Multivariate distribution functions F can be written in the form of a copula function, that is, if $F(v_1, \dots v_m)$ is a joint multivariate distribution function with univariate marginal distribution functions $F_1(v_1), \dots, F_m(v_m)$, thus there exists a copula function $C(u_1, \dots, u_m)$ such that,

$$F(v_1, \dots, v_m) = C(F_1(v_1), \dots, F_m(v_m)) \tag{1}$$

When the marginal distributions are continuous, a copula function always exists and can be found from the relation

$$C(u_1, \dots, u_m) = F(F_1^{-1}(u_1), \dots, F_m^{-1}(u_m)) \tag{2}$$

For the special case of bivariate distributions, we have $m = 2$. The approach to formulate a multivariate distribution using a copula is based on the idea that a simple transformation ($U = F_1(V_1)$ and $W = F_2(V_2)$) can be made of each marginal variable in such a way that each transformed marginal variable has a uniform distribution. Specifying dependence between V_1 and V_2 is the same as specifying dependence between U and W, thus the problem reduces to specifying a bivariate distribution between two uniform variables, that is a copula.

2.3.1. Model Considering Dependence Type FGM Copula

The third model considered for the study of the dependence structure for two tests, is based in the Farlie Gumbel Morgenstern (FGM) copula widely studied by authors as Nelsen (1999), Amblard & Girard (2002, 2005, 2008). The FGM copula is defined by,

$$C_I(u, w) = uw[1 + \varphi(1-u)(1-w)] \tag{3}$$

where $u = F_1(v_1)$, $w = F_2(v_2)$ and φ is a copula parameter such that $-1 \leq \varphi \leq 1$. If $\varphi = 0$, we have two independent marginal random variables. We assume different parameters φ_D and φ_{ND} for diseased and non-diseased individuals, respectively.

From (3) the cumulative joint distribution and the join survival function for the random variables V_1 and V_2 is given by,

$$\begin{aligned} F_I(v_1, v_2) &= C_I(F_1(v_1), F_2(v_2)) \\ &= F_1(v_1)F_2(v_2)[1 + \varphi(1 - F_1(v_1))(1 - F_2(v_2))] \end{aligned} \tag{4}$$

$$S(v_1, v_2) = P(V_1 > v_1, V_2 > v_2) = 1 - F_1(v_1) - F_2(v_2) + F(v_1, v_2) \tag{5}$$

Within the diseased individuals group, we have,

$$\begin{aligned} F_1^D(\xi_1) &= P(V_1 \leq \xi_1 | D = 1) = 1 - S_1 \\ F_2^D(\xi_2) &= P(V_2 \leq \xi_2 | D = 1) = 1 - S_2 \end{aligned}$$

From (4), we have

$$\begin{aligned} F_D(\xi_1, \xi_2) &= F_1^D(\xi_1) F_2^D(\xi_2)[1 + \varphi(1 - F_1^D(\xi_1))(1 - F_2^D(\xi_2))] \\ &= (1 - S_1)(1 - S_2)(1 + \varphi_D S_1 S_2) \end{aligned}$$

and from (5) we have,

$$\begin{aligned} P(T_1 = 1, T_2 = 1 | D = 1) &= S_D(\xi_1, \xi_2) \\ &= 1 - (1 - S_1) - (1 - S_2) + (1 - S_1)(1 - S_2)(1 + \varphi_D S_1 S_2) \end{aligned}$$

That is,

$$P(T_1 = 1, T_2 = 1 | D = 1) = S_1 S_2 (1 + \varphi_D (1 - S_1)(1 - S_2))$$

and

$$P(T_1 = 1, T_2 = 1, D = 1) = p S_1 S_2 (1 + \varphi_D (1 - S_1)(1 - S_2))$$

Observe that, if $\varphi_D = 0$ (independent test outcomes), we have

$$P(T_1 = 1, T_2 = 1, D = 1) = p S_1 S_2$$

as given in Table 2.

Also,

$$\begin{aligned} P(T_1 = 1, T_2 = 0, D = 1) &= P(D = 1) P(T_1 = 1, T_2 = 0 | D = 1) \\ &= p P(V_1 > \xi_1, V_2 \leq \xi_2 | D = 1) \end{aligned}$$

On the other hand,

$$\begin{aligned} P(V_1 > \xi_1, V_2 \leq \xi_2 | D = 1) &= P(V_2 \leq \xi_2 | D = 1) - P(V_1 \leq \xi_1, V_2 \leq \xi_2 | D = 1) \\ &= F_2^D(\xi_2) - F_D(\xi_1, \xi_2) \end{aligned}$$

Thus,

$$P(T_1 = 1, T_2 = 1, D = 1) = p(1 - S_2) S_1 [1 - \varphi_D S_2 (1 - S_1)]$$

If $\varphi_D = 0$, we have

$$P(T_1 = 1, T_2 = 0, D = 1) = p S_1 (1 - S_2)$$

as in the independent case (see Table 2).

Similarly,

$$\begin{aligned} P(T_1=0,T_2=1,D=1) &= P(D=1)P(T_1=0,T_2=1|D=1) \\ &= pP(V_1\leq \xi_1, V_2>\xi_2|D=1) \end{aligned}$$

Since,

$$\begin{aligned} P(V_1\leq \xi_1, V_2>\xi_2|D=1) &= P(V_1\leq \xi_1|D=1) - P(V_1\leq \xi_1, V_2\leq \xi_2|D=1) \\ &= F_1^D(\xi_1) - F_D(\xi_1,\xi_2) \end{aligned}$$

then,

$$P(T_1=0,T_2=1,D=1) = p(1-S_1)S_2[1-\varphi_D S_1(1-S_2)]$$

When $\varphi_D=0$ we have $P(T_1=0,T_2=1,D=1)=pS_2(1-S_1)$ as in the independent case (see Table 2).

We also have,

$$\begin{aligned} P(T_1=0,T_2=0,D=1) &= P(D=1)P(T_1=0,T_2=0|D=1) \\ &= pP(V_1\leq \xi_1, V_2\leq \xi_2|D=1) \\ &= pF_D(\xi_1,\xi_2), \end{aligned}$$

that is,

$$P(T_1=0,T_2=0,D=1) = p(1-S_1)(1-S_2)[1+\varphi_D S_1 S_2]$$

Within the non-diseased individuals group, we have:

$$\begin{aligned} P(T_1=1,T_2=1,D=0) &= P(D=0)P(T_1=1,T_2=1 \mid D=0) \\ &= (1-p)P(V_1>\xi_1, V_2>\xi_2|D=0) \\ &= (1-p)S_{ND}(\xi_1,\xi_2) \\ &= (1-p)(1-F_1^{ND}(\xi_1) - F_2^{ND}(\xi_2) + F_{ND}(\xi_1,\xi_2) \end{aligned}$$

Observe that,

$$\begin{aligned} P(T_1=0|D=0) = P(V_1\leq \xi_1|D=0) = F_1^{ND}(\xi_1) = E_1 \qquad \text{and} \\ P(T_2=0|D=0) = P(V_2\leq \xi_2|D=0) = F_2^{ND}(\xi_1) = E_2 \end{aligned}$$

Using (4), we have

$$F_{ND}(\xi_1,\xi_2) = E_1E_2[1+\varphi_{ND}(1-E_1)(1-E_2)]$$

That is,

$$P(T_1=1,T_2=1,D=0) = (1-p)(1-E_1)(1-E_2)[1+\varphi_{ND}E_1E_2]$$

The contributions to the likelihood for all situations with diseased and non-diseased individuals are given in Table 3.

2.3.2. Model Considering Dependence Type Gumbel Copula

The last considered model, is derived from Gumbel copula function defined as;

$$C_{II}(u,w) = u + w - 1 + (1-u)(1-w)\exp\{-\phi\log(1-u)\log(1-w)\} \qquad (6)$$

In this model, the joint cumulative distribution function for the random variables V_1 and V_2 is given by,

$$\begin{aligned} F_{II}(v_1 v_2) = F_1(v_1) + F_2(v_2) - 1 + (1 - F_1(v_1))(1 - F_2(v_2)) \\ \exp\{-\phi\log(1-F_1(v_1))\log(1-F_2(v_2))\} \qquad (7) \end{aligned}$$

As pointed out by (Gumbel 1960) for this copula model, when $\phi = 1$ the Pearson correlation linear coefficient (ρ) takes the value -0.40365. In this case, the parameter of the Gumbel copula, does not models positive linear correlations. Also, when the two variables are independent, ϕ takes the zero value.

Employing the same arguments considered with the FGM copula and using (7) we obtain all the contributions for the likelihood function when it is assumed a Gumbel copula dependence structure (Table 3).

Table 3: Likelihood contributions of all possible combinations of outcomes of T_1, T_2 and D when the dependence has the "FGM copula" or "Gumbel copula" structure. (f_i = number of individuals in the cell i; $i = 1, 2, \dots, 8$. Values in brackets are unknown under verification bias).

					Contribution to likelihood	
i	D	T_1	T_2	f_i	"FGM copula"	"Gumbel copula"
1	1	1	1	a	$pS_1S_2[1 + \varphi_D(1-S_1)(1-S_2)]$	$pS_1S_2Q_1$
2	1	1	0	b	$pS_1(1-S_2)[1 - \varphi_D(1-S_1)S_2]$	$pS_1[1 - S_2Q_1]$
3	1	0	1	c	$p(1-S_1)S_2[1 - \varphi_D S_1(1-S_2)]$	$pS_2[1 - S_1Q_1]$
4	1	0	0	[d]	$p(1-S_1)(1-S_2)[1 + \varphi_D S_1S_2]$	$p[1 - S_1 - S_2 + S_1S_2Q_1]$
5	0	1	1	e	$(1-p)(1-E_1)(1-E_2)[1 + \varphi_{ND}E_1E_2]$	$(1-p)(1-E_1)(1-E_2)Q_2$
6	0	1	0	f	$(1-p)(1-E_1)E_2[1 - \varphi_{ND}E_1(1-E_2)]$	$(1-p)(1-E_1)[1 - (1-E_2)Q_2]$
7	0	0	1	g	$(1-p)E_1(1-E_2)[1 - \varphi_{ND}E_2(1-E_1)]$	$(1-p)(1-E_2)[1 - (1-E_1)Q_2]$
8	0	0	0	[h]	$(1-p)E_1E_2[1 + \varphi_{ND}(1-E_1)(1-E_2)]$	$(1-p)[E_1 + E_2 - 1 + (1-E_1)(1-E_2)Q_2]$

Observe that; $Q_1 = \exp(-\phi_D \log S_1 \log S_2)$, $Q_2 = \exp(-\phi_{ND}\log(1-E_1)\log(1-E_2))$

3. Bayesian Approach

For a Bayesian analysis of the proposed models, we consider different Beta prior distributions on the prevalence, performance measure parameters (sensitivities and specificities) and the copula parameters. In some cases, we could have some prior information on the parameters from experts in diagnostic medical tests or from previous studies on the subject.

For a Bayesian analysis of the models, we assumed positive dependence between the diagnostic tests in the same way as it was considered by Dendukuri & Joseph (2001) (therefore $P(\varphi < 0) = 0$ and $P(\psi < 0) = 0$) we could assume uniform U(a,b) as non-informative prior distributions and Beta(a,b) distributions for the informative situation for FGM and Gumbel dependence parameters and for prevalence and performance test parameters. If we need to elicit informative prior distributions for binary covariance, we could use the Generalized Beta(a,b) distribution in the same way that was considered by Martinez et al. (2005). For the non-informative case the Uniform U(0,1) distribution should be a good option.

Usually, we do not have any kind of information about the copula parameters, that is, for both copula dependence parameters. In this case, we used the procedure developed by Tovar (2012) to obtain the prior hyperparameters and we assume that the dependence takes values within of some interval (θ_1, θ_2) within of parametric space. In this way, if we assumed that the dependence is weak, the parameter could belong to the interval $(0, 1/4)$; when the dependence is moderate the parameter should be in to the interval $(1/4, 3/4)$ and when the dependence is strong, the parameter should be in to the interval $(3/4, 1)$. To obtain the hyperparameter values, we take the midpoint of the interval as the mean $E(\theta)$ and we apply the Chebychev's inequality to approximate the variance $V(\theta)$, as follows:

$$\begin{aligned} P(|\theta - E(\theta)| \geq k\sigma) &\leq \frac{1}{k^2} = \gamma \\ P([\theta - E(\theta)]^2 \geq k^2\sigma^2) &\leq \gamma \\ P(\alpha[\theta - \theta_0]^2 \geq \sigma^2) &\leq \gamma \end{aligned} \tag{8}$$

where γ is the prior probability of θ do not belong to the constructed interval.

Therefore, the variance will be a function of the prior established probability to interval values of the unknown quantity and the distance between θ_0 and a percentile of the distribution. If it is replaced θ by some of the known values θ_1 or θ_2 in the equation (8) it is easy to obtain a approximated value for the variance of the Beta prior distribution, as follows;

$$\sigma^2 \leq \alpha[\theta_1 - \theta_0]^2 \cong \frac{ab}{(a+b)^2(a+b+1)} \tag{9}$$

And as the mean $\theta_0 = E(\theta)$ and the variance $\sigma^2 = V(\theta)$ can be written as functions of the Beta prior hyperparameters, it is necessary to solve a system of two equations with two unknowns to find values of a and b i.e:

$$\begin{aligned} \omega &= \frac{\theta_0}{(1-\theta_0)} \\ a &= \omega b \\ b &= \frac{\omega - [(\omega+1)^2\sigma^2]}{(\omega^3 + 3\omega^2 + 3\omega + 1)\sigma^2} \end{aligned} \tag{10}$$

In this way, assuming $\gamma = 0.05$ in (8), for the FGM and Gumbel dependence parameters we have evaluated a Beta(17, 122) distribution, a Beta(39.5, 39.5) distribution and a Beta(122, 17) distribution as informative prior distributions and finally we have selected as selection criteria the Deviance Information Criteria DIC Spiegelhalter, Thomas, Best & Lunn (2003) obtained within the WinBUGS environment and a heuristic procedure that assumes two criteria: quality in the convergence of the MCMC procedure and concentration of the posterior distribution using the coefficient of variation (CV). The best model should have the lower DIC, the best performance in MCMC convergence and highest concentration around the posterior mean (lowest CV).

We have seven parameters to be estimated, two sensitivities, two specificities, one prevalence, one dependence parameter for diseased individuals and another one for non-diseased individuals. If we assume a design with the presence of verification bias, we have only four degrees of freedom for the estimation process and if we assume a design without verification bias, we have six information components. Therefore, in both cases the model is non-identified. Using a classical approach, the problem has been addressed giving fixed values to a subset of parameters and estimating the remaining unconstrained parameters (Vacek 1985), but since all parameters are typically unknown, the division into constrained and unconstrained sets is often quite arbitrary. Since the Bayesian paradigm some authors as Joseph, Gyorkos & Coupal (1995), have proposed to construct informative prior distribution over a subset or over all unknown quantities. In accordance with Dendukuri & Joseph (2001), informative priors would be needed on at least as many parameters as would be constrained when using the most frequent approach. In this approach, the prior information is used to distinguish between the numerous possible solutions for the non-identifiable problem. This approach is approximately numerically equivalent to the most frequent approach when a degenerate (point mass) distribution is used that matches the constrained parameter values and diffuse prior distributions are used for the non-constrained parameters. In order to treat the non-identifiability problem, first, we assume informative prior distributions over the subset of dependence parameters and non-informative prior distributions on prevalence and performance test parameters and next, we assume informative prior distributions on all set of parameters in accordance with what was suggested by Joseph et al. (1995).

As the posterior distributions do not have closed forms, we have used MCMC methods, especially Metropolis-Hastings algorithm to obtain estimates for the parameters. For all models, $500,000$ Gibbs samples were simulated from the conditional distributions. From these generated samples, we discarded the first $50,000$ samples to eliminate the effect of the initial values and we also considered a spacing of 100. Convergence of the algorithm was verified graphically and also using standard existing methods implemented in the software CODA (Best, Cowles & Vines 1995).

4. Examples

4.1. Cancer Data

As a first example, we have used a data set introduced by (Smith et al. 1997). They screened 19,476 men for prostate cancer using Digital Rectal Examination (DRE) and Prostate-Specific Antigen (PSA) in serum. The PSA level was considered suspicious for cancer if it exceeded 4.0 ng/ml. Subjects with positive results on either DRE or PSA were submitted to an ultrasound guided needle biopsy test which was considered as "gold standard". This data set obtained under verification bias is related to approximately 20,000 individuals, as such, it may be considered as a large sample size.

For prior distribution elicitation, we have used the results introduced by Böhning & Patilea (2008). We get the values for the δ and λ indexes and from these results, we estimated the quantities d and h of non-verified subjects given in Table 1. (See Table 4).

Table 4: Estimated values for the dependence indexes and quantities of non-verified individuals using Böhning's results. The values in brackets were calculated using δ_i index, the another one using λ_i index.

	Diseased subjects $\lambda_1 = 2.42$, $\delta_1 = 3.08$			Non-diseased subjects $\lambda_0 = 2.40$, $\delta_0 = 3.03$		
	$DRE+$	$DRE-$	Total	$DRE+$	$DRE-$	Total
$PSA+$	189	292	481	141	755	896
$PSA-$	145	1431[691]	1576[836]	1002	15521[16261]	16523[17263]
Total	334	1723[983]	2057[1317]	1143	16276[17016]	17419[18159]

Using the data in Table 4 we assumed prior independence between the components of the parameter vector $[\theta_1 = S_1, \theta_2 = S_2, \theta_3 = E_1, \theta_4 = E_2, \theta_5 = p]$ to obtain estimates and intervals where it is possible assume to find each component with a probability $1 - \gamma = 0.95$. (See Table 5).

Table 5: Informative prior distribution hyperparameters for performance test parameters, prevalence and covariance (Martinez's prior informative distributions for ψ).

PARAMETER	INTERVAL	E(θ)	a_θ	b_θ
S_1	0.236 - 0.365	0.3006	303	704
S_2	0.162 - 0.254	0.2080	324	1232
E_1	0.949 - 0.951	0.950	902500	47500
E_2	0.934 - 0.937	0.9355	501758	34595
p	0.068 - 0.106	0.0866	379	4002
ψ_D	0.004659 - 0.004719	0.004689	486303	103225102
ψ_{ND}	0.080 - 0.133	0.1722	289	2421

Assuming prior independence, for each interval we take the midpoint of each interval as the expected value of the prior distribution and we use the Chebychev inequality to get approximations for the variance of each parameter in the way that was described in Section 3 and we obtained the hyperparameter values that appear in Table 5. For this set of parameters we have used U(0,1) distributions as non-informative priors.

To elicit binary covariance prior distributions, we have used the results obtained by Martinez et al. (2005). They estimated the covariance parameter for the same cancer data under a Bayesian approach assuming non-informative prior distributions for ψ_D and ψ_{ND}. We have used the 95% credible regions obtained by them and we applied the same procedure employed with the test parameters and prevalence. As non-informative distributions we have used GenBeta(1/2, 1/2) distributions.

For the copula parameters $\theta_2 = [\varphi_D, \varphi_{ND}, \phi_D, \phi_{ND}]$ we assumed the Beta distributions Beta(17, 122), Beta(39.5, 39.5) and Beta(122, 17) as prior distributions and Uniform U(0,1) as non-informative prior distributions. To address the lack identifiability problem of we have putting informative prior distributions on a subset or on the complete set of parameters considering two set of models as follows:

- Set 1 of models: informative prior distribution for the copula parameters and non-informative prior distributions for the prevalence and test parameters
- Set 2 of models: informative prior distributions for all parameters (See Table 6)

Table 6: Bayesian posterior summaries obtained by analyzing the data considering independence between tests assumption and different dependence structures. (Posterior means and 95% credible intervals (95% CrI) for each parameter of interest).

Set 1 of models				Set 2 of models			
Model	Parameter	Means	95% CrI	Model	Parameter	Means	95% CrI
$M_{1,1}$ $DIC = 180.4$	S_1	0.567	0.529 - 0.605	$M_{2,1}$ $DIC = 337.1$	S_1	0.258	0.252 - 0.264
	S_2	0.394	0.363 - 0.394		S_2	0.226	0.208 - 0.244
	E_1	0.952	0.950 - 0.954		E_1	0.948	0.946 - 0.950
	E_2	0.946	0.943 - 0.949		E_2	0.947	0.944 - 0.950
	p	0.044	0.041 - 0.047		p	0.080	0.075 - 0.085
$M_{1,2}$ $DIC = 55.2$	ψ_D	0.0316	0.019 - 0.046	$M_{2,2}$ $DIC = 54.4$	ψ_D	0.046	0.037 - 0.055
	ψ_{ND}	0.005	0.004 - 0.006		ψ_{ND}	0.005	0.004 - 0.006
	S_1	0.470	0.380 - 0.548		S_1	0.295	0.274 - 0.316
	S_2	0.335	0.273 - 0.393		S_2	0.211	0.196 - 0.227
	E_1	0.951	0.948 - 0.955		E_1	0.950	0.950 - 0.950
	E_2	0.937	0.933 - 0.940		E_2	0.936	0.935 - 0.936
	p	0.051	0.044 - 0.062		p	0.082	0.076 - 0.088
$M_{1,3}$ $DIC = 156.5$	φ_D	0.156	0.136 - 0.176	$M_{2,3}$ $DIC = 225.5$	φ_D	0.135	0.123 - 0.148
	φ_{ND}	0.040	0.036 - 0.044		φ_{ND}	0.041	0.039 - 0.043
	S_1	0.538	0.480 - 0.595		S_1	0.320	0.300 - 0.343
	S_2	0.384	0.339 - 0.430		S_2	0.225	0.209 - 0.242
	E_1	0.952	0.948 - 0.955		E_1	0.950	0.949 - 0.950
	E_2	0.937	0.933 - 0.941		E_2	0.936	0.935 - 0.936
	p	0.045	0.040 - 0.050		p	0.074	0.069 - 0.079
$M_{1,4}$ $DIC = 192.7$	ϕ_D	0.120	0.072 - 0.179	$M_{2,4}$ $DIC = 294.4$	ϕ_D	0.047	0.028 - 0.072
	ϕ_{ND}	0.017	0.010 - 0.026		ϕ_{ND}	0.018	0.011 - 0.027
	S_1	0.593	0.540 - 0.645		S_1	0.330	0.307 - 0.353
	S_2	0.424	0.379 - 0.469		S_2	0.228	0.211 - 0.245
	E_1	0.952	0.948 - 0.955		E_1	0.950	0.949 - 0.951
	E_2	0.937	0.933 -0.941		E_2	0.936	0.935 - 0.936
	p	0.040	0.037 - 0.044		p	0.072	0.0671 - 0.0771

$M_{j,1}$, $j = 1, 2$: Models under assumption of independence between tests
$M_{j,2}$, $j = 1, 2$: Covariance parameters with informative prior distributions
$M_{j,3}$, $j = 1, 2$: FGM dependence parameters with Beta(122, 17) prior distributions
$M_{j,4}$, $j = 1, 2$: Gumbel dependence parameters with Beta (17, 122) prior distributions

From the results in Table 6, we observe that in this example with a large sample size (almost 20,000 individuals), we have great differences in the posterior summaries of interest, especially for the sensitivities S_ν, $\nu = 1, 2$ of the tests considering different priors for the parameters and different modeling structures. It is also interesting to observe that the specificities $E_\nu \quad \nu = 1, 2$, that is, the probabilities of negative tests given that the individuals are not diseased, are almost not affected by the different priors and different modeling structures in presence or not of an dependence parameter. These results could be of great interest for medical diagnostic tests.

We also observe a large variability on the obtained DIC values considering each assumed model. The smallest DIC values are obtained for the class of models with a bivariate binary structure.

4.2. Urinary Tract Infection (UTI)

In this example, we consider a data set introduced by Ali et al. (2007) who evaluated a fast method to detect urinary tract infection. In this case, we can suspect an association between tests, since the results of the tests are more likely to be positive when the individual has a greater presence of infection. The authors considered the presence of nitrites ($N = test1$), and the levels of leukocyte esterase in urine ($LE = test2$) as screening tests and a bacterial culture as the confirmatory test. They applied the three methods in 132 children of both genders with ages ranging from three days to 11 years. The obtained performance test and prevalence estimates were compared with those obtained in other five studies. Since one of those studies had incomplete data, we only considered the results of the four complete studies to elicit our prior distributions. For each estimated parameter, we calculated the mean and variance of the results obtained in the five studies (including Ali's study) and used them as prior means and variances of the parameters. Thus, the informative prior distributions for prevalence and performance test parameters are given by:

$$S_1 \sim Beta(4.15, 4.5), \quad S_2 \sim Beta(15.7, 2.4)$$

$$E_1 \sim Beta(0.5, 13), \quad E_2 \sim Beta(8.3, 2.8)$$

and

$$p \sim Beta(2.1, 22.3)$$

For copula and covariance parameters, we assume the same informative priors used for copula parameters considered in the first example. We also assume uniform U(0,1) prior distributions for the performance test parameters as non-informative priors and applied the same procedure for the estimation process used in the cancer data example. The results obtained are given in Table 7.

In this example with a small sample size, but not including missing data, we observe from Table 7, that the sensitivities $S_\nu \quad \nu = 1, 2$ were not greatly affected by the choice of prior distributions (informative or not) and modeling structures, but the specificities $E_\nu \quad \nu = 1, 2$ have a great variability considering the different modeling structures. We also observe that the prevalences have similar posterior summaries considering each model and the DIC values do not present great differences for each modeling structure.

Table 7: Bayesian posterior summaries obtained by analyzing the data considering independence between tests assumption and different dependence structures. (Posterior means and 95% credible intervals (95% CrI) for each parameter of interest).

Set 1 of models				Set 2 of models			
Model	Parameter	Means	95% CrI	Model	Parameter	Means	95% CrI
$M_{1,1}$ $DIC = 36.6$	S_1	0.387	0.318 - 0.457	$M_{2,1}$ $DIC = 40.3$	S_1	0.387	0.318 - 0.457
	S_2	0.855	0.803 - 0.901		S_2	0.855	0.803 - 0.901
	E_1	0.875	0.799 - 0.935		E_1	0.769	0.682 - 0.846
	E_2	0.513	0.402 - 0.625		E_2	0.544	0.438 - 0.648
	p	0.673	0.616 - 0.728		p	0.625	0.568 - 0.679
$M_{1,1}$ $DIC = 36.4$	ψ_D	0.029	0.016 - 0.048	$M_{2,1}$ $DIC = 47.9$	ψ_D	0.028	0.015 - 0.049
	ψ_{ND}	0.036	0.014 - 0.067		ψ_{ND}	0.077	0.049 - 0.110
	S_1	0.384	0.287 - 0.484		S_1	0.392	0.298 - 0.489
	S_2	0.847	0.767 - 0.912		S_2	0.857	0.785 - 0.916
	E_1	0.870	0.756 - 0.949		E_1	0.702	0.582 - 0.809
	E_2	0.567	0.424 - 0.702		E_2	0.541	0.420 - 0.660
	p	0.672	0.590 - 0.748		p	0.583	0.505 - 0.658
$M_{1,2}$ $DIC = 52.6$	φ_D	0.050	0.027 - 0.794	$M_{2,2}$ $DIC = 40.0$	φ_D	0.053	0.028 - 0.085
	φ_{ND}	0.161	0.104 - 0.208		φ_{ND}	0.068	0.025 - 0.130
	S_1	0.392	0.289 - 0.487		S_1	0.385	0.289 - 0.487
	S_2	0.855	0.783 - 0.915		S_2	0.845	0.764 - 0.911
	E_1	0.681	0.556 - 0.795		E_1	0.866	0.753 - 0.948
	E_2	0.610	0.479 - 0.735		E_2	0.578	0.433 - 0.716
	p	0.583	0.505 - 0.658		p	0.672	0.590 - 0.748
$M_{1,3}$ $DIC = 42.6$	ϕ_D	0.118	0.071 - 0.175	$M_{2,3}$ $DIC = 55.3$	ϕ_D	0.118	0.071 - 0.175
	ϕ_{ND}	0.119	0.072 - 0.177		ϕ_{ND}	0.121	0.073 - 0.179
	S_1	0.387	0.290 - 0.488		S_1	0.392	0.298 - 0.490
	S_2	0.847	0.766 - 0.913		S_2	0.857	0.784 - 0.916
	E_1	0.864	0.751 - 0.946		E_1	0.679	0.553 - 0.792
	E_2	0.576	0.431 - 0.715		E_2	0.625	0.494 - 0.748
	p	0.672	0.590 - 0.748		p	0.582	0.504 - 0.658

$M_{j,1}$, $j = 1, 2$: Models under assumption of independence between tests
$M_{j,2}$, $j = 1, 2$: Models using $GenBeta(39.5, 39.5)$ prior distributions for the covariance parameters
$M_{j,3}$, $j = 1, 2$: Models taken $GenBeta(122, 17)$ prior distributions for the association FGM parameter
$M_{j,4}$, $j = 1, 2$: Models with $Beta(17, 122)$ prior distributions for association Gumbel parameter

Considering DIC as discrimination criteria, we could assume a model with independence between the diagnostic tests considering informative or non-informative prior distributions or a model with dependence between tests given by a bivariate Bernoulli distribution (small DIC and similar performance test parameter estimates).

In this case, the copula parameter in non-diseased individuals presents an important change when we use informative priors over all parameters, while in the other group it remains unchanged. The specificity of the test N (test1) shows changes in the three models whether we use or do not use informative priors over the vector of non-dependence parameters. For the binary covariance and Gumbel models, the E_1 estimate with informative priors is lower than in the other models while in FGM model we observed an opposite behavior. When we have small sample size, the FGM model shows a more unstable behavior in the estimation of association parameter for non-diseased individuals. The DIC values for the different models do not show important changes. It is important, to observe that the DIC value of the FGM model with informative priors over all parameters is very similar with the DIC value of the Gumbel model when we use non-informative priors over test parameters. On the other hand, the DIC value obtained assuming non-informative priors over test parameters in one model is similar with those obtained using informative priors over complete set in the other one. It is also interesting to see that the behavior of the FGM model with small sample size data is similar to the behavior observed in the Gumbel model when we have a large sample size.

5. Conclusion and Remarks

The main goal of this paper was to develop a Bayesian procedure to estimate the prevalence, performance test and copula parameters of two diagnostic tests in presence of verification bias and considering the dependence between test results.

We proposed the use of copula structure models to get the estimation of the parameters under dependence assumption and specifically, we have used the Farlie Gumbel Morgenstern (FGM) and the Gumbel copula models to compare the obtained results with a model under independence assumption between tests and another one assuming dependent binary tests in designs that consider two diagnostic tests with continuous outcome for screening, a perfect “gold standard” and verification bias presence. The estimation model obtained under verification bias presence, implies a lack of identifiability problem, because we have more parameters than informative pieces in the likelihood function. Given that, our approach considers the continuous dependence structure in the data but the estimation process is made with the binary observations in presence of verification bias, we consider that to estimate the parameters under the Bayesian approach is easier than under the frequentist approach, because many times it is possible that we do not have the continuous values, for instance, when the measured continuous traits are non-observable (they are latent variables).

We illustrated the procedure using two published data sets: one with a large sample size and another one with a small sample size of individuals. In both cases, the better fit for the data was obtained assuming binary associated tests and taking the covariance as a parameter. The FGM model showed better fit when compared to the Gumbel copula, regardless the sample size. With a large sample size, the FGM model presented DIC values lower when it was fitted assuming non-informative prior distributions on test parameters and the estimates are very close with those obtained using maximum likelihood method, reflecting the effect that has the observed data in the estimation process.

However, to use informative prior on all parameters allow us to obtain sensitivity estimates with shorter credibility regions which is very good if we consider that within the large sample used, the true positives are a small part. The previous conclusion is reinforced by the results observed with the data of the small sample size, which the informative prior on all parameters gave better fit. With the Gumbel model, we obtained similar results with large sample size, but the use of non-informative prior distributions on the test parameters gave better fit with small sample size. For binary covariance models the choice of prior distribution plays an important role in the estimation procedure, especially with large sample sizes, where the posterior summaries of interest do not have important changes assuming informative or non-informative prior distributions. With small sample sizes and binary covariance structure, we observed better fit assuming non-informative prior distributions on the test parameters and informative prior distributions on covariance parameter.

It is important to point out that we could consider other copula families introduced in the literature to model dependence between diagnostic tests. A special case is given by the Clayton copula which is useful when the dependence is mainly concentrated in the lower tail or the Frank copula which is radial symmetric. The use of these other copulas in dependent diagnostic tests will be the goal of a future work, since an appropriate choice is essential in order to get an optimal result.

Acknowledgments

The second author was supported by grants of Conselho Nacional de Pesquisas (CNPq), Brazil. This paper is a part of the doctoral in statistics project of the first author.

[Recibido: noviembre de 2011 — Aceptado: abril de 2012]

References

Ali, S., Moodambail, A., Hamrah, E., Bin-Nakhi, H. & Sadeq, S. (2007), 'Reliability of rapid dipstick test in detecting urinary tract infection in symptomatic children', *Kuwait Medical Journal* **39**, 36–38.

Amblard, C. & Girard, S. (2002), 'Symmetry and dependence properties within a semiparametric family of bivariate copulas', *Journal of Non-parametric Statistics* **14**, 715–727.

Amblard, C. & Girard, S. (2005), 'Estimation procedures for semiparametric family of bivariate copulas', *Journal of Computational and Graphical Statistics* **14**, 363–377.

Amblard, C. & Girard, S. (2008), 'A new extension of bivariate FGM copulas', *Metrika* **70**, 1–17.

Best, N., Cowles, M. & Vines, S. (1995), *CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, version 0.3;*, MRC Biostatistics Unit, Cambridge, U.K.

Böhning, D. & Patilea, V. (2008), 'A capture-recapture approach for screening using two diagnostic tests with availability of disease status for the positives only', *Journal of the American Statistical Association* **103**, 212–221.

Brenner, H. (1996), 'How independent are multiple diagnosis classifications?', *Statistics in Medicine* **15**, 1377–1386.

Dendukuri, N. & Joseph, L. (2001), 'Bayesian approaches to modelling the conditional dependence between multiple diagnostic tests', *Biometrics* **57**, 158–167.

Georgiadis, M., Johnson, W. & Gardner, I. (2003), 'Correlation adjusted estimation of sensitivity and specificity of two diagnostic tests', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **52**, 63–76.

Gumbel, E. (1960), 'Bivariate exponential distributions', *Journal of the American Statistical Association* **55**, 698–707.

Joseph, L., Gyorkos, T. & Coupal, L. (1995), 'Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard', *American Journal of Epidemiology* **141**, 263–272.

Martinez, E., Achcar, J. & Louzada, N. (2005), 'Bayesian estimation of diagnostic tests accuracy for semi-latent data with covariates', *Journal of Biopharmaceutical Statistics* **15**, 809–821.

Nelsen, R. (1999), *An Introduction to Copulas*, Springer Verlag, New York.

Qu, Y. & Hadgu, A. (1998), 'A model for evaluating sensitivity and specificity for correlated diagnostic test in efficacy studies with an imperfect reference test', *Journal of the American Statistical Association* **93**, 920–928.

Smith, D., Bullock, A. & Catalona, W. (1997), 'Racial differences in operating characteristics of prostate cancer screening tests', *Journal of Urology* **158**, 1861–1865.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003), *Winbugs User Manual version 1.4*, MRC Biostatistics Unit, Cambridge, U.K.

Thibodeau, L. (1981), 'Evaluating diagnostic tests', *Biometrics* **37**.

Torrance-Rynard, V. & Walter, S. (1997), 'Effects of dependent errors in the assessment of diagnostic tests performance', *Statistics in Medicine* **16**.

Tovar, J. R. (2012), 'Eliciting beta prior distributions for binomial sampling', *Revista Brasileira de Biometría* **30**, 159–172.

Vacek, P. (1985), 'The effect of conditional dependence on the evaluation of diagnostic tests', *Biometrics* **41**.

Walter, S. & Irwig, L. (1988), 'Estimation of test error rates disease prevalence and relative risk from misclassified data: a review', *Journal of Clinical Epidemiology* **41**.

Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 349 a 370

Fatigue Statistical Distributions Useful for Modeling Diameter and Mortality of Trees

Distribuciones estadísticas de fatiga útiles para modelar diámetro y mortalidad de árboles

Víctor Leiva[1,a], M. Guadalupe Ponce[2,b], Carolina Marchant[1,c], Oscar Bustos[3,d]

[1] Departamento de Estadística, Universidad de Valparaíso, Valparaíso, Chile
[2] Instituto de Matemáticas y Física, Universidad de Talca, Talca, Chile
[3] Departamento de Producción Forestal, Universidad de Talca, Talca, Chile

Abstract

Mortality processes and the distribution of the diameter at breast height (DBH) of trees are two important problems in forestry. Trees die due to several factors caused by stress according to a phenomenon similar to material fatigue. Specifically, the force (rate) of mortality of trees quickly increases at a first stage and then reaches a maximum. In that moment, this rate slowly decreases until stabilizing at a constant value in the long term establishing a second stage of such a rate. Birnbaum-Saunders (BS) distributions are models that have received considerable attention currently due to their interesting properties. BS models have their genesis from a problem of material fatigue and present a failure or hazard rate (equivalent to the force of mortality) that has the same behavior as that of the DBH of trees. Then, BS distributions have arguments that transform them into models that can be useful in forestry. In this paper, we present a methodology based on BS distributions associated with this forest thematic. To complete this study, we perform an application of five real DBH data sets (some of them unpublished) that provides statistical evidence in favor of the BS methodology in relation to the forestry standard methodology. This application provides valuable financial information that can be used for making decisions in forestry.

***Key words*:** data analysis, force of mortality, forestry, hazard rate.

Resumen

[a] Professor. E-mail: victor.leiva@uv.cl
[b] Assistant profesor. E-mail: gponce@utalca.cl
[c] Assistant profesor. E-mail: carolina.marchant@uv.cl
[d] Assistant professor. E-mail: obustos@utalca.cl

Los procesos de mortalidad y la distribución del diámetro a la altura del pecho (DAP) de árboles son dos problemas importantes en el área forestal. Los árboles mueren debido a diversos factores causados por estrés mediante un fenómeno similar a la fatiga de materiales. Específicamente, la fuerza (tasa) de mortalidad de árboles crece rápidamente en una primera fase y luego alcanza un máximo, momento en el que comienza una segunda fase en donde esta tasa decrece lentamente estabilizándose en una constante en el largo plazo. Distribuciones Birnbaum-Saunders (BS) son modelos que han recibido una atención considerable en la actualidad debido a sus interesantes propiedades. Modelos BS nacen de un problema de fatiga de materiales y poseen una tasa de fallas (equivalente a la fuerza de mortalidad) que se comporta de la misma forma que ésa del DAP de árboles. Entonces, distribuciones BS poseen argumentos que las transforman en modelos que puede ser útiles en las ciencias forestales. En este trabajo, presentamos una metodología basada en la distribución BS asociada con esta temática forestal. Para finalizar, realizamos una aplicación con cinco conjuntos de datos reales (algunos de ellos no publicados) de DAP que proporciona una evidencia estadística en favor de la metodología BS en relación a la metodología estándar usada en ciencias forestales. Esta aplicación entrega información que puede ser valiosa para tomar decisiones forestales.

***Palabras clave*:** análisis de datos, fuerza de mortalidad, silvicultura, tasa de riesgo.

1. Introduction

The determination of the statistical distribution of the diameter at breast height (DBH) of trees, and its relationship to the age, composition, density and geographical location where a forest is localized are valuable information for different purposes (Bailey & Dell 1973, Santelices & Riquelme 2007). Specifically, the distribution of the DBH is frequently used to determine the volume of wood from a stand allowing us to make decisions about: (i) productivity (quantity); (ii) diversity of products (quality); (iii) tree ages (mortality); and (iv) harvest policy and trees pruning (regeneration). Then, to know the DBH distribution may help to plan biological and financial management aspects of a forest in a more efficient way (Rennolls, Geary & Rollinson 1985). For example, trees with a large diameter are used for wood production, while trees with a small diameter are used for cellulose production. Thus, the four mentioned concepts (quality, quantity, mortality and regeneration) propose a challenge to postulate models that allow us to describe the forest behavior based on the DBH distribution.

Several statistical distributions have been used in the forestry area mainly to model the DBH. These distributions (in chronological order) are the models:

(i) Exponential (Meyer 1952, Schmelz & Lindsey 1965);

(ii) Gamma (Nelson 1964);

(iii) Log-normal (Bliss & Reinker 1964);

(iv) Beta (Clutter & Bennett 1965, McGee & Della-Bianca 1967, Lenhart & Clutter 1971, Li, Zhang & Davis 2002, Wang & Rennolls 2005);

(v) Weibull (Bailey & Dell 1973, Little 1983, Rennolls et al. 1985, Zutter, Oderwald, Murphy & Farrar 1986, Borders, Souter, Bailey & Ware 1987, McEwen & Parresol 1991, Maltamo, Puumalinen & Päivinen 1995, Pece, de Benítez & de Galíndez 2000, García-Güemes, Cañadas & Montero 2002, Wang & Rennolls 2005, Palahí, Pukkala & Trasobares 2006, Podlaski 2006);

(vi) Johnson SB (Hafley & Schreuder 1977, Schreuder & Hafley 1977);

(vii) Log-logistic (Wang & Rennolls 2005);

(viii) Burr XII (Wang & Rennolls 2005) and

(ix) Birnbaum-Saunders (BS) (Podlaski 2008).

The most used distribution is the Weibull model and the most recent is the BS model. In spite of the wide use of different statistical distributions to describe the DBH, the model selection has been based in empirical arguments supported by goodness-of-fit methods and not by theoretical arguments that justify its use. In order to propose DBH distributions with better arguments, mortality models based on cumulative stress can be considered (Podlaski 2008).

A statistical distribution useful for describing non-negative data that has recently received considerable attention is the BS model. This two-parameter distribution is unimodal and positively skewed. For more details about the BS distribution, see Birnbaum & Saunders (1969) and Johnson, Kotz & Balakrishnan (1995, pp. 651-663). The interest for the BS distribution is due to its theoretical arguments based on the physics of materials, its properties and its relation to the normal distribution. Some extensions and generalization of the BS distributions are attributed to Díaz-García & Leiva (2005); Vilca & Leiva (2006); Guiraud, Leiva & Fierro (2009). In particular, the BS-Student-t distribution has been widely studied (Azevedo, Leiva, Athayde & Balakrishnan 2012). Although BS distributions have their origin in engineering, these have been applied in several other fields, such as environmental sciences and forestry (Leiva, Barros, Paula & Sanhueza 2008, Podlaski 2008, Leiva, Sanhueza & Angulo 2009, Leiva, Vilca, Balakrishnan & Sanhueza 2010, Leiva, Athayde, Azevedo & Marchant 2011, Vilca, Santana, Leiva & Balakrishnan 2011, Ferreira, Gomes & Leiva 2012, Marchant, Leiva, Cavieres & Sanhueza 2013). Podlaski (2008) employed the BS model to describe DBH data for silver fir (*Abies alba* Mill.) and European beech (*Fagus sylvatica* L.) from a national park in Poland, using theoretical arguments. In addition, based on goodness-of-fit methods, he discovered that the BS distribution was the model that best described these data, displacing the Weibull distribution.

The aims of the present work are: (i) to introduce a methodology based on BS distributions (one of them being novel) for describing DBH data that can be useful for making decisions in forestry and (ii) to carry out practical applications of real DBH data sets (some of them unpublished) that illustrate this methodology. The article is structured as follows: In the second section, we explain the methods employed in this study, including a theoretical justification for the use of the BS distribution to model DBH data. In the third section, we establish an application with five real data sets of DBH using a methodology based on BS distributions. This methodology furnishes statistical evidence in its favor, in relation to the standard methodology used in forestry. This application provides valuable financial information that can be used for making decisions in forestry. Finally, we sketch some discussions and conclusions.

2. Methods

2.1. A Fatigue Model

The BS distribution is based on a physical argument that produces fatigue in the materials (Birnbaum & Saunders 1969). This argument is the Miner or cumulative damage law (Miner 1945). Birnbaum & Saunders (1968) provided a probabilistic interpretation of this law. The BS or fatigue life distribution was obtained from a model that shows failures to occur due to the development and growing of a dominant crack provoked by stress. This distribution describes the total time elapsed until a type of cumulative damage inducted by stress exceeds a threshold of resistance of the material producing its failure or rupture. Birnbaum & Saunders (1969) demonstrated that the failure rate (hazard rate or force of mortality) associated with their model has two phases. During the first phase, this rate quickly increases until a maximum point (change or critical point) and then a second phase starts when the failure rate begins to slowly decrease until it is stabilized at a constant greater than zero. Fatigue processes have failure rates which usually present in this way. In addition, these processes can be divided into three stages:

(A1) The beginning of an imperceptible fissure;

(A2) The growth and propagation of the fissure, which provokes a crack in the material specimen due to cyclic stress and tension; and

(A3) The rupture or failure of the material specimen due to fatigue.

The stage (A3) occupies a negligible lifetime. Therefore, (A2) contains most of the time of the fatigue life. For this reason, statistical models for fatigue processes are primarily concerned with describing the random variation of lifetimes associated with (A2) through two-parameter life distributions. These parameters allow those specimens subject to fatigue to be characterized and at the same time predicting their behavior under different force, stress and tension patterns.

Having explained the physical framework of the genesis of the BS distribution, it is now necessary to make the statistical assumptions. Birnbaum & Saunders (1969) used the knowledge of certain type of materials failure due to fatigue to develop their model. The fatigue process that they used was based on the following:

(B1) A material specimen is subjected to cyclic loads or repetitive shocks, which produce a crack or wear in this specimen;

(B2) The failure occurs when the size of the crack in the material specimen exceeds a certain level of resistance (threshold), denoted by ω;

(B3) The sequence of loads imposed in the material is the same from one cycle to another;

(B4) The crack extension due to a load l_i (X_i say) during the jth cycle is a random variable (r.v.) governed by all the loads l_j, for $j < i$, and by the actual crack extension that precedes it;

(B5) The total size of the crack due to the jth cycle (Y_i say) is an r.v. that follows a statistical distribution of mean μ and variance σ^2; and

(B6) The sizes of cracks in different cycles are independent.

Notice that the total crack size due to the $(j+1)$th cycle of load is $Y_{j+1} = X_{jm+1} + \cdots + X_{jm+m}$, for $j, m = 0, 1, 2, \ldots$ Thus, the accumulated crack size at the end of the nth stress cycle is $S_n = \sum_{j=1}^{n} Y_j$. Then, based on it, (B1)-(B6) and the central limit theorem, we have $Z_n = [S_n - n\mu]/\sqrt{n\,\sigma^2} \dot{\sim} \mathrm{N}(0,1)$, as n approaches to ∞, i.e., Z_n follows approximately a standard normal distribution. Now, let N be the number of stress cycles until the specimen fails. The cumulative distribution function (c.d.f.) of N, based on the total probability theorem, is $\mathrm{P}(N \leq n) = \mathrm{P}(N \leq n, S_n > \omega) + \mathrm{P}(N \leq n, S_n \leq \omega) = \mathrm{P}(S_n > \omega) + \mathrm{P}(N \leq n, S_n \leq \omega)$. Notice that $\mathrm{P}(N \leq n, S_n \leq \omega) > 0$, because S_n follows approximately a normal distribution, but this probability is negligible, so that $\mathrm{P}(N \leq n) \approx \mathrm{P}(S_n > \omega)$, and hence

$$\mathrm{P}(N \leq n) \approx \mathrm{P}\left(\tfrac{S_n - n\mu}{\sigma\sqrt{n}} > \tfrac{\omega - n\mu}{\sigma\sqrt{n}}\right) = \Phi\left(\tfrac{\sqrt{\omega\mu}}{\sigma}\left[\sqrt{\tfrac{n}{\omega/\mu}} - \sqrt{\tfrac{\omega/\mu}{n}}\right]\right) \tag{1}$$

where $\Phi(\cdot)$ is the normal standard c.d.f. However, we must suppose the probability that Y_j given in (B5) takes negative values is zero. Birnbaum & Saunders (1969) used (1) to define their distribution, considering the discrete r.v. N as a continuous r.v. T, i.e., the number of stress cycles until to fail N is replaced by the total time until to fail T and the nth cycle by the time t. Thus, considering the reparameterization $\alpha = \sigma/\sqrt{\omega\mu}$ and $\beta = \omega/\mu$, and that (1) is exact instead of approximated, we obtain the c.d.f. of the BS distribution for the fatigue life with shape (α) and scale (β) parameters given by

$$F_T(t) = \Phi\left(\tfrac{1}{\alpha}\left[\sqrt{\tfrac{t}{\beta}} - \sqrt{\tfrac{\beta}{t}}\right]\right), \quad t > 0,\, \alpha > 0,\, \beta > 0 \tag{2}$$

To suppose (1) is exact, it means to suppose Y_j follows exactly a $\mathrm{N}(\mu, \sigma^2)$ distribution in (B5).

2.2. Birnbaum-Saunders Distributions

If an r.v. T has a c.d.f. as in (2), then it follows a BS distribution with shape ($\alpha > 0$) and scale ($\beta > 0$) parameters, which is denoted by $T \sim \mathrm{BS}(\alpha, \beta)$. Here, the parameter β is also the median. Hence, BS (T say) and normal standard (Z say) r.v.'s are related by

$$T = \beta \left[\tfrac{\alpha\, Z}{2} + \sqrt{\left\{ \tfrac{\alpha\, Z}{2} \right\}^2 + 1} \right]^2 \sim \mathrm{BS}(\alpha, \beta) \text{ and } Z = \tfrac{1}{\alpha} \left[\sqrt{\tfrac{T}{\beta}} - \sqrt{\tfrac{\beta}{T}} \right] \sim \mathrm{N}(0, 1) \qquad (3)$$

In addition, $W = Z^2$ follows a χ^2 distribution with one degree of freedom (d.f.), denoted by $W \sim \chi^2(1)$. The probability density function (p.d.f.) of T is

$$f_T(t) = \tfrac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2\,\alpha^2} \left[\tfrac{t}{\beta} + \tfrac{\beta}{t} - 2 \right] \right) \tfrac{1}{2\,\alpha\,\beta} \left[\left\{ \tfrac{t}{\beta} \right\}^{-1/2} + \left\{ \tfrac{t}{\beta} \right\}^{-3/2} \right], \quad t > 0 \quad (4)$$

The qth quantile of T is $t_q = \beta[\alpha z_q/2 + \sqrt{\{\alpha z_q/2\}^2 + 1}]^2$, for $0 < q < 1$, where $t_q = F_T^{-1}(q)$, with $F_T^{-1}(\cdot)$ being the inverse c.d.f. of T, and z_q the N(0, 1) qth quantile. The mean, variance and coefficient of variation (CV) of T are

$$\mathrm{E}[T] = \tfrac{\beta}{2} \left[2 + \alpha^2 \right], \ \mathrm{V}[T] = \tfrac{\beta^2 \alpha^2}{4} \left[4 + 5\alpha^2 \right] \text{ and } \mathrm{CV}[T] = \tfrac{\alpha\sqrt{4+5\,\alpha^2}}{2+\alpha^2} \qquad (5)$$

Although the BS distribution can be useful to model the DBH, there are several reasons to consider that the DBH distribution could start from a value greater than zero. In such a situation, a shifted version of the BS (ShBS) distribution, with shape ($\alpha > 0$), scale ($\beta > 0$) and shift ($\gamma \in \mathbb{R}$) parameters, is needed, which is denoted by $T \sim \mathrm{ShBS}(\alpha, \beta, \gamma)$. Leiva et al. (2011) characterized this distribution assuming that if $T = \beta[\alpha\, Z/2 + \sqrt{\{\alpha\, Z/2\}^2 + 1}]^2 \sim \mathrm{ShBS}(\alpha, \beta, \gamma)$, then, $Z = [1/\alpha][\sqrt{\{T - \gamma\}/\beta} - \sqrt{\beta/\{T - \gamma\}}] \sim \mathrm{N}(0, 1)$ and so again $W = Z^2 \sim \chi^2(1)$. Therefore, in this case, the p.d.f. and c.d.f. of T are

$$f_T(t) = \tfrac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2\,\alpha^2} \left[\tfrac{t-\gamma}{\beta} + \tfrac{\beta}{t-\gamma} - 2 \right] \right) \tfrac{1}{2\,\alpha\,\beta} \left[\left\{ \tfrac{t-\gamma}{\beta} \right\}^{-1/2} + \left\{ \tfrac{t-\gamma}{\beta} \right\}^{-3/2} \right] \quad (6)$$

and $F_T(t) = \Phi([1/\alpha][\sqrt{\{t - \gamma\}/\beta} - \sqrt{\beta/\{t - \gamma\}}])$, for $t > \gamma$, respectively. In addition, the qth quantile of T is similar to that from the non-shifted case plus the value γ at the end of such an expression. The mean, variance and CV of T are now

$$\mathrm{E}\,[T] = \tfrac{\beta}{2} \left[2 + \alpha^2 + \tfrac{2\gamma}{\beta} \right], \ \mathrm{V}[T] = \tfrac{\beta^2\, \alpha^2}{4} [4 + 5\alpha^2] \text{ and } \mathrm{CV}[T] = \tfrac{\alpha\beta\sqrt{4+5\,\alpha^2}}{\beta[2+\alpha^2]+2\gamma} \qquad (7)$$

2.3. Birnbaum-Saunders-t-Student Distributions

If an r.v. T follows a BS-t distribution with shape ($\alpha > 0$, $\nu > 0$) and scale ($\beta > 0$) parameters, then the notation $T \sim \text{BS-}t(\alpha, \beta; \nu)$ is used. Thus, if $T = \beta[\alpha Z/2 + \sqrt{\{\alpha Z/2\}^2 + 1}]^2 \sim \text{BS-}t(\alpha, \beta; \nu)$, then $Z = [1/\alpha][\sqrt{T/\beta} - \sqrt{\beta/T}] \sim t(\nu)$, with ν d.f., and $W = Z^2 \sim \mathcal{F}(1,\nu)$. Therefore, in this case, the p.d.f. and c.d.f. of T are

$$\begin{aligned} f_T(t;\nu) &= \tfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left[1 + \left\{\tfrac{t}{\beta} + \tfrac{\beta}{t} - 2\right\}/\{2\,\alpha^2\,\nu\}\right]^{-\frac{[\nu+1]}{2}} \tfrac{1}{2\,\alpha\,\beta} \left[\left\{\tfrac{t}{\beta}\right\}^{-1/2} + \left\{\tfrac{t}{\beta}\right\}^{-3/2}\right] \\ F_T(t;\nu) &= \Phi_t(t) = \tfrac{1}{2}\left[1 + I_{\frac{[1/\alpha^2][t/\beta+\beta/t-2]}{[1/\alpha^2][t/\beta+\beta/t-2]+\nu}}\left(\tfrac{1}{2},\tfrac{\nu}{2}\right)\right], \quad t > 0 \end{aligned} \tag{8}$$

respectively, where $I_x(a,b) = [\int_0^x t^{a-1}\{1-t\}^{b-1}\,\mathrm{d}t]/\int_0^1 t^{a-1}\{1-t\}^{b-1}\,\mathrm{d}t$ is the incomplete beta function ratio. The qth quantile of T is

$$t_q = \beta[\alpha z_q/2 + \sqrt{\{\alpha z_q/2\}^2 + 1}]^2,$$

where z_q is the qth quantile of the $t(\nu)$ distribution. The mean, variance and CV of T are now

$$\mathrm{E}[T] = \tfrac{\beta}{2}\left[2 + A\,\alpha^2\right],\ \mathrm{V}[T] = \tfrac{\beta^2\alpha^2}{4}\left[4A + 5B\,\alpha^2\right] \text{ and } \mathrm{CV}[T] = \tfrac{\alpha\sqrt{4A+5B\,\alpha^2}}{2+A\,\alpha^2} \tag{9}$$

where $A = \nu/[\nu-2]$, for $\nu > 2$, and $B = \nu^2[\nu-1]/[\{\nu-6\}\{\nu-2\}^2]$, for $\nu > 6$.

Such as in the case of the BS distribution, we can define a new shifted version of the BS-t (ShBS-t) distribution, with shape ($\alpha > 0$, $\nu > 0$), scale ($\beta > 0$) and shift ($\gamma \in \mathbb{R}$) parameters, which is denoted by $T \sim \text{ShBS-}t(\alpha,\beta,\gamma;\nu)$. Thus, if $T = \beta[\alpha\,Z/2 + \sqrt{\{\alpha\,Z/2\}^2+1}]^2 \sim \text{ShBS-}t(\alpha,\beta,\gamma;\nu)$, then $Z = [1/\alpha][\sqrt{\{T-\gamma\}/\beta} - \sqrt{\beta/\{T-\gamma\}}] \sim t(\nu)$ and so again $W = Z^2 \sim \mathcal{F}(1,\nu)$. Therefore, in this case, the p.d.f. and c.d.f. of T are

$$\begin{aligned} f_T(t;\nu) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left[1 + \left\{\frac{t-\gamma}{\beta} + \frac{\beta}{t-\gamma} - 2\right\}/\{2\,\alpha^2\,\nu\}\right]^{-\frac{[\nu+1]}{2}} \\ &\qquad \frac{1}{2\,\alpha\,\beta}\left[\left\{\frac{t-\gamma}{\beta}\right\}^{-1/2} + \left\{\frac{t-\gamma}{\beta}\right\}^{-3/2}\right] \\ F_T(t;\nu) &= \Phi_t(t-\gamma) = \tfrac{1}{2}\left[1 + I_{\frac{[1/\alpha^2][\{t-\gamma\}/\beta+\beta/\{t-\gamma\}-2]}{[1/\alpha^2][\{t-\gamma\}/\beta+\beta/\{t-\gamma\}-2]+\nu}}\left(\tfrac{1}{2},\tfrac{\nu}{2}\right)\right], \quad t > \gamma \end{aligned} \tag{10}$$

respectively. The qth quantile of T is obtained in an analogous way as in the ShBS case. The mean, variance and CV of T respectively are now

$$\mathrm{E}\,[T] = \tfrac{\beta}{2}\left[2 + A\alpha^2 + \tfrac{2\gamma}{\beta}\right], \quad \mathrm{V}[T] = \tfrac{\beta^2\,\alpha^2}{4}[4A + 5B\alpha^2] \text{ and } \mathrm{CV}[T] = \tfrac{\alpha\beta\sqrt{4A+5B\,\alpha^2}}{\beta[2+A\alpha^2]+2\gamma} \tag{11}$$

where A and B are as given in (9).

2.4. Force of Mortality

Hazard can be defined as the probability that a dangerous event that could develop into an emergency or disaster. Origin of this event can be provoked by an environmental agent that could have an adverse effect. Then, hazard is a chance and not a real fact. This means that hazard should be evaluated as the frequency or intensity of an r.v., e.g., the DBH. A useful function in hazard analysis is the hazard rate (h.r.) or force of mortality defined as $h_T(t) = f_T(t)/[1-F_T(t)]$, where $f_T(\cdot)$ and $F_T(\cdot)$ are the p.d.f. and c.d.f. of the r.v. T, respectively (Johnson et al. 1995). The h.r. can be interpreted as the velocity or propensity that a specific event occurs, expressed per unit of the r.v. (in general, time, but in the case of DBH is a unit of length). A characteristic of the h.r. is that it allows us to identify statistical distributions. For example, distributions with shapes similar for their p.d.f.'s could have h.r.'s which are totally different (such as is the case with the BS and Weibull distributions). As mentioned in Subsection 2.1, the BS distribution has a non-monotone h.r., because it is first increasing, until a critical point in its phase I and then it is decreasing until its stabilization at a positive constant greater than zero in its phase II. Specifically, for the BS case, if t approaches to ∞, then the h.r. $h_T(t)$ converges to the constant $1/[2\alpha^2\beta] > 0$, for $t > 0$. Figure 1(a) shows the behavior of the BS p.d.f. for some values of the shape parameter (α). Notice that, as α decreases, the shape of the BS p.d.f. is approximately symmetrical. Graphical plots for different values of the parameter β were not considered, because this parameter only modifies the scale. Figure 1(b) displays the behavior of the BS h.r. for some values of α. Notice that, as α decreases, the shape of the h.r. is approximately increasing. For a recent study of the BS-t h.r., the interested reader is referred to (Azevedo et al. 2012).

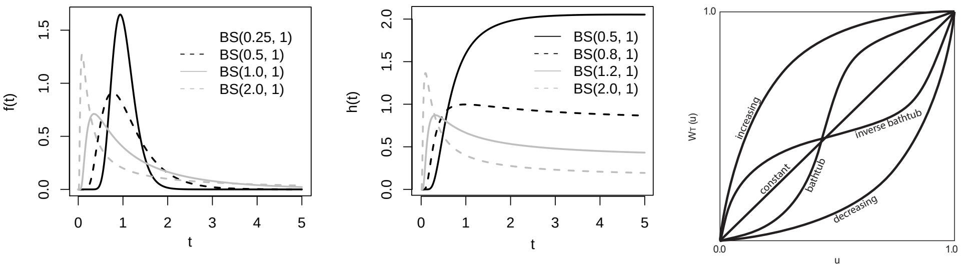


FIGURE 1: BS p.d.f. (left), BS h.r. (center) and theoretical TTT plots (right) for the indicated values.

When continuous data are analyzed (for example, DBH data) and we want to propose a distribution for modeling such data, one usually constructs a histogram. This graphical plot is an empirical approximation of the p.d.f. However, it is always convenient to look also for the h.r. of the data. The problem is that to approximate empirically the h.r. is not an easy task. A tool that is being used for this purpose is the total time on test (TTT) plot, which allows us to have an idea about the shape of the h.r. of an r.v. and, as consequence, about the distribution that the data follows. The TTT function of the r.v. T is given by $H_T^{-1}(u) = \int_0^{F_T^{-1}(u)}[1 - F_T(y)]\,dy$ and its scaled version by $W_T(u) = H_T^{-1}(u)/H_T^{-1}(1)$, for $0 \leq u \leq 1$, where once again $F_T^{-1}(\cdot)$ is the inverse c.d.f. of T. Now, $W_T(\cdot)$ can be approximated allowing us to construct the empirical scaled TTT curve by plotting the points $\big[k/n, W_n(k/n)\big]$, where

$W_n(k/n) = [\sum_{i=1}^{k} T_{(i)} + [n-k]\, T_{(k)}]/\sum_{i=1}^{n} T_{(i)}$, for $k = 1, \ldots, n$, with $T_{(i)}$ being the ith order statistic, for $i = 1, \ldots, n$. Specifically, if the TTT plot is concave (convex), then a model with increasing (decreasing) h.r. is appropriate. Now, if the TTT plot is first concave (convex) and then convex (concave), an inverse bathtub (IBT) shaped (bathtub –BT–) h.r. must be considered. If the TTT plot is a straight line, then the exponential distribution must be used. For example, the normal distribution is in the increasing h.r. class, while the gamma and Weibull distributions admit increasing, constant and decreasing h.r.'s. However, the BS and log-normal distributions have non-monotone h.r.'s, because these are initially increasing until their change points and then decreasing (IBT shaped h.r.) to zero, in the log-normal case, or to a constant greater than zero, in the BS case. This last case must be highlighted because biological entities (such as humans, insects and trees) have h.r.'s of this type (Gavrilov & Gavrilova 2001). In Figure 1(c), we see several theoretical shapes of the TTT plot, which correspond to a particular type of h.r. (Aarset 1987).

2.5. Model Estimation and Checking

Parameters of the BS, ShBS, BS-t and ShBS-t distributions can be estimated by the maximum likelihood (ML) method adapted by a non-failing algorithm (Leiva et al. 2011). To obtain the estimates of the parameters of these distributions, their corresponding likelihood functions must be constructed using (4), (6), (8) and (10), respectively. When these parameters have been estimated, we must check goodness-of-fit of the model to the data. Distributions used for describing DBH data can be compared using model selection criteria based on loss of information such as Akaike (AIC) and Bayesian (BIC) information criteria. AIC and BIC allows us to compare models for the same model and they are given by $\mathrm{AIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p$ and $\mathrm{BIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + p\log(n)$, where $\ell(\widehat{\boldsymbol{\theta}})$ is the logarithm of the likelihood function (log-likelihood) of the model with vector of parameters $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, n is the size of the sample and p is the number of model parameters. For the case of BS, ShBS, BS-t and ShBS-t models, as mentioned, $\ell(\boldsymbol{\theta})$ must be obtained by (4), (6), (8) and (10), respectively. AIC and BIC correspond to the log-likelihood function plus a component that penalizes such a function as the model has more parameters making it more complex. A model with a smaller AIC or BIC is better.

Differences between two values of the BIC are usually not very noticeable. Then, the Bayes factor (BF) can be used to highlight such differences, if they exist. Assume the data belongs to one of two possible models, according to probabilities P(Data | Model 1) and P(Data | Model 2), respectively. Given probabilities P(Model 1) and P(Model 2) = 1 − P(Model 1), the data produce conditional probabilities P(Model 1 | Data) and P(Model 2 | Data) = 1 − P(Model 1 | Data), respectively. The BF allows us to compare Model 1 (considered as correct) to Model 2 (to be contrasted with Model 1) and it is given by $B_{12} = \mathrm{P(Data \mid Model\ 1)}/\mathrm{P(Data \mid Model\ 2)}$, which can be approximated by $2\log(B_{12}) \approx 2\big[\ell(\widehat{\boldsymbol{\theta}}_1) - \ell(\widehat{\boldsymbol{\theta}}_2)\big] - [d_1 - d_2]\log(n)$, where $\ell(\widehat{\boldsymbol{\theta}}_k)$ is the log-likelihood function for the parameter $\boldsymbol{\theta}_k$ under the kth model evaluated at $\boldsymbol{\theta}_k = \widehat{\boldsymbol{\theta}}_k$, d_k is the dimension of $\boldsymbol{\theta}_k$, for $k = 1, 2$, and n is the sample size. Notice that the above approximation is computed by sustracting the BIC value from Model 2, given by $\text{BIC}_2 = -2\ell(\boldsymbol{\theta}_2) + d_2 \log(n)$, to the BIC value of Model 1, given by $\text{BIC}_1 = -2\ell(\boldsymbol{\theta}_1) + d_1 \log(n)$. In addition, notice that if Model 2 is a particular case of Model 1, then the procedure corresponds to applying the likelihood ratio (LR) test. In this case, $2\log(B_{12}) \approx \chi^2_{12} - \text{df}_{12}\, \log(n)$, where χ^2_{12} is the LR test statistic for testing Model 1 versus Model 2 and $\text{df}_{12} = d_1 - d_2$ are the d.f.'s associated with the LR test, so that one can obtain the corresponding p-value from $2\log(B_{12}) \dot{\sim} \chi^2(d_1 - d_2)$, with $d_1 > d_2$. The BF is informative, because it presents ranges of values in which the degree of superiority of one model with respect to another can be quantified. An interpretation of the BF is displayed in Table 1.

Table 1: Interpretation of $2\log(B_{12})$ associated with the BF.

$2\log(B_{12})$	Evidence in favor of Model 1
< 0	Negative (Model 2 is accepted)
$[0, 2)$	Weak
$[2, 6)$	Positive
$[6, 10)$	Strong
≥ 10	Very strong

2.6. Quantity and Quality of Wood

Because the DBH varies depending on the composition, density, geographic location and stand age, the diameter can be considered as an r.v. that we denote by T. As mentioned, information on the distribution of T in a forest plantation is an important element to quantify the products come from thinning and clearcutting activities. This information can help to plan the management and use of forest resources more efficiently. It is important to model the distribution of the DBH since this is the most relevant variable in determining the tree volume and then the forest production.

The forest volume quantification allows us to make decisions about the production and forest management, for example, to know when the forest should be harvested. However, the variable to maximize is diameter instead volume. Furthermore, the DBH is related to other variables such as cost of harvest, quality and product type. While the productivity is an important issue for timber industry, wood quality is also relevant in order to determine its use. Thus, volume and diameter distribution of trees determine what type of product will be obtained. For example, large diameter trees are used for saw wood and those of small diameter for pulpwood. This implies a financial analysis of forest harvest, i.e., how and when to harvest and what method to use. Studies from several types of climates and soils show trees growth as a function of the basal area. Making decisions using the forest basal area are related to pruning and thinning. These activities aim to improve tree growth and produce higher quality wood. The basal area of a tree is the imaginary basal area at breast height (1.3 m above ground level) given by

$$B = \frac{\pi}{4} T^2$$

where B is the basal area and T the DBH.

The sum of the individual basal area of all trees in one hectare leads to the basal area per hectare. However, it is the volume which allows for the planning of various forestry activities. There are several formulae to determine the volume of logs using the mean diameter measured without bark, and the log length. Volume allows for the planning of silvicultural and harvesting activities. In general, the formula used for the volume of a tree is given by

$$V = F\,B\,H = \frac{\pi}{4}\,F\,T^2\,H \tag{12}$$

where V is the tree volume, B its basal area, H its height and F the form factor, which is generally smaller than a value equal to one depending on the tree species.

2.7. Mortality and Tree Regeneration

The DBH is related to tree mortality, which is affected by stress factors such as light, nutrients, sunlight, temperature and water. The light and temperature can cause stress in minutes, whereas lack of water can cause stress in days or weeks. However, lack of nutrients in the soil can take months to generate stress. The mortality of a tree is similar to the material fatigue process described in Section 2.1, because the force of mortality of trees is growing rapidly in phase I, reaching a maximum and then decreases slowly until it is stabilized in phase II, which is consistent for almost all tree species.

Podlaski (2008) identified in a national park in Poland the following stress factors: (i) abiotic factors, such as severe weather (frost, hail, humidity, snow, temperature, wind), deficiency or excess of soil nutrients and toxic substances in air and soil, and (ii) biotic factors, such as bacteria (canker), fungi (dumping-off spots, root rots, rusts), insects and worms (nematodes), mycoplasma (elm phloem necrosis), parasitic plants (mistletoes) and viruses (elm mosaic). These factors caused the death of trees of the species *Abies alba.* From a theoretical point of view, the force of mortality of spruce could be more appropriately described by the h.r. of the BS distribution rather than using other distributions employed to model DBH. Podlaski (2008) indicated that mortality of spruce stand caused more openings within the stand and the canopy. Thus, with more spaces and gaps, trees of the species *Fagus sylvatica*, a kind that grows in temperate zones of the planet, tended to regenerate.

The regeneration process has been closely connected with the death of fir, whose speed in phase I also resulted in a rapid regeneration of beech, and the subsequent occurrence of understory vegetation in the stand. The decrease in the intensity of spruce mortality in phase II, as well as shading of soil by the understory, caused a gradual decrease in the intensity of the regeneration of beech. The stands generated by this process are characterized by a vertical structure of tree layers of different heights. These layers correspond to multiple layers of canopy whose statistical distribution of the DBH is asymmetric and positively skewed, as in the BS model. Most of the spruce stands had diameters of approximately 0.15 m to 0.35 m. The interruption of the regeneration process resulted in the death of these stands, which had a DBH of less than 0.1 m. The necessary condition for the creation of stands with DBH distributions approximated by the BS model is the simultaneous death of fir at all levels of the stand with regeneration of beech, i.e., a death that considers the different forest layers and has a similar degree of seasonality in the subsequent occurrence of the understory.

3. Application

Next, we apply the methodology outlined in this article using real data of the DBH and a methodology based on BS models. First, we perform an exploratory data analysis (EDA) of DBH. Then, based on this EDA, we propose statistical distributions to model the DBH. We use goodness-of-fit methods to find the more suitable distribution for modeling the DBH data under analysis. Finally, we make a confirmatory analysis and furnish information that can be useful to make financial and forestry decisions.

3.1. The Data Sets

The five DBH data sets to be analyzed are presented next. These data (all of them given in cm) are expressed in each case with the data frequency in parentheses and nothing when the frequency is equal to one.

Giant paradise (*Melia azedarach* L.) This is an exotic tree species originated from Asia and adapted to the province of Santiago del Estero, Argentina. Giant paradise produces wood of very good quality in a short time. We consider DBH data of giant paradise trees from four consecutive annual measurements collected since 1994 in 40 sites located at a stand in the Departamento Alberdi to the northwest of the province of Santiago del Estero, Argentina. Specifically, we use measurements collected at Site 7 due to the better conformation and reliability of the database (Pece et al. 2000). The data are: 16.5, 16.6, 17.8, 18.0 18.4, 18.5, 18.8, 18.9, 19.2, 19.3, 19.8, 20.3, 20.4, 20.6(2), 22.1, 22.2 23.5, 23.6, 26.7.

Silver fir (*Abies alba*). This is a species of tree of the pine family originated from mountainous regions in Europe. We consider DBH data of silver fir trees from 15 sites located at Świeta Katarzyna and Świety Krzyzÿ forest sections of the Świetokrzyski National Park, in Świetokrzyskie Mountains (Central Poland). Specifically, we use measurements collected at Site 10 due to similar reasons to that from *Melia azedarach* (Podlaski 2008). The data are: 11(2), 12, 13, 14(5), 15(4), 16(5), 17(4), 18(4), 19(3), 20(8), 21(4), 22(3), 23(4), 24(5), 25(6), 26(5), 27(5), 28(2), 29(5), 30(2), 31(7), 32(3), 33(2), 34(4), 35, 36(2), 37(2), 39(2), 40(3), 41(2), 42, 43(2), 44(3), 46(3), 47(2), 48, 50(2), 51, 52, 53, 54, 55, 56, 57, 59, 61, 66, 70, 89, 97.

Loblolly pine (*Pinus taeda* L.) This variety of tree is one of several native pines at the Southeastern of the United States (US). The data set corresponds to DBH of 20 year old trees from a plantation in the Western Gulf Coast of the US (McEwen & Parresol 1991). The data are: 6.2, 6.3, 6.4, 6.6(2), 6.7, 6.8, 6.9(3), 7.0(2), 7.1, 7.2(2), 7.3(3), 7.4(4), 7.6(2), 7.7(3), 7.8, 7.9(4), 8.1(4), 8.2(3), 8.3(3), 8.4, 8.5(3), 8.6(4), 8.7, 8.8(2), 8.9(3), 9.0(4), 9.1(5), 9.5(2), 9.6, 9.8(3), 10.0(2), 10.1, 10.3.

Ruíl (*Nothofagus alessandrii* Espinosa). This is an endemic species of central Chile, which is at risk of extinction. This tree variety is the older species of the family of the *Fagaceae* in the South Hemisphere, i.e., these stands are the older formations in South America. The data set of DBH was collected close to the locality of Gualleco, Región del Maule, Chile (Santelices & Riquelme 2007). The data are: 16(2), 18(2), 20(2), 22, 24, 26(2), 28, 30(2), 32, 34.

Gray birch (*Betula populifolia* Marshall). This is a perennial species from the US that has its best growth during spring and summer seasons. Gray birch has a short life in comparison with other plant species and a rapid growth rate. During its maturity (around 20 years), gray birch reaches an average height of 10 m. The data used for this study correspond to DBH of gray birch trees that are part of a natural forest of 16 hectares located at Maine, US. This data set was chosen because its collection is reliable and the database is complete, so it allows an adequate illustration for the purpose of this study. The data are: 10.5(5), 10.6, 10.7, 10.8(3), 10.9, 11.0, 11.2, 11.3(5), 11.4, 11.5(3), 11.6(2), 11.7(3), 11.9(2), 12.0(3), 12.1(3), 12.2(2), 12.3, 12.4(3), 12.5(3), 12.6, 12.7(2), 12.8(3), 12.9(5), 13.0(7), 13.1(4), 13.2(2), 13.3(3), 13.5(2), 13.6(3), 13.7(5), 13.8(2), 14.0(3), 14.1(4), 14.2(3), 14.3, 14.4(2), 14.5(5), 14.6(3), 14.8(4), 14.9(3), 15.0, 15.1(3), 15.2, 15.3(2), 15.6(2), 15.7(2), 15.8, 15.9(2), 16.0(2), 16.1(2), 16.4, 16.5, 16.6(2), 16.7, 16.9(2), 17.0(2), 17.5(2), 17.8(2), 18.3, 18.4, 18.5, 19.2, 19.4(2), 19.9(2), 20.0, 20.3, 20.5, 21.3, 21.9, 23.1, 24.4, 26.0, 28.4, 39.3.

We call S1, S2, S3, S4 and S5 to the DBH data sets of the varieties of *Melia azedarach*, *Abies alba*, *Pinus taeda*, *Nothofagus alessandrii*, and *Betula populifolia*, respectively.

3.2. Exploratory Data Analysis

Table 2 presents a descriptive summary of data sets S1-S5 that includes median, mean, standard deviation (SD), CV and coefficients of skewness (CS) and kurtosis (CK), among other indicators. Figure 2 shows histograms, usual and adjusted for asymmetrical data boxplots (Leiva et al. 2011) and TTT plots for S1-S5. From Table 2 and Figure 2, we detect distributions with positive skewness, different degrees of kurtosis, increasing and IBT shaped h.r.'s and a variable number of atypical DBH data. Specifically, the TTT plot of the DBH presented in Figure 2 (fifth panel) shows precisely a h.r. as those that the tree DBH should theoretically have and that coincides with the h.r. of the BS fatigue models. In addition, minimum values for S1-S5 indicate to us the necessity for considering a shift parameter in the modeling. As a consequence, based on this EDA, the different BS models presented in this paper seem to be good candidates for describing S1-S5, because they allow us to accommodate the different aspects detected in the EDA for these data sets. Particularly, BS-t and ShBS-t models allow us to accommodate atypical data in a robust statistically way. Also, BS distributions have a more appropriate h.r. to model such DBH data. This is a relevant aspect because DBH data have been widely modeled by the Weibull distribution. However, this distribution has a different h.r. to those that the tree DBH should theoretically have. Therefore, in the next section of model estimation and checking, we compare usual and shifted BS and Weibull models by means of a goodness-of-fit analysis in order to valuate whether this theoretical aspect is validated by the data or not.

Table 2: Descriptive summary of DBH for the indicated data set

Set	Median	Mean	SD	CV	CS	CK	Range	Minimum	Maximum	n
S1	19.55	20.09	2.53	12.58%	0.82	3.20	10.20	16.50	26.70	20
S2	27.00	30.68	14.85	48.42%	1.52	6.33	86.00	11.00	97.00	134
S3	8.20	8.19	1.01	12.37%	0.05	2.16	4.10	6.20	10.30	75
S4	24.00	24.00	5.95	24.80%	0.14	1.50	18.00	16.00	34.00	15
S5	13.70	14.54	3.61	24.85%	5.89	13.97	28.80	10.50	39.30	160

3.3. Model Estimation and Checking

As mentioned, the parameters of the BS, ShBS, BS-t, ShBS-t distributions can be estimated by the ML method adapted by a non-failing algorithm (Leiva et al. 2011). The estimation of the parameters of the BS distributions, as well as those of the usual and shifted Weibull distributions (as comparison), for S1-S5 are summarized in Table 3 together with the negative value of the corresponding log-likelihood function. In addition to the model selection criteria (AIC and BIC) presented in Section 2.1, the fit of the model to SI-S5 can be checked using the Kolmogorov-Smirnov test (KS). This test compares the empirical and theoretical c.d.f.'s (in this case of the BS and Weibull models). The p-values of the KS test, as well as the values of AIC, BIC and $2\log(B_{12})$ are also provided in Table 3. Based on the KS test and BF results presented in Table 3, we conclude that the BS distributions fit S1-S5 better than Weibull distributions. All this information supports the theoretical justification given in Section 2.

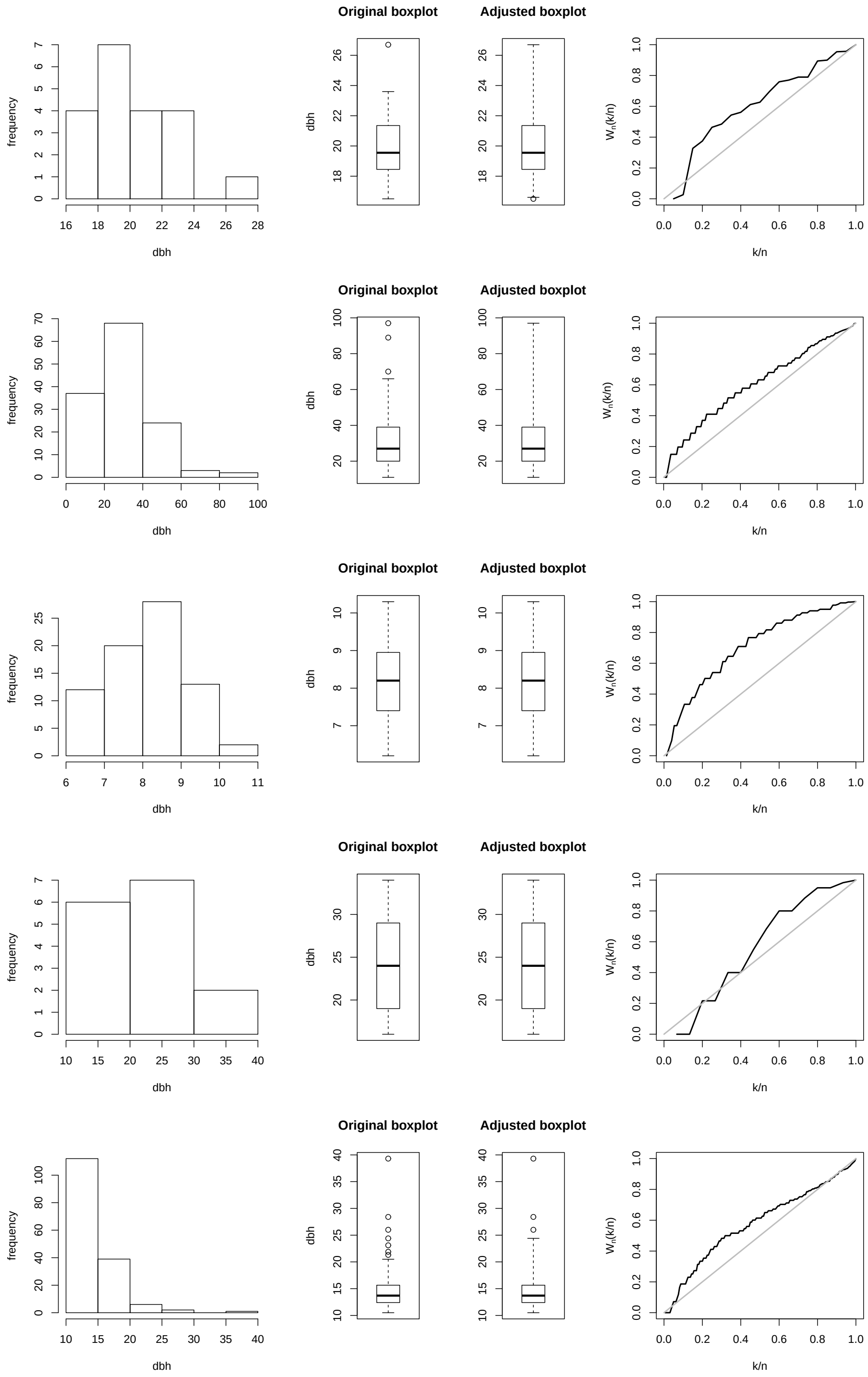


FIGURE 2: Histograms, usual and adjusted boxplots and TTT plots for S1 (first panel) to S5 (fifth panel).

Table 3: Indicators for the indicated data set and distribution.

Indicator	BS	BS-t	ShBS	ShBS-t	ShWeibull	Weibull
S1						
$\widehat{\alpha}$	0.118	0.117	0.374	0.443	1.587	7.736
$\widehat{\beta}$	19.950	19.934	6.161	3.260	4.810	21.228
$\widehat{\nu}$	-	87	-	1	-	-
$\widehat{\gamma}$	-	-	13.498	16.498	15.900	-
$-\ell(\widehat{\boldsymbol{\theta}})$	45.534	45.533	44.752	43.367	44.811	48.565
AIC	95.069	97.067	95.503	94.733	95.622	101.131
BIC	97.061	100.053	98.490	98.716	98.609	103.121
2 $\log(B_{12})$	-	2.992	1.429	1.655	1.548	6.060
KS p-value	0.806	0.829	0.986	0.963	0.882	0.385
S2						
$\widehat{\alpha}$	0.452	0.448	0.590	0.588	1.440	2.193
$\widehat{\beta}$	27.840	27.803	21.301	21.171	22.453	34.760
$\widehat{\nu}$	-	100	-	100	-	-
$\widehat{\gamma}$	-	-	5.666	5.778	10.358	-
$-\ell(\widehat{\boldsymbol{\theta}})$	525.820	525.889	524.255	524.306	524.772	540.361
AIC	1055.640	1057.777	1054.511	1056.612	1055.544	1084.721
BIC	1061.436	1066.472	1063.204	1068.203	1064.238	1090.518
2 $\log(B_{12})$	-	5.036	1.768	6.768	2.802	29.082
KS p-value	0.899	0.912	0.959	0.815	0.828	0.129
S3						
$\widehat{\alpha}$	0.124	0.123	0.124	0.124	2.514	8.952
$\widehat{\beta}$	8.125	8.127	8.125	8.127	2.635	8.636
$\widehat{\nu}$	-	100	-	100	-	-
$\widehat{\gamma}$	-	-	0.000	0.000	5.850	-
$-\ell(\widehat{\boldsymbol{\theta}})$	107.038	107.180	107.038	107.180	105.798	108.609
AIC	218.076	220.360	218.076	222.360	217.596	221.219
BIC	222.711	227.312	222.711	227.312	224.548	225.853
2 $\log(B_{12})$	-	4.601	-	4.601	1.837	3.142
KS p-value	0.876	0.874	0.876	0.874	0.918	0.840
S4						
$\widehat{\alpha}$	0.245	0.2445	0.394	0.585	2.837	4.685
$\widehat{\beta}$	23.298	23.304	14.625	9.207	16.568	26.282
$\widehat{\nu}$	-	100	-	1	-	-
$\widehat{\gamma}$	-	-	8.240	15.995	9.300	-
$-\ell(\widehat{\boldsymbol{\theta}})$	47.327	47.377	47.292	44.460	47.098	47.515
AIC	98.656	100.754	100.584	96.920	100.195	99.0294
BIC	100.070	102.878	102.708	99.752	104.084	100.446
2 $\log(B_{12})$	0.319	3.126	2.956	-	4.332	0.694
KS p-value	0.933	0.934	0.858	0.936	0.894	0.852
S5						
$\widehat{\alpha}$	0.208	0.151	0.727	0.563	1.502	3.467
$\widehat{\beta}$	14.230	13.817	3.774	4.232	4.749	15.920
$\widehat{\nu}$	-	4	-	8	-	-
$\widehat{\gamma}$	-	-	9.761	9.439	10.180	-
$-\ell(\widehat{\boldsymbol{\theta}})$	399.776	389.438	380.330	378.912	386.075	448.921
AIC	803.553	816.853	766.659	765.826	778.152	901.842
BIC	809.702	794.102	775.886	778.125	787.376	907.992
2 $\log(B_{12})$	33.817	18.216	-	2.239	11.490	132.107
KS p-value	0.052	0.400	0.530	0.773	0.467	< 0.001

Due to space limitations, in order to visualize the model fit to the DBH data, we only focus on S5. In addition, we only depict three plots corresponding to the shifted versions of the BS, BS-t and Weibull distributions, which are those that fit the data better. Comparison between the empirical (gray line) and ShBS, ShBS-t and ShWeibull theoretical (black dots) c.d.f.'s are shown in Figure 3. Histograms with the estimated ShBS, ShBS-t and ShWeibull p.d.f. curve are shown in Figure 4. Probability plots with "envelopes" based on the BS, BS-t and Weibull distributions for S5 are shown in Figure 5. The term "envelope" is a band for the probability plot built by means of a simulation process that facilitates the adjustment visualization. For example, for the BS distribution, this "envelope" is built using an expression given in (3). From Figure 5, we can see the excellent fit that the ShBS-t model provides to S5 and the bad fit provided by the ShWeibull model. Then, once the ShBS-t model has been considered as the most appropriate within the proposed distributions to model S5, we provide information that can be useful to make economical and forestry decisions based on this model and the methodology given in this study.

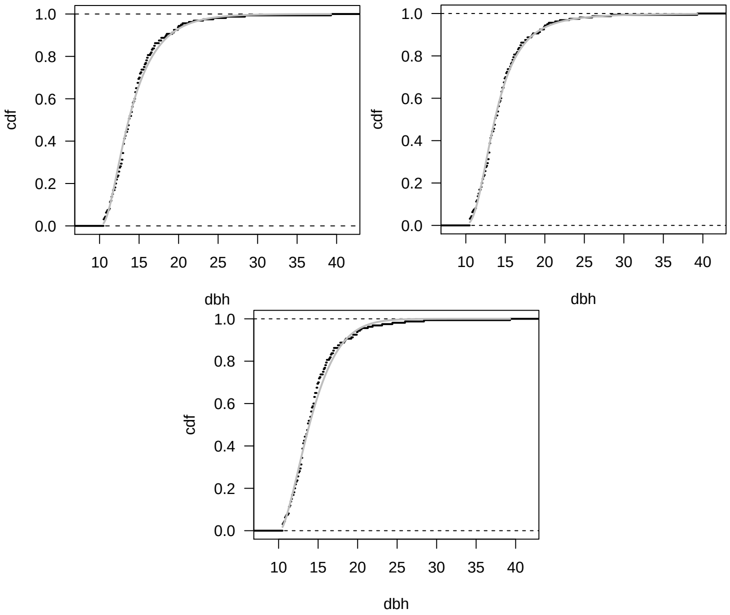


Figure 3: Empirical (bold) and theoretical (gray) c.d.f.'s for S5 using the ShBS, ShBS-t and ShWeibull distributions.

3.4. Financial Evaluation

We select S5 for carrying out a financial analysis. In this case, the ShBS-t distribution is considered as the best model. Then, we propose a forest production problem to illustrate the methodology presented in this article. Once the ShBS-t model parameters are estimated, we determine the mean volume per tree in a

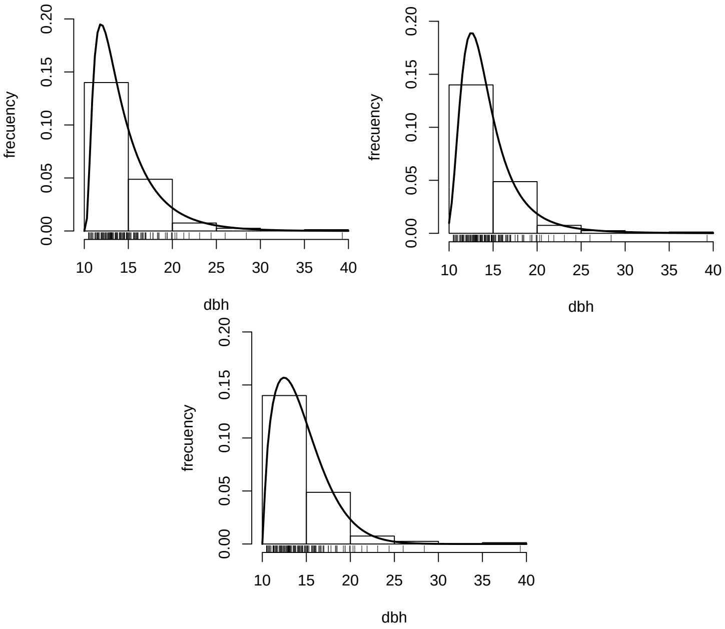


FIGURE 4: Histogram with ShBS, ShBS-t and ShWeibull p.d.f.'s for S5.

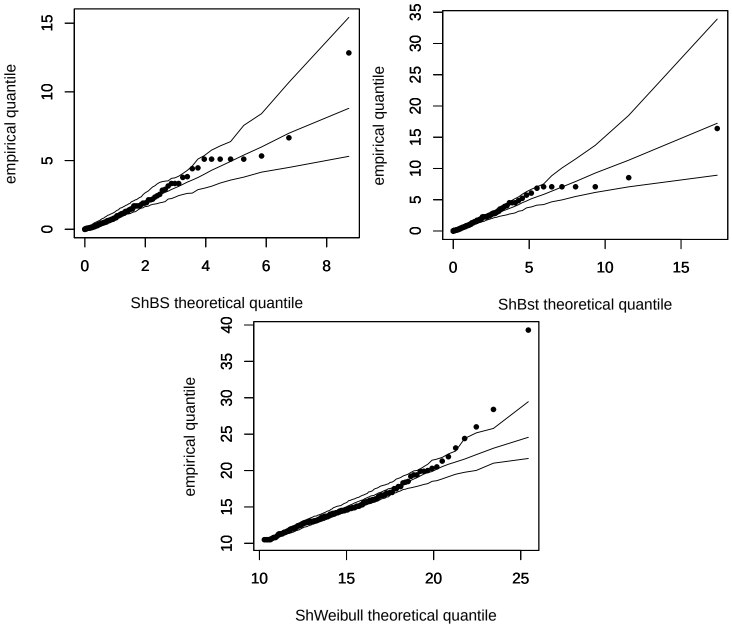


FIGURE 5: Probability plots with envelopes for S5 using the ShBS, ShBS-t and ShWeibull distributions.

stand by using (12) that leads to $\mathrm{E}[V] = (250/3)\,\pi\,\mathrm{E}[T^2]$, recalling that T is the DBH, H the known height of the tree equal to 10 m (1000 cm, because the data are expressed in cm) and F the form factor being it equal to 1/3 due to the birch case, which has conical shape, with the DBH equivalent to the diameter at the base of the cone. Using the expected value and variance of T given in (11), we get the expected volume as

$$\mathrm{E}[V] = \tfrac{250}{3}\,\pi\,\left[\beta^2\left\{1+\alpha^2(2A+\tfrac{5}{4}B\alpha^2+\tfrac{A^2\alpha^2}{4})\right\}+\gamma\left\{\gamma+\beta(2+A\alpha^2)\right\}\right]$$

The stand considered in this study only produces native wood that can be sold to sawmills at a price of US$250 (international price in US dollars) per cubic meter. This stand of Maine, US, had in the spring of 2004 an amount of 3327 trees, of which 160 (4.8%) were of the gray birch variety. Thus, the estimated expected economical value for gray birch wood of this forest (stand) based on the ShBS-t model is

$$\text{US\$0.25}\times\widehat{\mathrm{E}[V]}\times 160 = 10000\,\pi\,\big[\widehat{\beta}^2\left\{1+\widehat{\alpha}^2(2A+\frac{5}{4}B\widehat{\alpha}^2+\frac{A^2\widehat{\alpha}^2}{4})\right\}+\widehat{\gamma}\big\{\widehat{\gamma}+ \widehat{\beta}(2+A\widehat{\alpha}^2)\big\}\big] \quad (13)$$

being its estimation based on the proposed methodology and S5 of US$7,342,267.

4. Concluding Remarks

In this paper, we have presented, developed, discussed and applied a statistical methodology based on Birnbaum-Saunders distributions to address the problem of managing forest production. Specifically, we have linked a fatigue model to a forestry model through Birnbaum-Saunders distributions. This linkage has been possible because the hazard rate of this distribution has two clearly marked phases that coincide with the force of mortality of trees. This mortality is related to the diameter at breast height of trees. We have modeled the distribution of this diameter because this variable is the most relevant in determining the basal area of a tree. For its part, the basal area allows the volume of a tree to be determinated setting thus the production of a forest. Finally, we have shown the applicability of this model using five real data sets, obtaining for one of them financial information that may be valuable in forest decision making. The unpublished data used in the economical evaluation corresponded to the diameter at breast height of 10 m height mature gray birch trees collected in 2004, which are part of the inventory of a natural forest of area 16 hectares of different species located at Maine, US.

Acknowledgements

The authors wish to thank the Editor-in-Chief, Dr. Leonardo Trujillo, and two anonymous referees for their constructive comments on an earlier version of this manuscript which resulted in this improved version. Also, we wish to thank the colleagues Celia Benítez (Argentina), Rafal Podlaski (Poland) and Rómulo Santelices (Chile) for kindly providing us some of the data sets analyzed in this paper. This research work was partially supported by grant FONDECYT 1120879 from the Chilean government and by Universidad de Talca, Chile.

[Recibido: diciembre de 2011 — Aceptado: junio de 2012]

References

Aarset, M. V. (1987), 'How to identify a bathtub hazard rate', *IEEE Transaction on Reliability* **36**, 106–108.

Azevedo, C., Leiva, V., Athayde, E. & Balakrishnan, N. (2012), 'Shape and change point analyses of the Birnbaum-Saunders-t hazard rate and associated estimation', *Computational Statistics and Data Analysis* **56**, 3887–3897.

Bailey, R. & Dell, T. (1973), 'Quantifying diameter distributions with the Weibull function', *Forest Science* **19**, 97–104.

Birnbaum, Z. & Saunders, S. (1968), 'A probabilistic interpretation of miner's rule', *SIAM Journal of Applied Mathematics* **16**, 637–652.

Birnbaum, Z. & Saunders, S. (1969), 'A new family of life distributions', *Journal of Applied Probability* **6**, 319–327.

Bliss, C. & Reinker, K. (1964), 'A lognormal approach to diameter distributions in even-aged stands', *Forest Science* **10**, 350–360.

Borders, B., Souter, R., Bailey, R. & Ware, K. (1987), 'Percentile-based distributions characterize forest stand tables', *Forest Science* **33**, 570–576.

Clutter, J. & Bennett, F. (1965), Diameter distributions in old-field slash pine plantation, Report 13, US Forest Service.

Díaz-García, J. & Leiva, V. (2005), 'A new family of life distributions based on elliptically contoured distributions', *Journal of Statistical Planning and Inference* **128**, 445–457. (Erratum: *Journal of Statistical Planning and Inference*, **137**, 1512-1513).

Ferreira, M., Gomes, M. & Leiva, V. (2012), 'On an extreme value version of the Birnbaum-Saunders distribution', *Revstat-Statistical Journal* **10**, 181–210.

García-Güemes, C., Cañadas, N. & Montero, G. (2002), 'Modelización de la distribución diamétrica de las masas de *Pinus pinea* de Valladolid (España) mediante la función Weibull', *Investigación Agraria-Sistemas y Recursos Forestales* **11**, 263–282.

Gavrilov, L. & Gavrilova, N. (2001), 'The reliability theory of aging and longevity', *Journal of Theoretical Biology* **213**, 527–545.

Guiraud, P., Leiva, V. & Fierro, R. (2009), 'A non-central version of the Birnbaum-Saunders distribution for reliability analysis', *IEEE Transaction on Reliability* **58**, 152–160.

Hafley, W. & Schreuder, H. (1977), 'Statistical distributions for fitting diameter and height data in even-aged stands', *Canadian Journal of Forest Research* **7**, 481–487.

Johnson, N., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 2, Wiley, New York.

Leiva, V., Athayde, E., Azevedo, C. & Marchant, C. (2011), 'Modeling wind energy flux by a birnbaum-saunders distribution with unknown shift parameter', *Journal of Applied Statistics* **38**, 2819–2838.

Leiva, V., Barros, M., Paula, G. & Sanhueza, D. (2008), 'Generalized Birnbaum-Saunders distributions applied to air pollutant concentration', *Environmetrics* **19**, 235–249.

Leiva, V., Sanhueza, A. & Angulo, J. (2009), 'A length-biased version of the Birnbaum-Saunders distribution with application in water quality', *Stochastic Environmental Research and Risk Assessment* **23**, 299–307.

Leiva, V., Vilca, F., Balakrishnan, N. & Sanhueza, A. (2010), 'A skewed sinh-normal distribution and its properties and application to air pollution', *Communications in Statistics - Theory and Methods* **39**, 426–443.

Lenhart, J. & Clutter, J. (1971), Cubic foot yield tables for old-field loblolly pine plantations in the Georgia Piedmont, Report 22, US Forest Service.

Li, F., Zhang, L. & Davis, C. (2002), 'Modeling the joint distribution of tree diameters and heights by bivariate generalized Beta distribution', *Forest Science* **48**, 47–58.

Little, S. (1983), 'Weibull diameter distributions for mixed stands of western conifers', *Canadian Journal of Forest Research* **13**, 85–88.

Maltamo, M., Puumalinen, J. & Päivinen, R. (1995), 'Comparison of beta and Weibull functions for modelling basal area diameter in stands of *Pinus sylvestris* and *Picea abies*', *Scandinavian Journal of Forest Research* **10**, 284–295.

Marchant, C., Leiva, V., Cavieres, M. & Sanhueza, A. (2013), 'Air contaminant statistical distributions with application to PM10 in Santiago, Chile', *Reviews of Environmental Contamination and Toxicology* **223**, 1–31.

McEwen, R. & Parresol, B. (1991), 'Moment expressions and summary statistics for the complete and truncated Weibull dstribution', *Communications in Statistics - Theory and Methods* **20**, 1361–1372.

McGee, C. & Della-Bianca, L. (1967), Diameter distributions in natural yellow-poplar stands, Report 25, US Forest Service.

Meyer, H. (1952), 'Structure, growth, and drain in balanced uneven-aged forests', *Journal of Forestry* **50**, 85–92.

Miner, M. A. (1945), 'Cumulative damage in fatigue', *Journal of Applied Mechanics* **12**, 159–164.

Nelson, T. (1964), 'Diameter distribution and growth of loblolly pine', *Journal of Applied Mechanics* **10**, 105–115.

Palahí, M., Pukkala, T. & Trasobares, A. (2006), 'Modelling the diameter distribution of *Pinus sylvestris*, *Pinus nigra* and *Pinus halepensis* forest stands in Catalonia using the truncated Weibull function', *Forestry* **79**, 553–562.

Pece, M., de Benítez, C. & de Galíndez, M. (2000), 'Uso de la función Weibull para modelar distribuciones diamétricas en una plantación de *Melia azedarach*', *Revista Forestal Venezolana* **44**, 49–52.

Podlaski, R. (2006), 'Suitability of the selected statistical distribution for fitting diameter data in distinguished development stages and phases of near-natural mixed forests in the Swietokrzyski National Park (Poland)', *Forest Ecology and Management* **236**, 393–402.

Podlaski, R. (2008), 'Characterization of diameter distribution data in near-natural forests using the Birnbaum-Saunders distribution', *Canadian Journal of Forest Research* **18**, 518–527.

Rennolls, K., Geary, D. & Rollinson, T. (1985), 'Characterizing diameter distributions by the use of the Weibull distribution', *Forestry* **58**, 57–66.

Santelices, R. & Riquelme, M. (2007), 'Forest mensuration of *Nothofagus alessandri* of Coipué provenance', *Bosque* **28**, 281–287.

Schmelz, D. & Lindsey, A. (1965), 'Size-class structure of old-growth forests in Indiana', *Forest Science* **11**, 258–264.

Schreuder, H. & Hafley, W. (1977), 'A useful bivariate distribution for describing stand structure of tree heights and diameters', *Biometrics* **33**, 471–478.

Vilca, F. & Leiva, V. (2006), 'A new fatigue life model based on the family of skew-elliptical distributions', *Communications in Statistics - Theory and Methods* **35**, 229–244.

Vilca, F., Santana, L., Leiva, V. & Balakrishnan, N. (2011), 'Estimation of extreme percentiles in Birnbaum-Saunders distributions', *Computational Statistics and Data Analysis* **55**, 1665–1678.

Wang, M. & Rennolls, K. (2005), 'Tree diameter distribution modelling: introducing the logit-logistic distribution', *Canadian Journal of Forest Research* **35**, 1305–1313.

Zutter, B., Oderwald, R., Murphy, P. & Farrar, R. (1986), 'Characterizing diameter distributions with modified data types and forms of the Weibull distribution', *Forest Science* **32**, 37–48.

Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 371 a 384

A modified Cucconi Test for Location and Scale Change Alternatives

Un prueba de Cucconi modificada para alternativas de cambio en localización y escala

Marco Marozzi[a]

Dipartimento di Economia e Statistica, Università della Calabria, Italia

Abstract

The most common approach to develop a test for jointly detecting location and scale changes is to combine a test for location and a test for scale. For the same problem, the test of Cucconi should be considered because it is an alternative to the other tests as it is based on the squares of ranks and contrary-ranks. It has been previously shown that the Cucconi test is robust in level and is more powerful than the Lepage test, which is the most commonly used test for the location-scale problem. A modification of the Cucconi test is proposed. The idea is to modify this test consistently with the familiar approach which develops a location-scale test by combining a test for location and a test for scale. More precisely, we will combine the Cucconi test with the Wilcoxon rank test for location and a modified Levene test following the theory of the nonparametric combination. A power comparison of this modified Cucconi test with the original one, the Lepage test and the Podgor-Gastwirth $PG2$ test, shows that the modified Cucconi test is robust in size and markedly more powerful than the other tests for every considered type of distributions, from short- to normal- and long-tailed ones. A real data example is discussed.

***Key words*:** Combining tests, Location-scale model, Rank tests.

Resumen

La alternativa más común para implementar una prueba que detecta cambios en localización y escala conjuntamente es combinar una prueba de localización con una de escala. Para este problema, la prueba de Cucconi es considerada como una alternativa de otras pruebas que se basan en los cuadrados de los rangos y los contrarangos. Esta prueba es robusta en nivel y es más poderosa que la prueba de Lepage la cual es la más usada para el problema de localización-escala. En este artículo se propone una modificación de la prueba de Cucconi. La idea es modificar la prueba mediante

[a] Professor. E-mail: mmarozzi@unical.it

la combinación de una prueba de localización y uno de escala. Mas precisamente, se sugiere combinar la prueba de Cucconi con la prueba de rangos de Wilcoxon para localizacion y una prueba modificada de Levene siguiendo la teoría de la combinación no paramétrica. Una comparación de la potencia de esta prueba modificada de Cucconi con la prueba original, la prueba de Lepage y la prueba $PG2$ de Podgor-Gastwirth muestran que la prueba de Cucconi modificada es robusta en tamaño y mucho más poderosa que las anteriores para todas las distribuciones consideradas desde la normal hasta algunas de colas largas. Se hace una aplicación a datos reales.

***Palabras clave*:** combinación de pruebas, modelo de localización y escala, pruebas de rangos.

1. Introduction

The two sample Behrens-Fisher problem is to test that the locations, but not necessarily the scales, of the distribution functions associated to the populations behind the samples are equal. There exist situations of practical interest, however, when it is appropriate to jointly test for change in locations and change in scales. For example, Snedecor & Cochran (1989) emphasize that the application of a treatment (e.g. a drug) to otherwise homogeneous experimental units often results in the treated group differing not only in location but also in scales. The practitioner generally has no a prior knowledge about the distribution functions from which the data originate. Therefore, in such situations, an appropriate test does not require distributional assumptions. The test proposed by Perng & Littel (1976) for the equality of means and variances is not appropriate because is a combination of the t test and the F test, as the F test is not α robust for data from heavier than normal tailed distributions. According to Conover, Johnson & Johnson (1981) a test is α robust if its type one error rate is less than 2α. The cut off point is set to 1.5α by Marozzi (2011). As the Perng & Littel (1976) test which uses the Fisher combining function, the tests for the location-scale problem are generally expressed as functions of two tests, one sensitive to location changes and the other to scale changes. The corresponding statistics are generally obtained as direct combination of (i.e. by summing) a standardized statistic sensitive to location changes and a standardized statistic sensitive to scale changes. The most familiar test statistic for the location-scale problem, due to Lepage (1971), which is a direct combination of the squares of the standardized Wilcoxon and Ansari-Bradley statistics. It is important to note that Lepage-type tests can be obtained following Podgor & Gastwirth (1994). Marozzi (2009) compared several Podgor & Gastwirth (1994) efficiency robust tests and found that the $PG2$ test is the most powerful one. To perform the $PG2$ test it is necessary to regress the group indicator on the ranks and on the squares of the ranks of the data and to test that the two regression coefficients are zero. The $PG2$ test can be recast as a quadratic combination of the Wilcoxon test and the Mood squared rank test. For the same problem, the test of Cucconi (1968) should be considered because it is different from the other tests being not based on the combination of a test for location and a test for scale. It is a nonparametric test based on the squares of ranks and contrary-ranks. Marozzi (2009) computed for the very first time exact critical values for this test, compared its power to that of the Lepage and other tests that included several Podgor-Gastwirth tests and showed that the test of Cucconi maintains the size very close to the nominal level and is more powerful than the Lepage test. In this paper we are not interested in the general two sample problem, and therefore we do not consider tests like the Kolmogorov-Smirnov, Cramer-Von Mises or Anderson-Darling tests. In Section 2 we introduce a modification of the Cucconi test developed within the framework of the nonparametric combination of dependent tests (Pesarin 2001). A power comparison of this modified Cucconi test with the original one, the Lepage test and the Podgor-Gastwirth $PG2$ test is carried out in Section 3. These tests are applied to a real data set in Section 4. The conclusions are reported in Section 5.

2. The Modified Cucconi Test

In this section we introduce a modification of the Cucconi (Cucconi 1968) test. The idea is to modify this test consistently with the familiar approach which develops a location-scale test by combining a test for location and a test for scale. More precisely, following the theory of the nonparametric combination (Pesarin 2001) we will combine the Cucconi test with the Wilcoxon test for location and the modified Levene test for scale proposed by Brown & Forsythe (1974). We consider the Wilcoxon test and the modified Levene test because they have good properties in addressing the location and the scale problem respectively. Among other things, they are robust against non normality and they have good power, see Hollander & Wolfe (1999) and Marozzi (2011).

Let $\underline{\boldsymbol{X}}_1 = (X_{11}, \dots, X_{1n_1})$ and $\underline{\boldsymbol{X}}_2 = (X_{21}, \dots, X_{2n_2})$ be independent random samples of iid observations. Let F_1 and F_2 denote the absolutely continuous distribution functions associated to the populations underlying the samples. We wish to test

$$H_0 : F_1(g) = F_2(g) \text{ for all } g \in R \tag{1}$$

versus the location-scale alternative

$$H_1 : F_2(g) = F_1(\frac{g - \vartheta}{\tau}) \text{ with } \vartheta \in R, \tau > 0 \tag{2}$$

Note that for $\vartheta = 0$, H_1 reduces to a pure scale alternative and for $\tau = 1$ to a pure location alternative. Let μ_j and σ_j denote the location and scale of F_j, $j = 1, 2$. H_0 can be equivalently represented as

$$H_0 = H_{0l} \cap H_{0s} \text{ where } H_{0l} : \vartheta = \mu_1 - \mu_2 = 0 \text{ and } H_{0s} : \tau = \sigma_1/\sigma_2 = 1 \tag{3}$$

H_1 can be equivalently represented as

$$H_1 = H_{1l} \cup H_{1s} \text{ where } H_{1l} : \mu_1 - \mu_2 \neq 0 \text{ and } H_{1s} : \sigma_1/\sigma_2 \neq 1 \tag{4}$$

This representation of the system of hypotheses emphasizes that it is composed by two partial systems of hypotheses: the location and the scale one.

The test of Cucconi (1968) is based on

$$C = C\,(U, V) = \frac{U^2 + V^2 - 2\rho UV}{2(1-\rho^2)}$$

where

$$U = U(\underline{\boldsymbol{S}}_1) = \frac{6\sum\limits_{i=1}^{n_1} S_{1i}^2 - n_1(n+1)(2n+1)}{\sqrt{n_1 n_2 (n+1)(2n+1)(8n+11)/5}},$$

$$V = V(\underline{\boldsymbol{S}}_1) = \frac{6\sum\limits_{i=1}^{n_1} (n+1-S_{1i})^2 - n_1(n+1)(2n+1)}{\sqrt{n_1 n_2 (n+1)(2n+1)(8n+11)/5}}$$

$$n = n_1 + n_2, \quad \underline{\boldsymbol{S}}_1 = (S_{11}, \ldots, S_{1n_1})$$

S_{1i} denotes the rank of X_{1i} in the pooled sample

$$\underline{\boldsymbol{X}} = (\underline{\boldsymbol{X}}_1, \underline{\boldsymbol{X}}_2) = (X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}) = (X_1, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_n)$$

and $\rho = \frac{2(n^2-4)}{(2n+1)(8n+11)} - 1$. Note that U is based on the squares of the ranks S_{1i}, while V is based on the squares of the contrary-ranks $(n+1-S_{1i})$ of the first sample. Cucconi (1968) showed that under H_0 (U, V) has mean (0,0) because $E(\sum\limits_{i=1}^{n_1} S_{1i}^2) = n_1(n+1)(2n+1)/6$, and that $VAR(U) = VAR(V) = 1$ because $VAR(\sum\limits_{i=1}^{n_1} S_{1i}^2) = n_1 n_2 (n+1)(2n+1)(8n+11)/180$. Of course, it is $E(\sum\limits_{i=1}^{n_1}(n+1-S_{1i})^2) = E(\sum\limits_{i=1}^{n_1} S_{1i}^2)$ and $VAR(\sum\limits_{i=1}^{n_1}(n+1-S_{1i})^2) = VAR(\sum\limits_{i=1}^{n_1} S_{1i}^2)$. U and V are negatively correlated, more precisely, since $CORR(U,V) = COVAR(U,V) = \frac{2(n^2-4)}{(2n+1)(8n+11)} - 1 = \rho$ then $-1 \leq CORR(U,V) < -7/8$, where the minimum occurs when $n = 2$ and the supremum is reached when $n \to \infty$. It has been also shown that under H_0 if $n_1, n_2 \to \infty$ and $n_1/n \to \lambda \in]0,1[$ then $Pr(U \leq u) \to \Phi(u)$ and $Pr(U \leq v) \to \Phi(v)$, where Φ is the standard normal distribution function, moreover (U, V) converges in distribution to the bivariate normal with mean (0,0) and correlation $\rho_0 = -7/8$

$$\Pr(U \leq u, V \leq v) \to \int_{-\infty}^{u} \int_{-\infty}^{v} \frac{1}{2\pi\sqrt{1-\rho_0^2}} \exp\left(-\frac{q^2 + r^2 - 2\rho_0 qr}{2\,(1-\rho_0^2)}\right) dqdr$$

Therefore the points (u, v) outside the rejection region are close to (0,0), i.e. satisfy $\frac{1}{2\pi\sqrt{1-\rho_0^2}} \exp\left(-\frac{u^2+v^2-2\rho_0 uv}{2\left(1-\rho_0^2\right)}\right) \geq k$, where the constant k is chosen so that the type-one error rate is α. Let $k = \alpha\left(2\pi\sqrt{1-\rho_0^2}\right)^{-1}$, then it follows that if the point (u, v) is such that $\frac{u^2+v^2-2\rho_0 uv}{2\left(1-\rho_0^2\right)} < -\ln\alpha$ then we failed to have evidence against H_0. It is interesting to note that the rejection region E of the test is the set of points (u, v) outside the ellipse $u^2 + v^2 - 2\rho_0 uv = -2(1-\rho_0^2)\ln\alpha$. The test has size α because $\int\int_E \frac{1}{2\pi\sqrt{1-\rho_0^2}} \exp\left(-\frac{q^2+r^2-2\rho_0 qr}{2\left(1-\rho_0^2\right)}\right) dqdr = \alpha$. Note that in practice, unless you have large samples, ρ_0 should be replaced by ρ. Cucconi (1968) proved also that the test is unbiased and consistent for the location-scale problem.

We develop the modified Cucconi MC^* test following the nonparametric combination of dependent tests theory, which operates within the permutation framework, by combining the permutation version of the Cucconi test with the permutation version of the Wilcoxon W test for comparing locations and the Levene $W50$ test for comparing scales. The Wilcoxon W test is based on

$$W = \frac{\left|\sum_{i=1}^{n_2} S_{2i} - n_2\,(n+1)\,/2\right|}{n_1 n_2\,(n+1)\,/12}$$

The Levene $W50$ test is based on the Student t statistic computed on $R_{ji} = |X_{ji} - \widetilde{X}_j|$ where $\widetilde{X}_j$ is the median of the jth sample. Let us denote the mean of $R_{ji}, i = 1, \dots, n_j$ by $\overline{R}_j$, $j = 1, 2$, the Levene statistic is

$$W50 = \frac{\left|\overline{R}_1 - \overline{R}_2\right|}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{\sum_{i=1}^{n_1}\left(R_{1i}-\overline{R}_1\right)^2 + \sum_{i=1}^{n_2}\left(R_{2i}-\overline{R}_2\right)^2}{n-2}}}$$

Large values of W and $W50$ are evidence of difference in locations and scales respectively. It is desirable that the good performance in detecting separately location and scale changes shown by the W and the $W50$ tests are transferred to the combined test resulting in an improved power for jointly detecting location and scale changes with respect to the original Cucconi test. It has been shown that the nonparametric combination of dependent tests theory is very useful to address the location problem, see Marozzi (2004*b*), Marozzi (2004*a*) and Marozzi (2007), and the scale problem, see Marozzi (2011) and Marozzi (2012). We would like to see whether this theory is also useful to address the location-scale problem.

We describe now the permutation version C^* of the Cucconi test. Let $\underline{\boldsymbol{X}}^* = (\underline{\boldsymbol{X}}_1^*, \underline{\boldsymbol{X}}_2^*) = (X_{u_1^*}, \dots, X_{u_n^*}) = (X_1^*, \dots, X_n^*)$ denote a random permutation of the combined sample, where $(u_1^*, \dots, u_n^*)$ is a permutation of $(1, \dots, n)$, and so $\underline{\boldsymbol{X}}_1^* = (X_{u_1^*}, \dots, X_{u_{n_1}^*})$ and $\underline{\boldsymbol{X}}_2^* = (X_{u_{n_1+1}^*}, \dots, X_{u_n^*})$ are the two permuted samples. The permutation version of the C statistic is

$$C^* = C\left(\underline{\boldsymbol{X}}_1^*\right) = C\left(U^*, V^*\right) = \frac{\left(U^*\right)^2 + \left(V^*\right)^2 - 2\rho U^* V^*}{2(1-\rho)}$$

where $U^* = U\left(\underline{\boldsymbol{S}}_1^*\right)$, $V^* = V\left(\underline{\boldsymbol{S}}_1^*\right)$ and $\underline{\boldsymbol{S}}_1^*$ contains the ranks of $\underline{X}_1^*$ elements. The observed value of C^* is ${}_0C = C(U, V)$. To compute the p-value we compute the permutation null distribution of the C statistic as the distribution function of its permutation values: ${}_1C^*, \dots, {}_kC^*, \dots, {}_KC^*$ where ${}_kC^* = C\left({}_k\underline{\boldsymbol{X}}_1^*\right)$, ${}_k\underline{\boldsymbol{X}}_1^*$ contains the first n_1 elements of the kth permutation of $\underline{\boldsymbol{X}}$ and $k = 1, \dots, K = n!/(n_1! n_2!)$.

Therefore the p-value is

$$L_{C^*}\left({}_0C\right) = \frac{1}{K}\sum_{k=1}^{K} I\left({}_kC^* \geq_0 C\right)$$

where $I(.)$ denotes the indicator function.

We briefly describe now the permutation version of the W and $W50$ tests. Let

$$\begin{aligned}\underline{\boldsymbol{Y}} &= \left(\underline{\boldsymbol{X}}_1/SD\left(\underline{\boldsymbol{X}}_1\right), \underline{\boldsymbol{X}}_2/SD\left(\underline{\boldsymbol{X}}_2\right)\right)\\ &= \left(X_1/SD\left(\underline{\boldsymbol{X}}_1\right), \ldots, X_n/SD\left(\underline{\boldsymbol{X}}_2\right)\right)\\ &= \left(Y_1, \ldots, Y_n\right)\end{aligned}$$

be the standardized pooled sample, and let

$$\begin{aligned}\underline{\boldsymbol{Z}} &= \left(\underline{\boldsymbol{X}}_1 - E\left(\underline{\boldsymbol{X}}_1\right), \underline{\boldsymbol{X}}_2 - E\left(\underline{\boldsymbol{X}}_2\right)\right)\\ &= \left(X_1 - E\left(\underline{\boldsymbol{X}}_1\right), \ldots, X_n - E\left(\underline{\boldsymbol{X}}_2\right)\right)\\ &= \left(Z_1, \ldots, Z_n\right)\end{aligned}$$

be the mean aligned pooled sample. Let $\underline{\boldsymbol{Y}}^*$ and $\underline{\boldsymbol{Z}}^*$ be a random permutation of $\underline{\boldsymbol{Y}}$ and $\underline{\boldsymbol{Z}}$ respectively, it is important to emphasize that the $\underline{\boldsymbol{Y}}$ and $\underline{\boldsymbol{Z}}$ elements are not exactly exchangeable under H_0 and so the permutation solution is approximate; however it becomes asymptotically exact. $\underline{\boldsymbol{Z}}$ elements would be exchangeable if μ_1 and μ_2 were known and used in place of $E(\underline{\boldsymbol{X}}_1)$ and $E(\underline{\boldsymbol{X}}_2)$, see Pesarin & Salmaso (2010, pp. 73-74) and Good (2000, pp. 38-41). $\underline{\boldsymbol{Y}}$ elements would be exchangeable if σ_1 and σ_2 were known and used in place of $SD(\underline{\boldsymbol{X}}_1)$ and $SD(\underline{\boldsymbol{X}}_2)$, see Pesarin & Salmaso (2010, pp. 25 and 166-167). Alternatively, we considered also the median absolute deviation and the median in place of the standard deviation and the mean respectively in transforming $\underline{\boldsymbol{X}}$ and we obtained very similar results to those presented in section 3. It is also to be emphasized that, in order to preserve the within individual dependence on the transformed data $[\underline{\boldsymbol{X}}, \underline{\boldsymbol{Y}}, \underline{\boldsymbol{Z}}]$, the permutations must be carried on the n three-dimensional individual vectors $[(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)]$. So that $[\underline{\boldsymbol{X}}^*, \underline{\boldsymbol{Y}}^*, \underline{\boldsymbol{Z}}^*] = [(X_{u_i^*}, Y_{u_i^*}, Z_{u_i^*}), i = 1, \ldots, n]$.

In the permutation version W^* of the W test, the p-value is computed as $L_{W^*}({}_0W) = \frac{1}{K}\sum_{k=1}^{K} I\left({}_kW^* \geq_0 W\right)$, where ${}_0W$ is the observed value of the Wilcoxon statistic (that is computed on $\underline{\boldsymbol{Y}}$) and ${}_kW^*$ is the Wilcoxon statistic computed on the kth permutation ${}_k\underline{\boldsymbol{Y}}^*$ of $\underline{\boldsymbol{Y}}$. In the permutation version $W50^*$ of the $W50$ test, the p-value is computed as $L_{W50^*}({}_0W50) = \frac{1}{K}\sum_{k=1}^{K} I\left({}_kW50^* \geq_0 W50\right)$, where ${}_0W50$ is the observed value of the $W50$ statistic (that is computed on $\underline{\boldsymbol{Z}}$) and ${}_kW50^*$ is the $W50$ statistic computed on the kth permutation ${}_k\underline{\boldsymbol{Z}}^*$ of $\underline{\boldsymbol{Z}}$.

To obtain the MC^* test we combine the p-values of the C^*, W^* and $W50^*$ tests. This is equivalent to combine the test statistics being one to one decreasingly related to the p-values. Pesarin (2001, pp. 147-149) reports several combining functions, with the most familiar being

- the Fisher combining function $\ln\left(1/L_{C^*}\right) + \ln\left(1/L_{W^*}\right) + \ln\left(1/L_{W50^*}\right)$;

- the Tippett combining function $\max(1 - L_{C^*}, 1 - L_{W^*}, 1 - L_{W50^*})$;
- the Liptak combining function

$$\Phi^{-1}\left(1 - L_{C^*}\right) + \Phi^{-1}\left(1 - L_{W^*}\right) + \Phi^{-1}\left(1 - L_{W50^*}\right);$$

and noted that the Tippett combining function has a good power behavior when only one among the partial alternatives is true; that the Liptak combining function is generally good when the partial alternatives are jointly true; that the Fisher combining function has an intermediate behavior with respect to the Tippett and Liptak ones and therefore it is suggested when nothing is expected about the partial alternatives. Since we would like a combined test that is sensitive in all the three alternative situations: that are when H_{1l} alone is true, when H_{1s} alone is true, when H_{1l} and H_{1s} are jointly true, we use the Fisher combining function to obtain the test statistic for the null hypothesis $H_0 = H_{0l} \cap H_{0s}$

$$MC^* = \ln\left(1/L_{C^*}\right) + \ln\left(1/L_{W^*}\right) + \ln\left(1/L_{W50^*}\right)$$

Note that the Fisher combining function is used also by Perng & Littel (1976). The observed value of the MC^* statistic is ${}_0MC = \ln\left(1/L_{C^*}\left({}_0C\right)\right) + \ln\left(1/L_{W^*}\left({}_0W\right)\right) + \ln\left(1/L_{W50^*}\left({}_0W50\right)\right)$. The null distribution of the MC^* statistic is the distribution function of ${}_1MC^*, \ldots, {}_k MC^*, \ldots, {}_K MC^*$ where ${}_kMC^* = \ln\left(1/L_{C^*}\left({}_kC^*\right)\right) + \ln\left(1/L_{W^*}\left({}_kW^*\right)\right) + \ln\left(1/L_{W50^*}\left({}_kW50^*\right)\right)$. Large values of ${}_0MC$ are evidence against H_0, that should be rejected if $L_{MC^*}({}_0MC) \leq \alpha$ where $L_{MC^*}({}_0MC) = \frac{1}{K}\sum_{k=1}^{K} I\left({}_kMC^* \geq_0 MC\right)$. According to Pesarin (2001) it is possible to combine even a large, although finite, number of tests. In our case, we limit the number of tests to be combined to avoid the possibility that the type one error rate of the combined test may inflate too much, because under H_0 $\underline{\boldsymbol{Y}}$ and $\underline{\boldsymbol{Z}}$ elements are only approximately exchangeable.

3. Size and Power Study

We investigate via Monte Carlo simulation (5000 replications) the robustness of the significance level and the power of the modified Cucconi MC^* test in detecting location and scale changes, and we made comparisons with the classical Cucconi C test, the Lepage L test and the $PG2$ test. The Lepage test is based on

$$L = W^2 + \frac{\left(A - E(A)\right)^2}{VAR(A)}$$

where $A = \sum_{i=1}^{n_2} A_{2i}$ is the Ansari-Bradley statistic, A_{ji} denotes the Ansari-Bradley score of X_{ji} in the combined sample. To compute the A_{ji}s assign the score 1 to both the smallest and largest observations in the pooled sample, the score 2 to the second smallest and second largest, and so on. $E(A)$ and $VAR(A)$ denote the expected value and variance of A under H_0. Since the scoring depends on whether n is even or odd, two cases should be distinguished, $E(A) = n_2(n+2)/4$ and $VAR(A) = n_1 n_2\left(n+2\right)\left(n-2\right)/\left(48\left(n-1\right)\right)$ when n is even, and

$E(A) = n_2(n+1)^2/\left(4n\right)$ and $VAR(A) = n_1 n_2\left(n+1\right)\left(3+n^2\right)/\left(48n^2\right)$ when n is odd.

Let I_i $i = 1, \dots, n$ be a group indicator so that $I_i = 1$ when the ith element of the combined sample belongs to the first sample, $I_i = 0$ otherwise. The $PG2$ test statistic is the F statistic with 2 and $n-3$ df computed by regressing group indicators I_i on the ranks S_{ji} and the squared ranks S_{ji}^2 of the observations in the combined sample

$$PG2 = \frac{\left(\underline{\boldsymbol{b}}^T \underline{\boldsymbol{S}}^T \underline{\boldsymbol{I}} - n_1^2/n\right)/2}{\left(n_1 - \underline{\boldsymbol{b}}^T \underline{\boldsymbol{S}}^T \underline{\boldsymbol{I}}\right)/\left(n-3\right)}$$

where ^{T} denotes the transpose operator, $\underline{\boldsymbol{b}}$ is the 3×1 column vector of the OLS estimate of the intercept term and the regression coefficients, $\underline{\boldsymbol{S}}$ is a $n \times 3$ matrix with the first column of 1s, the second column of S_{ji} and the third column of S_{ji}^2, $i = 1, \dots, n_j$, $j = 1, 2$, $\underline{\boldsymbol{I}}$ is the nx1 column of the group indicators $I_1, \dots, I_n$.

The nominal 5% level is used throughout. We consider the following distributions that cover a wide range from short-tailed to very long-tailed distributions:

1. standard normal N(0,1);
2. uniform between $-\sqrt{3}$ and $\sqrt{3}$;
3. bimodal obtained as a mixture of a N(-1.5,1) with probability 0.5 and a N(1.5,1) with 0.5;
4. Laplace double exponential with scale parameter of $1/\sqrt{2}$;
5. 10% outlier obtained as a mixture of a N(0,1) with probability 0.9 and a N(1,10) with 0.1;
6. 30% outlier obtained as a mixture of a N(0,1) with probability 0.7 and a N(1,10) with 0.3;
7. Student's t with 2 df;
8. standard Cauchy, which corresponds to a Student's t with 1 df.

Note that distributions 7 and 8 have infinite second moment, and that distribution 8 has an undefined first moment. We consider only symmetric distributions because if one considers skewed distributions, a change in location is not qualitatively different with respect to a change in scale and therefore the location-scale alternative is not well specified in terms of $\mu_1 - \mu_2$ and σ_1/σ_2. We consider the balanced cases $(n_1, n_2) = (10, 10)$ and $(30, 30)$ as well as the unbalanced cases $(n_1, n_2) = (10, 30)$ and $(30, 10)$. We emphasize that p-values of the $PG2$ test have been computed exactly for all the sample size settings. p-values of the Lepage and Cucconi tests have been computed exactly for $(n_1, n_2) = (10, 10)$ and have been estimated by considering a random sample of 1 million permutations in the remaining settings. p-values of the MC^* test have been estimated by considering a random sample of 1000 permutations. The results in terms of the proportion of times H_0 is rejected are reported in Table 1 and Table 2 for the estimates of the size and power. The first two lines of the tables display the parameter choice: in the first column we are under H_0, while in the others we are under H_1. Note that all the tests are robust in size because their maximum estimated significance level (MESL) does not exceed 0.07. More precisely the MESL is 0.067, 0.058, 0.057 and 0.058 for the MC^*, L, C and $PG2$ tests respectively. It is important to note that the MESL of all the tests is greater than .05 and that the MESL of the MC^* test is the greatest one. Note that the cut-off point for the robustness in size is set to 0.1 by Conover et al. (1981) and more stringently to 0.075 by Marozzi (2011). Even if we caution that the results are obtained via simulations, they are very clear and show that the MC^* test is more powerful than the other tests for all distribution and sample size settings considered here. The results show that the combination of the Cucconi test with the Wilcoxon test for location and the modified Levene test for scale markedly improve the power of the Cucconi test in detecting separately location and scale changes, and in jointly detecting location and scale changes, for distributions that range from light-, to normal- and heavy-tailed distributions. The cost to be paid is the slightly liberality of the test that has a MESL of .067 (the other tests have a MESL between .057 and .058).

4. Application

Table 3 shows expenditure in Hong Kong dollars of 20 single men and 20 single women on the commodity group housing including fuel and light. This real data example is taken from Hand, Daly, Lunn, McConway & Ostrowski (1994, p. 44). Figure 1 presents the box plots of the data.

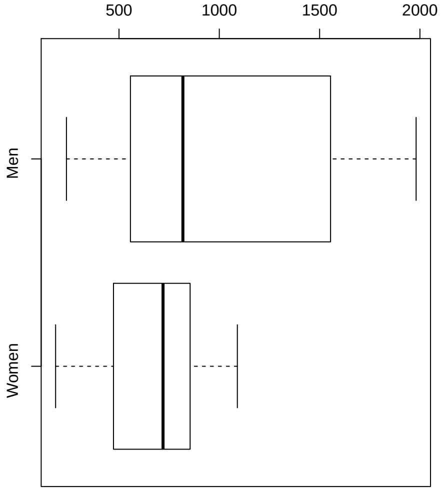


Figure 1: Box-plot of household expenditures.

We see from the box plots that the distributions of the data in the two groups seem to have different locations as well as different scales. This example illustrates

Table 1: Size and power of some tests for location and scale changes, $(n_1, n_2) = (10, 10)$ and $(10, 30)$.

	$(n_1, n_2) = (10, 10)$						$(n_1, n_2) = (10, 30)$				
	Normal						Normal				
$\mu_1 - \mu_2$	0	0	1	1	1	$\mu_1 - \mu_2$	0	0	0.75	0.75	0.75
σ_1/σ_2	1	2	2	1	3	σ_1/σ_2	1	1.5	1.5	1	2.5
MC^*	0.055	0.423	0.646	0.595	0.821	MC^*	0.057	0.349	0.628	0.544	0.896
L	0.050	0.249	0.383	0.415	0.585	L	0.044	0.201	0.427	0.383	0.690
C	0.052	0.281	0.414	0.410	0.639	C	0.048	0.257	0.473	0.388	0.780
$PG2$	0.053	0.286	0.418	0.413	0.642	$PG2$	0.046	0.253	0.467	0.381	0.775
	Uniform						Uniform				
$\mu_1 - \mu_2$	0	0	1	1	1	$\mu_1 - \mu_2$	0	0	0.75	0.75	0.75
σ_1/σ_2	1	2	2	1	3	σ_1/σ_2	1	1.5	1.5	1	2.5
MC^*	0.065	0.582	0.730	0.533	0.912	MC^*	0.063	0.519	0.688	0.518	0.966
L	0.053	0.381	0.435	0.348	0.683	L	0.051	0.324	0.430	0.340	0.778
C	0.053	0.456	0.489	0.327	0.764	C	0.050	0.430	0.503	0.343	0.882
$PG2$	0.054	0.462	0.494	0.331	0.767	$PG2$	0.049	0.424	0.497	0.339	0.879
	Bimodal						Bimodal				
$\mu_1 - \mu_2$	0	0	2.5	1.5	1.5	$\mu_1 - \mu_2$	0	0	2	1	1
σ_1/σ_2	1	1.5	1.5	1	2.5	σ_1/σ_2	1	1.5	1.5	1	1.5
MC^*	0.062	0.285	0.718	0.431	0.824	MC^*	0.061	0.489	0.801	0.356	0.634
L	0.048	0.174	0.453	0.261	0.587	L	0.051	0.305	0.555	0.222	0.379
C	0.047	0.203	0.441	0.251	0.652	C	0.053	0.396	0.611	0.222	0.459
$PG2$	0.048	0.206	0.446	0.253	0.657	$PG2$	0.050	0.389	0.605	0.216	0.453
	Laplace						Laplace				
$\mu_1 - \mu_2$	0	0	1	1	1	$\mu_1 - \mu_2$	0	0	0.75	0.75	0.75
σ_1/σ_2	1	2	2	1	3	σ_1/σ_2	1	1.5	1.5	1	2.5
MC^*	0.064	0.293	0.616	0.689	0.741	MC^*	0.054	0.243	0.681	0.690	0.844
L	0.057	0.164	0.435	0.543	0.539	L	0.058	0.144	0.537	0.563	0.682
C	0.055	0.175	0.449	0.547	0.572	C	0.053	0.177	0.554	0.560	0.739
$PG2$	0.056	0.176	0.452	0.549	0.576	$PG2$	0.051	0.174	0.548	0.554	0.735
	10% outlier						10% outlier				
$\mu_1 - \mu_2$	0	0	1.5	1	1	$\mu_1 - \mu_2$	0	0	1	0.75	0.75
σ_1/σ_2	1	2.2	2.2	1	3.5	σ_1/σ_2	1	2	2	1	2.2
MC^*	0.056	0.306	0.542	0.434	0.593	MC^*	0.063	0.355	0.606	0.443	0.557
L	0.055	0.225	0.408	0.303	0.493	L	0.054	0.331	0.513	0.289	0.482
C	0.050	0.235	0.423	0.310	0.501	C	0.053	0.370	0.549	0.294	0.526
$PG2$	0.051	0.238	0.427	0.312	0.505	$PG2$	0.051	0.365	0.541	0.288	0.521
	30% outlier						30% outlier				
$\mu_1 - \mu_2$	0	0	3.6	1.3	1.3	$\mu_1 - \mu_2$	0	0	1.8	1	1
σ_1/σ_2	1	3	3	1	6	σ_1/σ_2	1	2.2	2.2	1	3
MC^*	0.055	0.296	0.617	0.351	0.618	MC^*	0.057	0.306	0.608	0.350	0.569
L	0.047	0.238	0.491	0.260	0.520	L	0.052	0.242	0.503	0.239	0.464
C	0.046	0.224	0.502	0.270	0.488	C	0.052	0.259	0.506	0.243	0.480
$PG2$	0.047	0.227	0.504	0.271	0.494	$PG2$	0.049	0.255	0.500	0.238	0.475
	Student						Student				
$\mu_1 - \mu_2$	0	0	2	1	1	$\mu_1 - \mu_2$	0	0	1.1	0.8	0.8
σ_1/σ_2	1	2.4	2.4	1	3.6	σ_1/σ_2	1	1.8	1.8	1	2.2
MC^*	0.055	0.369	0.669	0.376	0.671	MC^*	0.058	0.310	0.608	0.410	0.605
L	0.046	0.242	0.506	0.252	0.490	L	0.049	0.234	0.474	0.264	0.464
C	0.047	0.255	0.521	0.262	0.509	C	0.050	0.272	0.500	0.274	0.515
$PG2$	0.048	0.258	0.525	0.263	0.514	$PG2$	0.047	0.267	0.495	0.267	0.508
	Cauchy						Cauchy				
$\mu_1 - \mu_2$	0	0	3	1.5	1.5	$\mu_1 - \mu_2$	0	0	1.5	1	1
σ_1/σ_2	1	3	3	1	5	σ_1/σ_2	1	2	2	1	3
MC^*	0.053	0.320	0.591	0.425	0.577	MC^*	0.062	0.218	0.536	0.395	0.521
L	0.046	0.255	0.490	0.318	0.495	L	0.051	0.195	0.457	0.249	0.483
C	0.048	0.250	0.494	0.321	0.488	C	0.051	0.217	0.466	0.244	0.503
$PG2$	0.049	0.255	0.498	0.324	0.493	$PG2$	0.050	0.214	0.460	0.239	0.496

TABLE 2: Size and power of some tests for location and scale changes, $(n_1, n_2) = (30, 10)$ and $(30, 30)$.

	$(n_1, n_2) = (30, 10)$						$(n_1, n_2) = (30, 30)$				
			Normal						Normal		
$\mu_1 - \mu_2$	0	0	1	0.75	0.75	$\mu_1 - \mu_2$	0	0	0.5	0.5	0.5
σ_1/σ_2	1	1.8	1.8	1	2.5	σ_1/σ_2	1	1.3	1.3	1	1.75
MC^*	0.059	0.416	0.726	0.512	0.854	MC^*	0.060	0.256	0.578	0.470	0.858
L	0.046	0.240	0.427	0.391	0.579	L	0.050	0.144	0.374	0.357	0.641
C	0.045	0.240	0.431	0.397	0.612	C	0.053	0.164	0.394	0.353	0.715
$PG2$	0.042	0.230	0.417	0.390	0.599	$PG2$	0.053	0.165	0.396	0.355	0.716
			Uniform						Uniform		
$\mu_1 - \mu_2$	0	0	1	0.75	0.75	$\mu_1 - \mu_2$	0	0	0.5	0.5	0.5
σ_1/σ_2	1	1.8	1.8	1	2.5	σ_1/σ_2	1	1.3	1.3	1	1.75
MC^*	0.059	0.600	0.808	0.472	0.954	MC^*	0.053	0.450	0.662	0.464	0.957
L	0.053	0.395	0.488	0.343	0.782	L	0.049	0.272	0.425	0.340	0.796
C	0.053	0.458	0.470	0.345	0.844	C	0.051	0.376	0.478	0.327	0.896
$PG2$	0.052	0.446	0.458	0.339	0.836	$PG2$	0.052	0.379	0.480	0.330	0.896
			Bimodal						Bimodal		
$\mu_1 - \mu_2$	0	0	2	1	1	$\mu_1 - \mu_2$	0	0	1.1	0.75	0.75
σ_1/σ_2	1	1.5	1.5	1	1.75	σ_1/σ_2	1	1.3	1.3	1	1.4
MC^*	0.067	0.328	0.740	0.318	0.671	MC^*	0.054	0.385	0.742	0.356	0.724
L	0.057	0.218	0.487	0.217	0.403	L	0.047	0.242	0.512	0.246	0.493
C	0.056	0.216	0.421	0.212	0.391	C	0.048	0.298	0.547	0.231	0.555
$PG2$	0.054	0.209	0.410	0.208	0.378	$PG2$	0.049	0.301	0.549	0.234	0.558
			Laplace						Laplace		
$\mu_1 - \mu_2$	0	0	1	0.75	0.75	$\mu_1 - \mu_2$	0	0	0.5	0.5	0.5
σ_1/σ_2	1	1.8	1.8	1	2.5	σ_1/σ_2	1	1.3	1.3	1	1.75
MC^*	0.055	0.285	0.727	0.648	0.734	MC^*	0.055	0.184	0.637	0.632	0.783
L	0.048	0.145	0.515	0.561	0.446	L	0.050	0.108	0.481	0.519	0.593
C	0.048	0.129	0.531	0.554	0.469	C	0.052	0.118	0.482	0.514	0.624
$PG2$	0.046	0.124	0.522	0.550	0.455	$PG2$	0.052	0.119	0.485	0.517	0.627
			10% outlier						10% outlier		
$\mu_1 - \mu_2$	0	0	1.5	0.75	0.75	$\mu_1 - \mu_2$	0	0	0.75	0.5	0.5
σ_1/σ_2	1	2	2	1	3	σ_1/σ_2	1	1.5	1.5	1	1.8
MC^*	0.066	0.329	0.646	0.365	0.632	MC^*	0.059	0.234	0.599	0.383	0.560
L	0.053	0.240	0.524	0.290	0.557	L	0.048	0.217	0.509	0.269	0.511
C	0.052	0.218	0.523	0.295	0.514	C	0.051	0.233	0.514	0.272	0.518
$PG2$	0.050	0.211	0.514	0.288	0.507	$PG2$	0.051	0.235	0.516	0.273	0.520
			30% outlier						30% outlier		
$\mu_1 - \mu_2$	0	0	3	1	1	$\mu_1 - \mu_2$	0	0	1.2	0.7	0.7
σ_1/σ_2	1	2.5	2.5	1	4.5	σ_1/σ_2	1	1.8	1.8	1	2.3
MC^*	0.054	0.283	0.664	0.324	0.617	MC^*	0.057	0.315	0.610	0.334	0.603
L	0.047	0.249	0.513	0.258	0.541	L	0.055	0.261	0.500	0.240	0.521
C	0.047	0.189	0.513	0.260	0.454	C	0.057	0.246	0.487	0.241	0.487
$PG2$	0.046	0.184	0.506	0.254	0.447	$PG2$	0.058	0.247	0.491	0.244	0.489
			Student						Student		
$\mu_1 - \mu_2$	0	0	1.7	0.8	0.8	$\mu_1 - \mu_2$	0	0	0.8	0.6	0.6
σ_1/σ_2	1	2.2	2.2	1	3	σ_1/σ_2	1	1.6	1.6	1	1.8
MC^*	0.060	0.414	0.712	0.344	0.703	MC^*	0.056	0.331	0.660	0.406	0.632
L	0.048	0.278	0.506	0.260	0.515	L	0.051	0.248	0.512	0.300	0.502
C	0.050	0.244	0.527	0.263	0.500	C	0.049	0.262	0.521	0.298	0.515
$PG2$	0.048	0.238	0.515	0.258	0.493	$PG2$	0.049	0.265	0.525	0.299	0.518
			Cauchy						Cauchy		
$\mu_1 - \mu_2$	0	0	2.5	1	1	$\mu_1 - \mu_2$	0	0	1.2	0.8	0.8
σ_1/σ_2	1	2.5	2.5	1	4	σ_1/σ_2	1	1.8	1.8	1	2.2
MC^*	0.063	0.322	0.588	0.297	0.567	MC^*	0.055	0.269	0.608	0.424	0.557
L	0.050	0.258	0.457	0.238	0.511	L	0.046	0.250	0.530	0.302	0.520
C	0.048	0.208	0.473	0.238	0.441	C	0.044	0.245	0.519	0.300	0.505
$PG2$	0.046	0.202	0.465	0.232	0.435	$PG2$	0.044	0.246	0.522	0.302	0.508

Table 3: Household expenditures (Honk Kong dollars) of a group of men and a group of women.

Men									
497	839	798	892	1585	755	388	617	248	1641
1180	619	253	661	1981	1746	1865	238	1199	1524
Women									
820	184	921	488	721	614	801	396	864	845
404	781	457	1029	1047	552	718	495	382	1090

that in practice we may have situations where F_1 and F_2 are different in both location and scale. With the aim at finding out whether household expenditures differ from men to women, we use the modified Cucconi test. By considering a random sample of 1 million permutations, the estimated p-value of the MC^* test is 0.0105, that suggests to reject the null hypothesis at level 5%. This result is consistent with the results obtained using the original Cucconi test and the $PG2$ test whose p-values are 0.0446 (estimated by considering a random sample of 1 million permutations) and 0.0441 (exact computation) respectively. The estimated p-value of the Lepage test is 0.0896 and suggests to reject H_0 at level 10%. At the basis of these results we conclude that household expenditures of men and women differ. It is worth noting that, with respect to the MC^* test, the other tests need a higher level in order to reject H_0. This might suggest a gain in power of the modified Cucconi test with respect to the original one and to the other tests.

5. Conclusion

We introduced a modification of the Cucconi test. The main objetive was to modify this test consistently with the familiar approach which develops a location-scale test by combining a test for location and a test for scale. More precisely we combined the Cucconi test with the Wilcoxon test for location and the modified Levene test for scale proposed by Brown & Forsythe (1974) following the theory of the nonparametric combination (Pesarin 2001). We compared the performance of the modified Cucconi test with the original one, the Lepage test and the Podgor-Gastwirth $PG2$ test in separately detecting location and scale changes as well as in jointly detecting location and scale changes. The results show that the combination of the Cucconi test with the Wilcoxon test for location and the modified Levene test for scale gives rise to a test which is slightly more liberal and markedly more powerful than the other tests for all the considered distributions, from short- to normal- and long-tailed ones. In the light of our findings, we recommend the practitioner to use the modified Cucconi test to address the location-scale problem, with caution on its type-one error rate.

[Recibido: diciembre de 2011 — Aceptado: agosto de 2012]

References

Brown, M. B. & Forsythe, A. B. (1974), 'Robust tests for the equality of variantes', *Journal of the American Statistical Association* **69**, 364–367.

Conover, W. J., Johnson, M. E. & Johnson, M. M. (1981), 'A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data', *Technometrics* **23**, 351–361.

Cucconi, O. (1968), 'Un nuovo test non parametrico per il confronto tra due gruppi campionari', *Giornale degli Economisti* **27**, 225–248.

Good, P. (2000), *Permutation Tests, a Practical Guide to Resampling Methods for Testing Hypotheses*, 2 edn, Springer-Verlag, New York.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. & Ostrowski, E. (1994), *A Handbook Of Small Data Sets*, Chapman and Hall, London.

Hollander, M. & Wolfe, D. A. (1999), *Nonparametric Statistical Methods*, 2 edn, Wiley, New York.

Lepage, Y. (1971), 'A combination of Wilcoxon and Ansari-Bradley statistics', *Biometrika* **58**, 213–217.

Marozzi, M. (2004*a*), 'A bi-aspect nonparametric test for the multi-sample location problem', *Computational Statistics and Data Analysis* **46**, 81–92.

Marozzi, M. (2004*b*), 'A bi-aspect nonparametric test for the two-sample location problem', *Computational Statistics and Data Analysis* **44**, 639–648.

Marozzi, M. (2007), 'Multivariate tri-aspect non-parametric testing', *Journal of Nonparametric Statistics* **19**, 269–282.

Marozzi, M. (2009), 'Some notes on the location-scale Cucconi test', *Journal of Nonparametric Statistics* **21**, 629–647.

Marozzi, M. (2011), 'Levene type tests for the ratio of two scales', *Journal of Statistical Computation and Simulation* **81**, 815–826.

Marozzi, M. (2012), 'Combined interquantile range tests for differences in scale', *Statistical Papers* **53**, 61–72.

Perng, S. K. & Littel, R. C. (1976), 'A test of equality of two normal population means and variances', *Journal of the American Statistical Association* **71**, 968–971.

Pesarin, F. (2001), *Multivariate Permutation Tests With Applications In Biostatistics*, Wiley, Chichester.

Pesarin, F. & Salmaso, S. (2010), *Permutation Tests for Complex Data*, Wiley, Chichester.

Podgor, M. J. & Gastwirth, J. L. (1994), 'On non-parametric and generalized tests for the two-sample problem with location and scale change alternatives', *Statistics in Medicine* **13**, 747–758.

Snedecor, G. W. & Cochran, W. G. (1989), *Statistical Methods*, 8 edn, Iowa State University Press, Ames.

Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 385 a 407

geofd: An R Package for Function-Valued Geostatistical Prediction

geofd: un paquete R para predicción geoestadística de datos funcionales

Ramón Giraldo[1,a], Jorge Mateu[2,b], Pedro Delicado[3,c]

[1]Department of Statistics, Sciences Faculty, Universidad Nacional de Colombia, Bogotá, Colombia

[2]Department of Mathematics, Universitat Jaume I, Castellón, Spain

[3]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain

Abstract

Spatially correlated curves are present in a wide range of applied disciplines. In this paper we describe the R package geofd which implements ordinary kriging prediction for this type of data. Initially the curves are pre-processed by fitting a Fourier or B-splines basis functions. After that the spatial dependence among curves is estimated by means of the trace-variogram function. Finally the parameters for performing prediction by ordinary kriging at unsampled locations are by estimated solving a linear system based estimated trace-variogram. We illustrate the software analyzing real and simulated data.

***Key words*:** Functional data, Smoothing, Spatial data, Variogram.

Resumen

Curvas espacialmente correlacionadas están presentes en un amplio rango de disciplinas aplicadas. En este trabajo se describe el paquete R geofd que implementa predicción por kriging ordinario para este tipo de datos. Inicialmente las curvas son suavizadas usando bases de funciones de Fourier o B-splines. Posteriormente la dependencia espacial entre las curvas es estimada por la función traza-variograma. Finalmente los parámetros del predictor kriging ordinario son estimados resolviendo un sistema de ecuaciones basado en la estimación de la función traza-variograma. Se ilustra el paquete analizando datos reales y simulados.

***Palabras clave*:** datos funcionales, datos espaciales, suavizado, variograma.

[a]Associate professor. E-mail: rgiraldoh@unal.edu.co
[b]Professor. E-mail: mateu@mat.uji.es
[c]Associate professor. E-mail: pedro.delicado@upc.edu

1. Introduction and Overview

The number of problems and the range of disciplines where the data are functions has recently increased. This data may be generated by a large number of measurements (over time, for instance), or by automatic recordings of a quantity of interest. Since the beginning of the nineties, functional data analysis (FDA) has been used to describe, analyze and model this kind of data. Functional versions for a wide range of statistical tools (ranging from exploratory and descriptive data analysis to linear models and multivariate techniques) have been recently developed (see an overview in Ramsay & Silverman 2005). Standard statistical techniques for FDA such as functional regression (Malfait & Ramsay 2003) or functional ANOVA (Cuevas, Febrero & Fraiman 2004) assume independence among functions. However, in several disciplines of the applied sciences there exists an increasing interest in modeling correlated functional data: This is the case when functions are observed over a discrete set of time points (temporally correlated functional data) or when these functions are observed in different sites of a region (spatially correlated functional data). For this reason, some statistical methods for modeling correlated variables, such as time series (Box & Jenkins 1976) or spatial data analysis (Cressie 1993), have been adapted to the functional context. For spatially correlated functional data, Yamanishi & Tanaka (2003) developed a regression model that enables to model the relationship among variables over time and space. Baladandayuthapani, Mallick, Hong, Lupton, Turner & Caroll (2008) showed an alternative for analyzing an experimental design with a spatially correlated functional response. For this type of modeling an associate software in MATLAB (MATLAB 2010) is available at http://odin.mdacc.tmc.edu/∼vbaladan. Staicu, Crainiceanu & Carroll (2010) propose principal component-based methods for the analysis of hierarchical functional data when the functions at the lowest level of the hierarchy are correlated. A software programme accompanying this methodology is available at http://www4.stat.ncsu.edu/∼staicu. Delicado, Giraldo, Comas & Mateu (2010) give a review of some recent contributions in the literature on spatial functional data. In the particular case of data with spatial continuity (geostatistical data) several kriging and cokriging predictors (Cressie 1993) have been proposed for performing spatial prediction of functional data. In these approaches a smoothing step, usually achieved by means of Fourier or B-splines basis functions, is initially carried out. Then a method to establish the spatial dependence between functions is proposed and finally a predictor for carrying out spatial prediction of a curve on a unvisited location is considered. Giraldo, Delicado & Mateu (2011) propose a classical ordinary kriging predictor, but considering curves instead of one-dimensional data; that is, each curve is weighted by a scalar parameter. They called this method “ordinary kriging for function-valued spatial data” (OKFD). This predictor was initially considered by Goulard & Voltz (1993). On the other hand; Giraldo, Delicado & Mateu (2010) solve the problem of spatial prediction of functional data by weighting each observed curve by a functional parameter. Spatial prediction of functional data based on cokriging methods are given in Giraldo (2009) and Nerini, Monestiez & Manté (2010). All of abovementioned approaches are important from a theoretical and applied perspective.

A comparison of these methods based on real data suggests that all of them are equally useful (Giraldo 2009). However, from a computational point of view the approach based on OKFD is the simplest because the parameters to estimate are scalars. In other cases the parameters are functions themselves and in addition it is necessary to estimate a linear model of coregionalization (Wackernagel 1995) for modeling the spatial dependence among curves, which could be restrictive when the number of basis functions used for smoothing the data set is large. For this reason the current version of the package geofd implemented within the statistical environment R (R Development Core Team 2011) only contains functions for doing spatial prediction of functional data by OKFD. However, the package will be progressively updated including new R functions.

It is important to clarify that the library geofd allows carrying out spatial prediction of functional data (we can predict a whole curve). This software cannot be used for doing spatio-temporal prediction. There is existing software that analyzes and models space-time data by considering a space-time covariance model and using to make this model predictions. There is no existing software for functional spatial prediction except the one we present in this paper. We believe there is no reason for confusion and the context gives us the necessary information to use existing space-time software or our software.

The package geofd has been designed mainly to support teaching material and to carry out data analysis and simulation studies for scientific publications. Working in geofd with large data sets can be a problem because R has limited memory to deal with such a large object. A solution can be use R packages for big data support such as bigmemory (http://www.bigmemory.org) or ff (http://ff.r-forge.r-project.org/).

This work is organized as follows: Section 2 gives a brief overview of spatial prediction by means of OKFD method, Section 3 describes the use of the package geofd based on the analysis of real and simulated data and conclusions are given in Section 4.

2. Ordinary Kriging for Functional Data

Ferraty & Vieu (2006) define a *functional variable* as a random variable X taking values in an infinite dimensional space (or functional space). *Functional data* is an observation x of X. A *functional data set* $x_1, \dots, x_n$ is the observation of n functional variables $X_1, \dots, X_n$ distributed as X. Let $T = [a, b] \subseteq \mathbb{R}$. We work with functional data that are elements of

$$L_2(T) = \{X : T \to \mathbb{R}, \text{ such that } \int_T X(t)^2 dt < \infty\}$$

Note that $L_2(T)$ with the inner product $\langle x, y\rangle = \int_T x(t)y(t)dt$ defines an Euclidean space.

Following Delicado et al. (2010) we define a functional random process as $\{X_s(t) : s \in D \subseteq \mathbb{R}^d, t \in T \subseteq \mathbb{R}\}$, usually $d = 2$, such that $X_s(t)$ is a functional variable for any $s \in D$. Let $s_1, \dots, s_n$ be arbitrary points in D and assume that we can observe a realization of the functional random process $X_s(t)$ at these n sites, $x_{s_1}(t), \dots, x_{s_n}(t)$. OKFD is a geostatistical technique for predicting $X_{s_0}(t)$, the functional random process at s_0, where s_0 is a unsampled location.

It is usually assumed that the functional random process is second-order stationary and isotropic, that is, the mean and variance functions are constant and the covariance depends only on the distance between sampling points (however, the methodology could also be developed without assuming these conditions). Formally, we assume that

1. $E(X_s(t)) = m(t)$ and $V(X_s(t)) = \sigma^2(t)$ for all $s \in D$ and all $t \in T$.
2. $COV(X_{s_i}(t), X_{s_j}(t)) = C(\|s_i - s_j\|)(t) = C_{ij}(h,t), s_i, s_j \in D, t \in T$, where $h = \|s_i - s_j\|$.
3. $\frac{1}{2}V(X_{s_i}(t) - X_{s_j}(t)) = \gamma(\|s_i - s_j\|)(t) = \gamma(h,t), s_i, s_j \in D, t \in T$, where $h = \|s_i - s_j\|$.

These assumptions imply that $V(X_{s_i}(t) - X_{s_j}(t)) = E(X_{s_i}(t) - X_{s_j}(t))^2$ and $\gamma\|s_i - s_i\|(t) = \sigma^2(t) - C(\|s_i - s_j\|)(t)$.

The OKFD predictor is defined as (Giraldo et al. 2011)

$$\widehat{X}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i X_{s_i}(t), \;\; \lambda_1, \dots, \lambda_n \in \mathbb{R} \tag{1}$$

The predictor (1) has the same expression as the classical ordinary kriging predictor (Cressie 1993), but considering curves instead of variables. The predicted curve is a linear combination of observed curves. Our approach considers the whole curve as a single entity, that is, we assume that each measured curve is a complete datum. The kriging coefficients or weights λ in Equation (1) give the influence of the curves surrounding the unsampled location where we want to perform our prediction. Curves from those locations closer to the prediction point will naturally have greater influence than others more far apart. These weights are estimated in such a way that the predictor (1) is the best linear unbiased predictor (BLUP). We assume that each observed function can be expressed in terms of K basis functions, $B_1(t), \dots, B_K(t)$, by

$$x_{s_i}(t) = \sum_{l=1}^{K} a_{il} B_l(t) = \boldsymbol{a}_i^T \boldsymbol{B}(t), \; i = 1, \dots, n \tag{2}$$

where $\boldsymbol{a}_i = (a_{i1}, \dots, a_{iK})$, $\boldsymbol{B}(t) = (B_1(t), \dots, B_K(t))$

In practice, these expressions are truncated versions of Fourier series (for periodic functions, as it is the case for Canadian temperatures) or B-splines expansions. Wavelets basis can also be considered (Giraldo 2009).

To find the BLUP, we consider first the unbiasedness. From the constant mean condition above, we require that $\sum_{i=1}^{n} \lambda_i = 1$. In a classical geostatistical setting we assume that the observations are realizations of a random field

$\{X_s : s \in D, D \in \mathbb{R}^d\}$. The kriging predictor is defined as $\widehat{X}_{s_0} = \sum_{i=1}^n \lambda_i X_{s_i}$, and the BLUP is obtained by minimizing

$$\sigma^2_{s_0} = V(\widehat{X}_{s_0} - X_{s_0})$$

subject to $\sum_{i=1}^n \lambda_i = 1$. On the other hand in multivariable geostatistics (Myers 1982, Ver Hoef & Cressie 1993, Wackernagel 1995) the data consist of $\{\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_n}\}$, that is, we have observations of a spatial vector-valued process $\{\mathbf{X}_s : s \in D\}$, where $\mathbf{X}_s = (X_s(1), \ldots, X_s(m))$ and $D \in \mathbb{R}^d$. In this context $V(\widehat{\boldsymbol{X}}_{s_0} - \boldsymbol{X}_{s_0})$ is a matrix, and the BLUP of m variables at an unsampled location s_0 can be obtained by minimizing

$$\sigma^2_{s_0} = \sum_{i=1}^m V\left(\widehat{X}_{s_0}(i) - X_{s_0}(i)\right)$$

subject to constraints that guarantee unbiasedness conditions, that is, minimizing the trace of the mean-squared prediction error matrix subject to some restrictions given by the unbiasedness condition (Myers 1982). Extending the criterion given in Myers (1982) to the functional context by replacing the summation by an integral, the n parameters in Equation (1) are obtained by solving the following constrained optimization problem (Giraldo et al. 2011)

$$\min_{\lambda_1,\ldots,\lambda_n} \int_T V(\widehat{X}_{s_0}(t) - X_{s_0}(t))dt, \text{ s.t.} \sum_{i=1}^n \lambda_i = 1 \quad (3)$$

which after some algebraic manipulation can be written as

$$\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \int_T C_{ij}(h,t)dt + \int_T \sigma^2(t)dt - 2\sum_{i=1}^n \int_T C_{i0}(h,t)dt + 2\mu(\sum_{i=1}^n \lambda_i - 1) \quad (4)$$

where μ is the Lagrange multiplier used to take into account the unbiasedness restriction. Minimizing (4) with respect to $\lambda_1, \ldots, \lambda_n$ and μ, we find the following linear system which enables to estimate the parameters

$$\begin{pmatrix} \int_T \gamma\|s_1 - s_1\|(t)dt & \cdots & \int_T \gamma\|s_1 - s_n\|(t)dt & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \int_T \gamma\|s_n - s_1\|(t)dt & \cdots & \int_T \gamma\|s_n - s_n\|(t)dt & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \int_T \gamma\|s_0 - s_1\|(t)dt \\ \vdots \\ \int_T \gamma\|s_0 - s_n\|(t)dt \\ 1 \end{pmatrix} \quad (5)$$

The function $\gamma(h) = \int_T \gamma\|s_i - s_j\|(t)dt$, is called the trace-variogram. In order to solve the system in (5), an estimator of the trace-variogram is needed. Given that we are assuming that $X_s(t)$ has a constant mean function $m(t)$ over D, $V(X_{s_i}(t) - X_{s_j}(t)) = E[(X_{s_i}(t) - X_{s_j}(t))^2]$. Note that, using Fubini's theorem

$$\gamma(h) = \frac{1}{2}E\left[\int_T (X_{s_i}(t) - X_{s_j}(t))^2 dt\right], \text{ for } s_i, s_j \in D \text{ with } h = \|s_i - s_j\| \quad (6)$$

Then an adaptation of the classical method-of-moments (MoM) for this expected value, gives the following estimator

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \int_T (X_{s_i}(t) - X_{s_j}(t))^2 dt \tag{7}$$

where $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$, and $|N(h)|$ is the number of distinct elements in $N(h)$. For irregularly spaced data there are generally not enough observations separated by exactly a distance h. Then $N(h)$ is modified to $\{(s_i, s_j) : \|s_i - s_j\| \in (h - \varepsilon, h + \varepsilon)\}$, with $\varepsilon > 0$ being a small value.

Once we have estimated the trace-variogram for a sequence of K values h_k, a parametric model $\gamma(h; \theta)$ such as spherical, Gaussian, exponential or Matérn (Ribeiro & Diggle 2001) must be fitted.

The prediction trace-variance of the functional ordinary kriging based on the trace-variogram is given by

$$\sigma^2_{s_0} = \int_T V(\widehat{X}_{s_0}(t) - X_{s_0}(t))dt = \sum_{i=1}^{n} \lambda_i \int_T \gamma \|s_i - s_0\|(t) dt - \mu \tag{8}$$

This parameter should be considered as a global uncertainty measure, in the sense that it is an integrated version of the classical pointwise prediction variance of ordinary kriging. For this reason its estimation cannot be used to obtain a confidence interval for the predicted curve. There is not, to the best of our knowledge, a method which allows us to do spatial prediction of functional data with an estimation of a prediction variance curve. We must take into account that we predict a whole curve and is not possible with this methodology to get point-wise confidence intervals, as we can obtain by using space or space-time models. It is clear that spatial-functional data and spatial temporal models have a common link in the sense that we have evolution of a spatial process through time or through any other characteristic. But at the same time there is an important difference. Spatial temporal models consider the evolution of a spatial process through time and models the interdependency of space and time. In this case we have $X(s, t)$ a single variable and we want to predict a variable at an unsampled location. In the spatial-functional case $X_s(t)$ is itself a function and thus we aim at predicting a function.

3. Illustration

Table 1 summarizes the functions of the package geofd. To illustrate its use we analyze real and simulated data. Initially in Sections 3.1 and 3.2 we apply the methodology to temperature measurements recorded at 35 weather stations located in the Canadian Maritime Provinces (Figure 1, left panel). Then the results with a simulated data set are shown in Section 3.3

The Maritime Provinces cover a region of Canada consisting of three provinces: Nova Scotia (NS), New Brunswick (NB), and Prince Edward Island (PEI). In particular, we analyze information of daily mean temperatures averaged over the

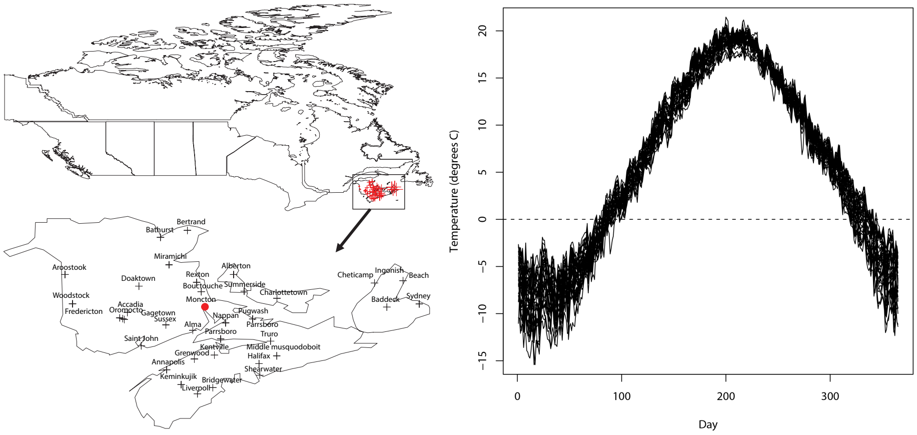


Figure 1: Averages (over 30 years) of daily temperature curves (right panel) observed at 35 weather stations of the Canadian Maritime provinces (left panel).

Table 1: Summary of the geofd functions.

Function	Description
`fit.tracevariog`	Fits a parametric model to the trace-variogram
`.geofd.viewer`	Graphical interface to plot multiple predictions
`l2.norm`	Calculates the L_2 norm between all pairs of curves
`maritimes.data`	Temperature values at 35 weather stations of Canada
`maritimes.avg`	Average temperature at Moncton station
`okfd`	Ordinary kriging for function-value data
`okfd.cv`	Cross-validation analysis for ordinary kriging for function-value data
`plot.geofd`	Plot the trace-variogram function and some adjusted models
`trace.variog`	Calculates the trace-variogram function

years 1960 to 1994 (February 29th combined with February 28th) (Figure 1, right panel). The data for each station were obtained from the Meteorological Service of Canada (http://www.climate.weatheroffice.ec.gc.ca/climateData/). Our package makes use of the R libraries fda (Ramsay, Hooker & Graves 2009) for smoothing data (by Fourier or B-splines basis) and geoR (Ribeiro & Diggle 2001) for fitting a variogram model to the estimated trace-variogram function. The temperature data set considered (Figure 1, right panel) is periodic and consequently a Fourier basis function is the most appropriate choice for smoothing it (Ramsay & Silverman 2005). However for illustrative purposes we also use a B-spline basis function. We can make a prediction at only one site or at multiple locations. Both alternatives are considered in the examples (Figure 2). In Section 3.1 we smooth the temperature data using a B-splines basis and, make a prediction at an unvisited location (left panel, Figure 2). In Section 3.2 we smooth the data using a Fourier basis and predict the temperature curves at ten randomly chosen sites (right panel, Figure 2).

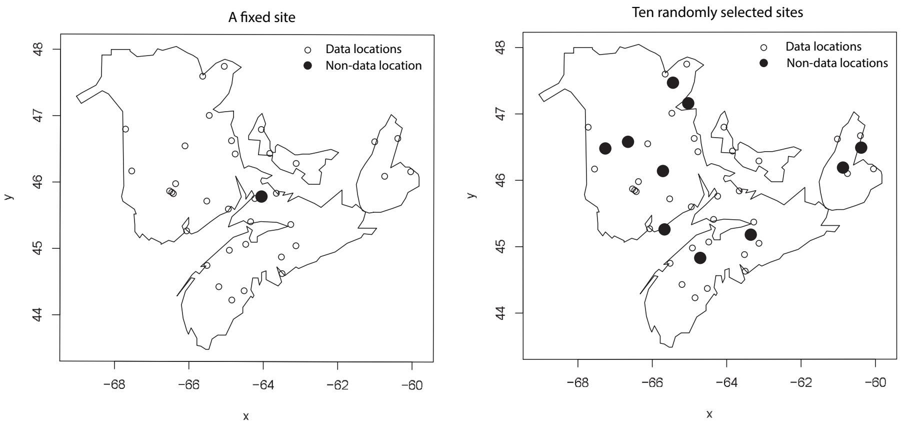


Figure 2: Prediction sites. A fixed site considered in the first example (left panel) and ten randomly selected sites considered in the second one (right panel).

3.1. Using a B-splines Basis

The following code illustrates how to use the library geofd for predicting a temperature curve at an unsampled location when the data are smoothed by using a B-splines basis. Initially we read and plot the data set (Figure 1, right panel), plot the coordinates of visited sites and choose a site for carrying out a prediction (Figure 2, left panel). The R code is the following.

```
R> library (geofd)
R> data(maritimes)
```

The `library(geofd)` command loads the package geofd (and other dependent packages) into the R computing environment. The `data(maritimes)` command loads the maritimes data set containing 35 temperature curves obtained at the same number of weather stations of the maritime provinces of Canada. The first five temperature values for four weather stations are

```
R> head(maritimes.data[,1:4], n=5)

     Fredericton       Halifax Sydney    Miramichi
[1,]             -7.9    -4.4       -3.8      -8.60
[2,]             -7.5    -4.2       -3.5      -8.32
[3,]             -9.3    -5.3       -4.6      -9.87
[4,]             -8.7    -5.4       -5.0      -9.55
[5,]             -9.1    -5.6       -4.1      -9.58
```

The next five lines of commands allow to plot the data and the coordinates.

```
R> matplot(maritimes.data,type="l",xlab="Day",ylab="degress C")
R> abline(h=0, lty=2)
```

```
R> plot(maritimes.coords)
R> coord.cero <- matrix(c(-64.06, 45.79),nrow=1,ncol=2)
R> points(coord.cero, col=2, lwd=3)
```

The main function of geofd is `okfd` (Table 1). This function allows to carry out predictions by ordinary kriging for function-valued data by considering a Fourier or a B-splines basis as methods for smoothing the observed data set. This covers from the smoothing step and trace-variogram estimation to data prediction. Although the estimation of the trace-variogram can be obtained by directly using the function `okfd`, it is also possible to estimate it in a sequential way by using the functions `l2.norm`, `trace.vari` and `fit.tracevariog`, respectively (Table 1). Now we give an illustration in this sense. In this example the data set is smoothed by using a B-splines basis with 65 functions without penalization (Figure 3, left panel). The number of basis functions was chosen by cross-validation (Delicado et al. 2010). We initially define the parameters for smoothing the data. We use here the fda library. An overview of the smoothing functional data by means of B-splines basis using the library fda library can be found in (Ramsay, Wickham, Graves & Hooker 2010). The following code illustrates how to run this process with the maritime data set.

```
R> n<-dim(maritimes.data)[1]
R> argvals<-seq(1,n, by=1)
R> s<-35
R> rangeval <- range(argvals)
R> norder   <- 4
R> nbasis   <- 65
R> bspl.basis <- create.bspline.basis(rangeval, nbasis, norder)
R> lambda <-0
R> datafdPar <- fdPar(bspl.basis, Lfdobj=2, lambda)
R> smfd <- smooth.basis(argvals,maritimes.data,datafdPar)
R> datafd <- smfd$fd
R> plot(datafd, lty=1, xlab="Day", ylab="Temperature (degrees C)")
```

The smoothed curves are shown in the left panel of Figure 3. Once wi have smoothed the data, we can use the functions above for estimating the trace-variogram. First we have to calculate the L_2 norm between the smoothed curves using the function `l2.norm`. The arguments for this function are the number `s` of sites where curves are observed, `datafd` a functional data object representing a smoothed data set and `M` a symmetric matrix of order equal to the number of basis functions defined by the B-splines basis object, where each element is the inner product of two basis functions after applying the derivative or linear differential operator defined by Lfdobj (Ramsay et al. 2010).

```
R> M <- bsplinepen(bspl.basis,Lfdobj=0)
R> L2norm <- l2.norm(s, datafd, M)
```

In the above commands the results are assigned to the variable `L2norm`. This one stores a matrix whose values correspond to the L_2 norm between each pair

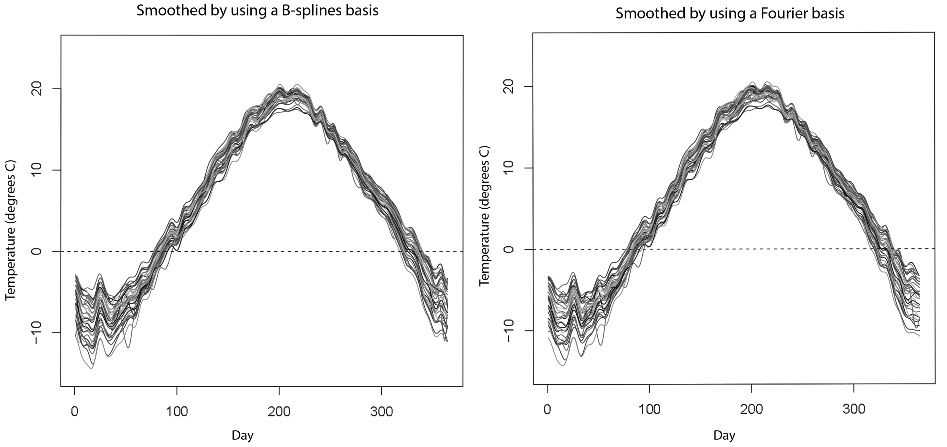


FIGURE 3: Smoothed data of daily temperature by using a B-splines basis (left panel) and a Fourier basis (right panel) with 65 functions.

of functional data into the data set. This matrix is then passed to the function `trace.variog` for estimating the trace-variogram function. The output can be returned as a trace-variogram “cloud" or as a binned trace-variogram (see Equation 7). The following code shows how this function can be used in combination with `fit.tracevariog` for fitting a model to the trace-variogram function obtained with the maritime data set. The main arguments of the function `trace.variog` are `coords` the geographical coordinates in decimal degrees where data where recorded, L_2 `norm` a matrix whose values are the L_2 norm between all pair of smoothed functions (an output from the function `l2.norm`), `bin` which is a logical argument indicating whether the output is a binned variogram, `maxdist` a numerical value defining the maximum distance for calculating the trace-variogram. Other arguments such as `uvec`, `breaks` and `nugget.tolerance` are defined as in the function `variog` of the package geoR. In order to fit a theoretical model (exponential, Gaussian, spherical or Matern) to the estimated trace-variogram we can use the function `fit.tracevariog`. This function makes use of the function `variofit` of geoR. The arguments of these functions are the estimations of the trace-variogram function (an output of the function `trace.variog`), `model` a list with the models that we want to fit, and some initial values for the parameters in these models. The command lines below show the use of these functions.

```
R> dista=max(dist(maritimes.coords))*0.9
R> tracev=trace.variog(maritimes.coords, L2norm, bin=FALSE,
+  max.dist=dista,uvec="default",breaks="default",nugget.tolerance)
R> models=fit.tracevariog(tracev, models=c("spherical","exponential",
+  "gaussian","matern"),sigma2.0=2000, phi.0=4, fix.nugget=FALSE,
+  nugget=0, fix.kappa=TRUE, kappa=1, max.dist.variogram=dista)
```

The variable `tracev` above stores the output of the function `trace.variog` which is used posteriorly in the function `plot.geofd` for plotting the trace-variogram

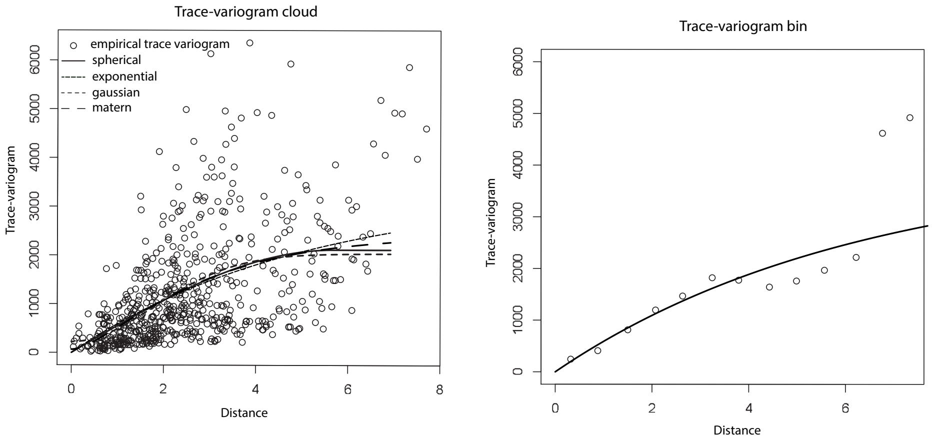


FIGURE 4: Estimated trace-variogram "cloud" and four fitted models (left panel). Estimated trace-variogram "bin" and the best fitted model (right panel).

"cloud". On the other hand the variable `model` stores the results obtained with the function `fit.tracevariog`. The use of the function `plot.geofd` in combination with the command `lines (models$fitted)` produces the plot shown in Figure 4 (left panel), this is, the estimated trace-variogram "cloud" and the four fitted models (exponential, Gaussian, spherical and Matern).

```
R> plot(tracev, xlab="Distancia", ylab="Trace-Variogram")
R> lines(models$fitted[[1]], lwd=2)
R> lines(models$fitted[[2]], lwd=2, col=4)
R> lines(models$fitted[[3]], lwd=2, col=7)
R> lines(models$fitted[[4]], lwd=2, col=6)
R> legend("topleft",   c("empirical trace variogram", "spherical",
+  "exponential", "gaussian",   "matern"), lty=c(-1,1,1,1,1),
+  col=c(1,1,4,7,6), pch=c(1,-1,-1,-1,-1))
```

In Figure 4 (right panel) the estimated trace-variogram "bin" and the best fitted model are shown. This plot is obtained by using the code below. In this case we use the option `bin=TRUE` in the function `trace.variog`, and the command line `lines(models$fitted[[2]], lwd=2, col=4)` to plot the exponential model.

```
R> tracevbin=trace.variog(maritimes.coords, L2norm, bin=TRUE,
+  max.dist=dista)
R> plot(tracevbin$u, tracevbin$v, ylim=c(0,3000), xlim=c(0, 7),
+  xlab="Distance", ylab="Trace-Variogram")
R> lines(models$fitted[[2]], lwd=2, col=4)
```

The numerical results of the function `fit.tracevariog` are stored in the object `models`. This list contains the estimations of the parameters (τ^2, σ^2, and ϕ) for each trace-variogram model and the minimized sum of squared errors (see `variofit` from geoR). According to the results below we can observe that the best model (least sum of squared errors) is the exponential model.

```
R>models

[[1]]
variofit: model parameters estimated by OLS (ordinary least squares):
covariance model is: spherical
fixed value for tausq =  0
parameter estimates:
  sigmasq       phi
3999.9950   12.0886
Practical Range with cor=0.05 for asymptotic range: 12.08865
variofit: minimised sum of squares = 529334304
[[2]]
variofit: model parameters estimated by OLS (ordinary least squares):
covariance model is: exponential
fixed value for tausq =  0
parameter estimates:
  sigmasq       phi
4000.0003    6.2689
Practical Range with cor=0.05 for asymptotic range: 18.77982
variofit: minimised sum of squares = 524840646
[[3]]
variofit: model parameters estimated by OLS (ordinary least squares):
covariance model is: gaussian
fixed value for tausq =  0
parameter estimates:
  sigmasq       phi
2092.8256    2.2886
Practical Range with cor=0.05 for asymptotic range: 3.961147
variofit: minimised sum of squares = 541151209
fitted[[4]]
variofit: model parameters estimated by OLS (ordinary least squares):
covariance model is: Matern with fixed kappa = 1
fixed value for tausq =  0
parameter estimates:
  sigmasq       phi
2693.1643    1.9739
Practical Range with cor=0.05 for asymptotic range: 7.892865
variofit: minimised sum of squares = 529431348
```

Once fitted, the best trace-variogram model we can use the `okfd` function for performing spatial prediction at an unvisited location. The arguments of this function are `new.coords` an $n \times 2$ matrix containing the coordinates of the new n prediction sites, `coords` an $s \times 2$ matrix containing the coordinates of the s sites where functional data were recorded, `data` an $m \times s$ matrix with values for the observed functions, `smooth.type` a string with the name of smoothing method to be used (B-splines or Fourier), `nbasis` a numeric value defining the number of basis functions used to smooth the discrete data set recorded at each site, `argvals` a vector containing argument values associated with the values to be smoothed, `lambda` (optional) a penalization parameter for smoothing the observed functions, and `cov.model` a string with the name of the correlation function (see `variofit` from geoR). Other additional arguments are `fix.nugget`, `nugget value`, `fix.kappa`, `kappa` (related to the parameters of the correlation model), and `max.dist.variogram` a numerical value defining the maximum distance considered when fitting the variogram model. The code below allows to predict a temperature curve at the Moncton weather station (see Figure 1).

```
R> okfd.res<-okfd(new.coords=coord.cero, coords=maritimes.coords,
+  cov.model="exponential",  data=maritimes.data, nbasis=65,
+  argvals=argvals, fix.nugget=TRUE)
R> plot(okfd.res$datafd, lty=1,col=8, xlab="Day",
+  ylab="Temperature (degrees C)",
+  main="Prediction at Moncton")
R> lines(okfd.res$argvals, okfd.res$krig.new.data, col=1, lwd=2,
+  type="l", lty=1, main="Predictions", xlab="Day",
+  ylab="Temperature (Degrees C)")
R> lines(maritimes.avg,  type="p", pch=20,cex=0.5, col=2, lwd=1)
```

A graphical comparison between real data (see `maritimes.avg` in Table 1) and the predicted curve (Figure 5) allows to conclude that the method OKFD has a good performance with this data set.

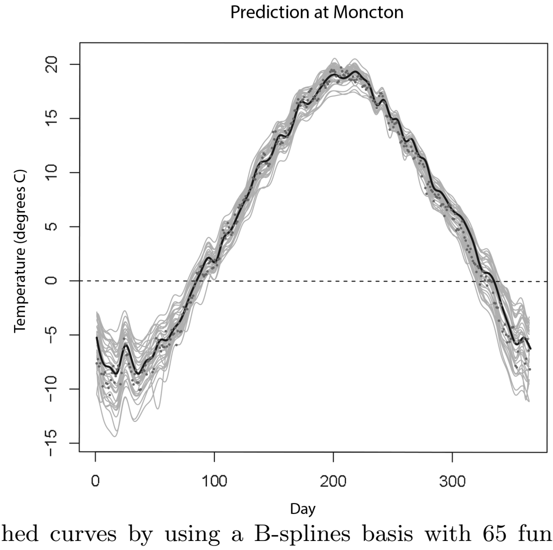


Figure 5: Smoothed curves by using a B-splines basis with 65 functions (*gray*), real data at Moncton weather station (red dots) and prediction at Moncton by ordinary kriging for function-value spatial data.

3.2. Using a Fourier basis

Now we use the package geofd for carrying out spatial prediction of temperature curves at ten randomly selected locations in the Canadian Maritimes Provinces (Figure 2, right panel). We use a Fourier basis with 65 functions for smoothing the data set (the same number of basis functions K as in Section 3.1) . In this example we show how the function `okfd` allows both smoothing the data and estimating directly a trace-variogram model. Posteriorly the estimation is used for performing spatial predictions of temperature curves on the ten locations already mentioned. The R code is the following

```
R> argvals<-seq(1,n, by=1)
R> col1<-sample((min(maritimes.coords[,1])*100):
(max(maritimes.coords[,1]) +  *100),10, replace=TRUE)/100
R> col2<-sample((min(maritimes.coords[,2])*100):
(max(maritimes.coords[,2]) +  *100),10, replace=TRUE)/100
R> new.coords <- cbind(col1,col2)
```

The variable `argvals` contains argument values associated with the values to be smoothed by using a Fourier basis. The variables `col1`, `col2`, and `new.coords` are used for defining the prediction locations (Figure 2, right panel). The variable `argvals` and `new.coords` are used as arguments of the function `okfd` in the code below

```
R> okfd.res<-okfd(new.coords=new.coords, coords=maritimes.coords,
+  data=maritimes.data, smooth.type="fourier", nbasis=65,
+  argvals=argvals, kappa=0.7)
```

In this example the arguments `smooth.type="fourier"` and `nbasis=65` in the function `okfd` allows us to smooth the data by using a Fourier basis with 65 functions (the number of basis functions was determined by cross-validation). In the example in Section 3.1 we use directly `cov.model="exponential"` in the function `okfd` because we chose this model previously by using the functions `trace.variog` and `fit.tracevariog`. If we do not specify a covariance model the function `okfd` estimates several models and selects the model with the least sum of squared errors. The parameter `kappa=.7` indicates that in addition to the spherical, exponential and Gaussian model, a Matern model with $\kappa = .7$ is also fitted.

A list with the objects stored in the variable `okfd.res` is obtained with the command line

```
R> names(okfd.res)
```

```
 [1] "coords"                         "data"
 [3] "argvals"                        "nbasis"
 [5] "lambda"                         "new.coords"
 [7] "emp.trace.vari"                 "trace.vari"
 [9] "new.Eu.d"                       "functional.kriging.weights"
```

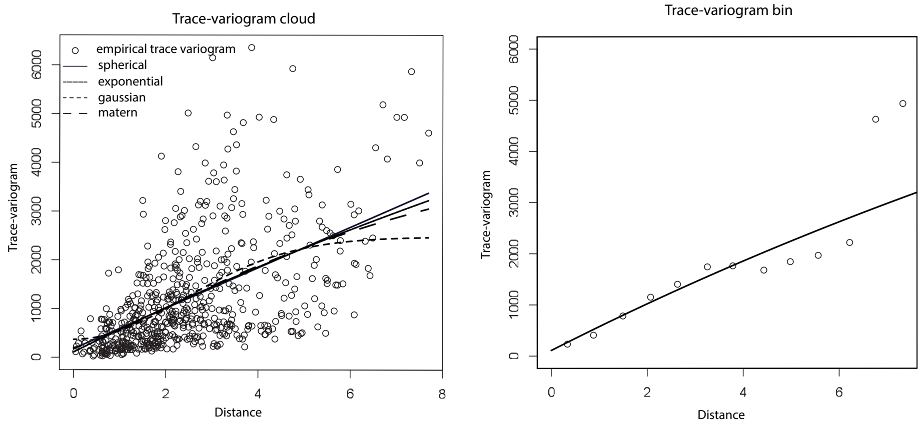


FIGURE 6: Estimated trace-variogram “cloud” and four fitted models (left panel). Estimated trace-variogram “bin” and the best fitted model (right panel).

```
[11] "krig.new.data"              "pred.var"
[13] "trace.vari.array"           "datafd"
```

We can use these objects for plotting the trace-variogram function, the estimated models and the predictions. A plot with the four fitted models and the best model is shown in Figure 6. We obtain this figure by using the command lines

```
R> plot(okfd.res, ylim=c(0,6000))
R> trace.variog.bin<-trace.variog(okfd.res$coords,
+  okfd.res$emp.trace.vari$L2norm, bin=TRUE)
R> plot(trace.variog.bin, ylim=c(0,6000), xlab="Distance",
+  ylab="Trace-variogram", main="Trace-variogram Bin")
R> lines(okfd.res$trace.vari, col=4, lwd=2)
```

Numerical results of the trace-variogram fitted models are obtained by using the command line

```
okfd.res$trace.vari.array

[[1]]
variofit: model parameters estimated by OLS (ordinary least squares):
covariance model is: spherical
parameter estimates:
      tausq     sigmasq         phi
   178.4011 644834.9056   2328.6674
Practical Range with cor=0.05 for asymptotic range: 2328.667
variofit: minimised sum of squares = 539799716
[[2]]
variofit: model parameters estimated by OLS (ordinary least squares):
```

```
covariance model is: exponential
parameter estimates:
     tausq    sigmasq        phi
  109.9118 11006.6152    23.1467
Practical Range with cor=0.05 for asymptotic range: 69.34139
variofit: minimised sum of squares = 539566326
[[3]]
variofit: model parameters estimated by OLS (ordinary least squares):
covariance model is: gaussian
parameter estimates:
    tausq   sigmasq       phi
 369.1311 2103.8708    3.3617
Practical Range with cor=0.05 for asymptotic range: 5.818404
variofit: minimised sum of squares = 552739397
[[4]]
variofit: model parameters estimated by OLS (ordinary least squares):
covariance model is: matern with fixed kappa = 0.7
parameter estimates:
    tausq   sigmasq       phi
 200.4886 4486.5365    5.8946
Practical Range with cor=0.05 for asymptotic range: 20.31806
variofit: minimised sum of squares = 541310787
```

The model with least sum of squared errors is again a exponential model (Figure 6, right panel). Consequently the function okfd above uses this model for solving the system in Equation 5 and for carrying out the predictions. Numerical values of predictions and prediction variances can be checked by using the commands

```
R> okfd.res[11]
R> okfd.res[12]
```

The predictions can be plotted by using the following command line

```
R>.geofd.viewer(okfd.res, argnames=c("Prediction","Day",
"Temperature"))
```

The function .geofd.viewer implements a Tcl/Tk interface (Grosjean 2010) for showing OKFD prediction results. This viewer presents two frames, the left one presents the spatial distribution of the prediction sites. The right one presents the selected prediction curve based on the point clicked by the user on the left frame. In Figure 7 we show the result of using this function. In the left panel a scatterplot with the coordinates of the prediction locations are shown. The dark point in the left panel is the clicked point and, the curve in the right panel shows the prediction at this site.

On the other hand if we want to plot all the predicted curves and analyze them simultaneously we can use the following command line

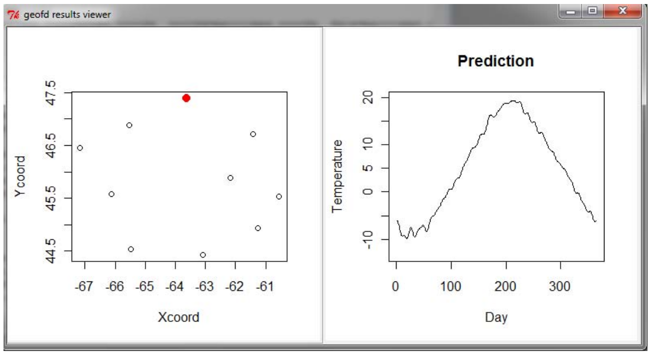


FIGURE 7: An example of the function `.geofd.viewer`. Left panel: Scatterplot with the coordinates of prediction locations. Right panel: Prediction on a clicked point (red point in left panel).

```
R> matplot(okfd.res$argvals, okfd.res$krig.new.data, col=1, lwd=1,
 type="l", +  lty=1, main="Predictions", xlab="Day",
 ylab="Temperature (degrees C)")
```

We can observe that the predicted curves (Figure 8) are consistent with the behavior of the original data set (Figure 1). This result indicates empirically that the OKFD method shows a good performance.

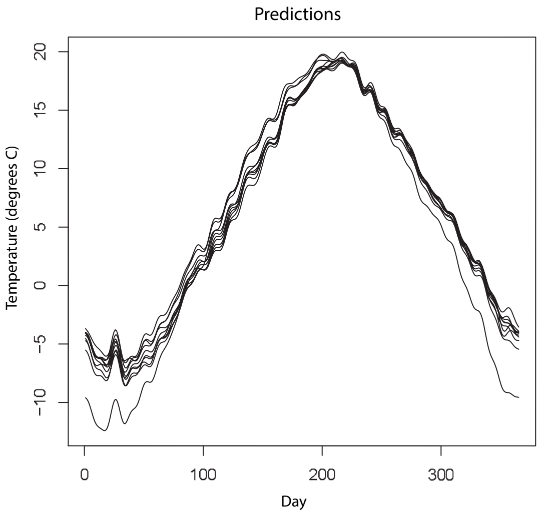


FIGURE 8: OKFD Predictions at ten randomly selected sites from Canadian Maritimes Provinces. Observed data were previously smoothed by using a Fourier basis.

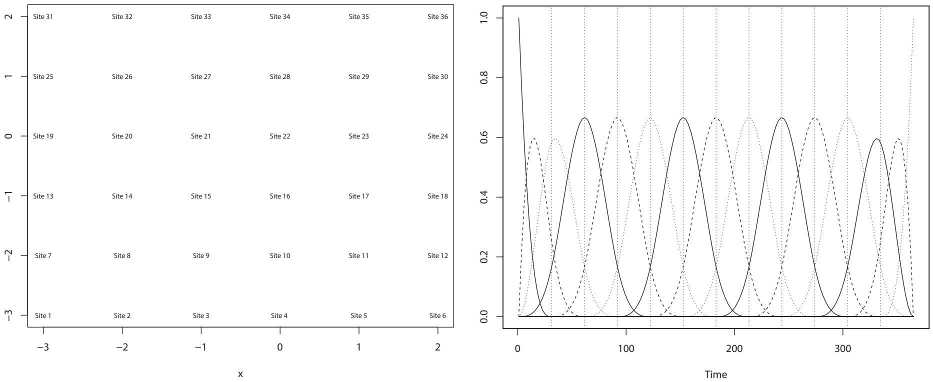


Figure 9: Left panel: Grid of simulated locations. Right panel: B-splines basis used in the simulation algorithm.

3.3. Using Simulated Data

In this section we discuss algorithms proposed in our package and evaluate the performance of the methodologies proposed in Section 2 by means of a simulation study.

We fixed the thirty six sites shown in Figure 9, and simulated a discretized set of spatially correlated functional data according to the model

$$X_{s_i}(t) = \sum_{l=1}^{15} a_{il} B_l(t) + \epsilon_i(t),\ i = 1, \dots, 36 \tag{9}$$

with $\mathbf{B}(t) = (B_1(t), \dots, B_{15}(t))$ a B-splines basis (see right panel Figure 9), a_{il}, a realization of a Gaussian random field $\mathbf{a_l} \sim N_{36}(10, \Sigma)$, where Σ is a 36×36 covariance matrix defined according to the exponential model $C(h) = 2\exp(\frac{-h}{8})$ with $h = \|s_i - s_j\|, i, j = 1, \dots, 36$, and $\epsilon(t) \sim N_{36}(0.09, 1)$ is a random error for each fixed t, with $t = 1, \dots, 365$. The number of basis functions and the parameters for simulating coefficients and errors were chosen empirically.

The R code for obtaining the simulated curves is the following

```
R> coordinates<-expand.grid(x= c(-3,-2, -1, 0, 1, 2),
+  y=c(-3,-2,-1,0, 1, 2))
R> mean.coef=rep(10,36)
R> covariance.coef <- cov.spatial(distance, cov.model=model,
+  cov.pars=c(2,8))
R> normal.coef=mvrnorm(15,mean.coef,covariance.coef)
R> mean.error<-rep(0, 36)
R> covariance.error <-cov.spatial(distance, cov.model=model,
+  cov.pars=c(0.09,0))
R> normal.error<-mvrnorm(365,mean.error,covariance.error)
R> argvals=seq(1, 365, len = 365)
```

```
R> nbasis=15
R> lambda=0
R> rangeval <- range(argvals)
R> norder   <- 4
R> bspl.basis <- create.bspline.basis(rangeval, nbasis,
+  norder)
R> data.basis=eval.basis(argvals, bspl.basis, Lfdobj=0)
R> func.data=t(normal.coef)%*%t(data.basis)
R> simulated.data= func.data+ normal.error
```

A plot with the simulated data and smoothed curves (by using a B-splines basis) is obtained with the following code

```
R> datafdPar <- fdPar(bspl.basis, Lfdobj=2, lambda)
R> smooth.datafd <- smooth.basis(argvals, simulated.data,
+  datafdPar)
R> simulated.smoothed=eval.fd(argvals, smooth.datafd$fd,
+  fdobj=0)
R> matplot(simulated.data, type="l", lty=1, xlab="Time",
+  ylab="Simulated data")
R> matplot(simulated.smoothed, lty=1, xlab="Time",
+  ylab="Smoothed data", type="l")
```

The simulated data are shown in the left panel of Figure 10. These data were smoothed by using a B-splines basis with 15 functions (right panel Figure 10). Once obtaining the smoothed curves we carry out a cross-validation prediction procedure. Each data location in Figure 9 is removed from the dataset and a smoothed curve is predicted at this location using OKFD based on the remaining smoothed functions.

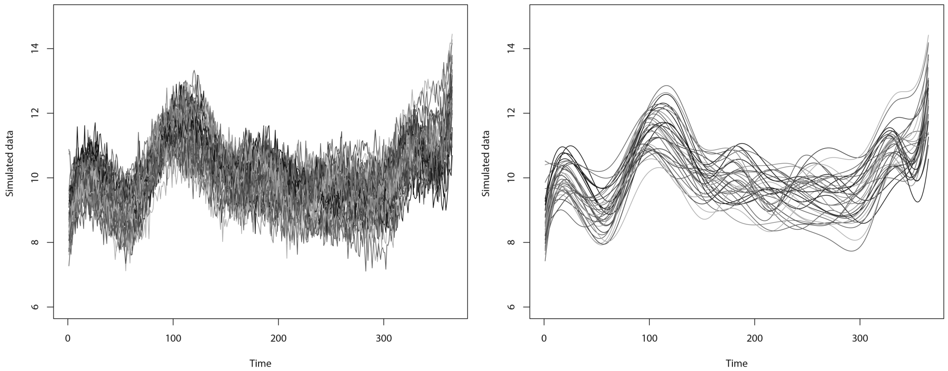


FIGURE 10: Left panel: Simulated data. Right panel: Smoothed curves (by using a B-splines basis).

The R code for obtaining the cross-validation predictions is

```
R> predictions= matrix(0, nrow=365, ncol=36)
```

```
R> for (i in 1:36)
R>     {
R>      coord.cero=matrix(coordinates[i,], nrow=1,ncol=2)
R>      okfd.res<-okfd(new.coords=coord.cero,
+       coords=coordinates[-i,], cov.model="exponential",
+       data=simulated.data[,-i], smooth.type="bsplines",
+       nbasis=15, argvals=argvals, fix.nugget=TRUE)
R>      predictions[,i]=okfd.res$krig.new.data
R>      }
```

We can plot the cross-validation predictions and the cross-validation residuals by using the following code

```
R> matplot(predictions, lty=1, xlab="Time",ylab="Predictions",
+  main="Cross-validation predictions", type="l")
R> cross.residuals=simulated.smoothed-predictions
R> matplot(cross.residuals, lty=1, xlab="Time",
+  ylab="Residuals", main="Cross-validation residuals",
+  type="l")
```

The cross-validation predictions (left panel Figure 11) shows that the predictions have the same temporal behavior as the smoothed curves (right panel Figure 10). Note also that the prediction curves have less variance. This is not surprising, because kriging is itself a smoothing method.

Figure 11 (right panel) shows cross-validation residuals. The predictions are plausible in all sites because all the residual curves are varying around zero.

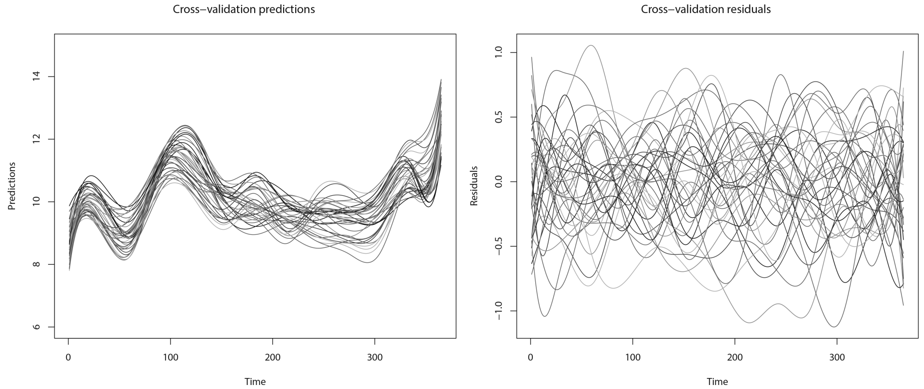


Figure 11: Left panel: Simulated data. Right panel: Smoothed curves (by using a B-splines basis).

The cross-validation results based on simulated data show a good performance of the proposed predictor, and indicate from a descriptive point of view that it can be adopted as a valid method for modeling spatially correlated functional data.

4. Conclusion

This paper introduces the R package geofd through an example. This package contains functions for modeling the trace-variogram function and for carrying out spatial prediction using the method of ordinary kriging for functional data. The advancements in this package would not be possible without several other important contributions to CRAN; these are reflected as geofd's package dependencies. The fda package by (Ramsay et al. 2010) provides methods for smoothing data by using basis functions. The geoR package (Ribeiro & Diggle 2001) provides functions to enable modeling the trace-variograma function. There remains scope for further extensions to geofd. We can consider other approaches for smoothing the data. For example, the use of wavelets could be useful for smoothing data with rapid changes in behavior. We plan to continue adding methods to the package. Continuous time varying kriging (Giraldo et al. 2010) and methods based on multivariable geostatistics (Giraldo 2009, Nerini et al. 2010) can be implemented in the package. However the use of these approaches could be restrictive when the number of basis functions used for smoothing the data set is large. Computationally efficient strategies are needed in this sense.

Acknowledgements

We would like to thank Andrés Pérez for his valuable contribution to upload the package geofd to CRAN. This work was partially supported by the Spanish Ministry of Education and Science through grants MTM2010-14961 and MTM2009-13985-C02-01.

[Recibido: octubre de 2011 — Aceptado: agosto de 2012]

References

Baladandayuthapani, V., Mallick, B., Hong, M., Lupton, J., Turner, N. & Caroll, R. (2008), 'Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinoginesis', *Biometrics* **64**, 64–73.

Box, G. & Jenkins, G. (1976), *Time Series Analysis.*, Holden Day, New York.

Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley & Sons, New York.

Cuevas, A., Febrero, M. & Fraiman, R. (2004), 'An ANOVA test for functional data.', *Computational Statistics and Data Analysis* **47**, 111–122.

Delicado, P., Giraldo, R., Comas, C. & Mateu, J. (2010), 'Statistics for spatial functional data: Some recent contributions', *Environmetrics* **21**, 224–239.

Ferraty, F. & Vieu, P. (2006), *Nonparametric Functional Data Analysis. Theory and Practice*, Springer, New York.

Giraldo, R. (2009), Geostatistical Analysis of Functional Data, PhD thesis, Universitat Politècnica de Catalunya.

Giraldo, R., Delicado, P. & Mateu, J. (2010), 'Continuous time-varying kriging for spatial prediction of functional data: An environmental application', *Journal of Agricultural, Biological, and Environmental Statistics* **15**(1), 66–82.

Giraldo, R., Delicado, P. & Mateu, J. (2011), 'Ordinary kriging for function-valued spatial data', *Environmental and Ecological Statistics* **18**(3), 411–426.

Goulard, M. & Voltz, M. (1993), Geostatistical interpolation of curves: A case study in soil science, *in* A. Soares, ed., 'Geostatistics Tróia 92', Vol. 2, Kluwer Academc Press, pp. 805–816.

Grosjean, P. (2010), *SciViews-R: A GUI API for R*, UMONS, Mons, Belgium.
*http://www.sciviews.org/SciViews-R

Malfait, N. & Ramsay, J. (2003), 'The historical functional linear model', *The Canadian Journal of Statistics* **31**(2), 115–128.

MATLAB (2010), *version 7.10.0 (R2010a)*, The MathWorks Inc., Natick, Massachusetts.

Myers, D. (1982), 'Matrix formulation of co-kriging', *Mathematical Geology* **14**(3), 249–257.

Nerini, D., Monestiez, P. & Manté, C. (2010), 'Cokriging for spatial functional data', *Journal of Multivariate Analysis* **101**(2), 409–418.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org.

Ramsay, J., Hooker, G. & Graves, S. (2009), *Functional Data Analysis with R and MATLAB*, Springer, New York.

Ramsay, J. & Silverman, B. (2005), *Functional Data Analysis. Second edition*, Springer, New York.

Ramsay, J., Wickham, H., Graves, S. & Hooker, G. (2010), *fda: Functional Data Analysis. R package version 2.2.6.*
*http://cran.r-project.org/web/packages/fda

Ribeiro, P. & Diggle, P. (2001), 'geoR: A package for geostatistical analysis', *R-NEWS* **1**(2), 15–18.
*http://cran.R-project.org/doc/Rnews

Staicu, A., Crainiceanu, C. & Carroll, R. (2010), 'Fast methods for spatially correlated multilevel functional data', *Biostatistics* **11**(2), 177–194.

Ver Hoef, J. & Cressie, N. (1993), 'Multivariable spatial prediction', *Mathematical Geology* **25**(2), 219–240.

Wackernagel, H. (1995), *Multivariate Geostatistics: An Introduction with Applications*, Springer-Verlag, Berlin.

Yamanishi, Y. & Tanaka, Y. (2003), 'Geographically weighted functional multiple regression analysis: A numerical investigation', *Journal of Japanese Society of Computational Statistics* **15**, 307–317.

Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 409 a 424

Goodness of Fit Tests for the Gumbel Distribution with Type II right Censored data

Pruebas de bondad de ajuste para la distribución Gumbel con datos censurados por la derecha tipo II

Víctor Salinas[a], Paulino Pérez[b], Elizabeth González[c], Humberto Vaquera[d]

Estadística, Socio Economía Estadística e Informática, Colegio de Postgraduados, Texcoco, México

Abstract

In this article goodness of fit tests for the Gumbel distribution with type II right censored data are proposed. One test is based in earlier works using the Kullback Leibler information modified for censored data. The other tests are based on the sample correlation coefficient and survival analysis concepts. The critical values of the tests were obtained by Monte Carlo simulation for different sample sizes and percentages of censored data. The powers of the proposed tests were compared under several alternatives. The simulation results show that the test based on the Kullback-Leibler information is superior in terms of power to the correlation tests.

***Key words*:** Correlation coefficient, Entropy, Monte Carlo simulation, Power of a test.

Resumen

En este artículo se proponen pruebas de bondad de ajuste para la distribución Gumbel para datos censurados por la derecha Tipo II. Una prueba se basa en trabajos previos en los que se modifica la información de Kullback-Leibler para datos censurados. Las otras pruebas se basan en el coeficiente de correlación muestral y en conceptos de análisis de supervivencia. Los valores críticos se obtuvieron mediante simulación Monte Carlo para diferentes tamaños de muestras y porcentajes de censura. La potencia de la pruebas se compararon bajo varias alternativas. Los resultados de la simulación muestran que la prueba basada en la Divergencia de Kullback-Leibler es superior a las pruebas de correlación en términos de potencia.

***Palabras clave*:** coeficiente de correlación, entropía, potencia de una prueba, simulación Monte Carlo.

[a] Student. E-mail: salinas.victor@colpos.mx
[b] Professor. E-mail: perpdgo@colpos.mx
[c] Professor. E-mail: egonzalez@colpos.mx
[d] Professor. E-mail: hvaquera@colpos.mx

1. Introduction

The Gumbel distribution is one of the most used models to carry out risk analysis in extreme events, in reliability tests, and in life expectancy experiments. This distribution is adequate to model natural phenomena, such as rainfall, floods, and ozone levels, among others. In the literature there exist some goodness of fit tests for this distribution, for example Stephens (1986), Lin, Huang & Balakrishnan (2008), Castro-Kuriss (2011). Several of these proposals modify well known tests, like the Kolmorogov-Smirnov and Anderson-Darling tests for type II censored data.

Ebrahimi, Habibullah & Soofi (1992), Song (2002), Lim & Park (2007), Pérez-Rodríguez, Vaquera-Huerta & Villaseñor-Alva (2009), among others, provide evidence that goodness of fit tests based on the Kullback-Leibler (1951) information show equal or greater power performance than tests based on the correlation coefficient or on the empirical distribution function. Motivated by this fact, in this article a goodness of fit test for the Gumbel distribution for type II right censored samples is proposed, using concepts from survival analysis and information theory.

This paper is organized as follows. Section 2 contains the proposed test based on Kullback-Leibler information as well as tables of critical values. In Section 3 we introduce two goodness of fit tests based on the correlation coefficient. Section 4 contains the results of a Monte Carlo simulation experiment performed in order to study the power and size of the tests against several alternative distributions. Section 5 presents two application examples with real datasets. Finally, some conclusions are given in Section 6.

2. Test Statistic Based on Kullback-Leibler Information

2.1. Derivation

Let X be a random variable with Gumbel distribution with location parameter $\xi \in \mathbb{R}$ and scale parameter $\theta > 0$, with probability density function (pdf) given by:

$$f_0(x;\xi,\theta) = \frac{1}{\theta} exp\left\{-\frac{x-\xi}{\theta} - exp\left\{-\frac{x-\xi}{\theta}\right\}\right\} I_{(-\infty,\infty)}(x) \tag{1}$$

Let $X_{(1)}, \dots, X_{(n)}$ be an ordered random sample of size n of an unknown distribution F, with density function $f(x) \in \mathbb{R}$ and finite mean. If only the first r (fixed) observations are available $X_{(1)}, \dots, X_{(r)}$ and the remaining $n-r$ are unobserved but are known to be greater than $X_{(r)}$ then we have type II right censoring. We are interested in testing the following hypotheses set:

$$H_0 : f(x;\cdot) = f_0(x;\xi,\theta) \tag{2}$$

$$H_1 : f(x;\cdot) \neq f_0(x;\xi,\theta) \tag{3}$$

That is, we wish to test if the sample comes from a Gumbel distribution with unknown parameters ξ and θ. To discriminate between H_0 and H_1, the Kullback-Leibler information for type II right censored data will be used, as proposed by Lim & Park (2007). To measure the distance between two known densities, $f(x)$ and $f_0(x)$, with $x < c$; the incomplete Kullback-Leibler information from Lim & Park (2007) can be considered, which is defined as:

$$KL(f, f_0 : c) = \int_{-\infty}^{c} f(x) \log \frac{f(x)}{f_0(x)} dx \tag{4}$$

In the case of complete samples, it is easy to see that $KL(f, f_0 : \infty) \geq 0$, and the equality holds if $f(x) = f_0(x)$ almost everywhere. However, the incomplete Kullback-Leibler information does not satisfy non-negativity any more. That is $KL(f, f_0 : c) = 0$ does not imply that $f(x)$ be equal to $f_0(x)$, for any x within the interval $(-\infty, c)$.

Lim & Park (2007) redefine the Kullback-Leibler information for the censored case as:

$$KL^{*}(f, f_0 : c) = \int_{-\infty}^{c} f(x) \log \frac{f(x)}{f_0(x)} dx + F_0(c) - F(c) \tag{5}$$

which has the following properties:

1. $KL^{*}(f, f_0 : c) \geq 0$.
2. $KL^{*}(f, f_0 : c) = 0$ if and only if $f(x) = f_0(x)$ almost everywhere for x in $(-\infty, c)$.
3. $KL^{*}(f, f_0 : c)$ is an increasing function of c.

In order to evaluate $KL^{*}(f, f_0 : c)$, f and f_0 must be determined. So it is necessary to propose estimators of these quantities based on the sample and considering the hypothesis of interest. From equation (5), using properties of logarithms we get:

$$KL^{*}(f, f_0 : c) = \int_{-\infty}^{c} f(x) \log f(x) dx - \underbrace{\int_{-\infty}^{c} f(x) \log f_0(x) dx}_{(\star)} + F_0(c) - F(c) \tag{6}$$

To estimate $f(x)$ for $x < c$, Lim & Park (2007) used the estimator proposed by Park & Park (2003), which is given by:

$$\hat{f}(x) = \begin{cases} 0 & \text{if } x < \nu_1 \\ n^{-1} \frac{2m}{x_{(i+m)} - x_{(i-m)}} & \text{if } \nu_i < x \leq \nu_{i+1},\ i = 1, \ldots,\ r \end{cases} \tag{7}$$

where ν_i=$(x_{(i-m)} + \cdots + x_{(i+m-1)})/(2m)$, $i = 1, \ldots, r$ and m is an unknown window size and a positive integer usually smaller than $n/2$. From (7) Lim &

Park (2007), built an estimator for $\int_{-\infty}^{c} f(x)\log f(x)dx = -H(f:c)$ in (6), which is given by:

$$H(m,n,r) = \frac{1}{n}\sum_{i=1}^{r}\log\left[\frac{n}{2m}\left(x_{(i+m)} - x_{(i-m)}\right)\right] \tag{8}$$

where $x_{(i)} = x_{(1)}$ for $i < 1$, $x_{(i)} = x_{(r)}$ for $i > r$.

To estimate ($\star$) in (6), Lim & Park (2007) proposed $\int_{-\infty}^{\nu_{r+1}} f(x)\log f_0(x)dx$, which can be written as:

$$\begin{aligned}\int_{-\infty}^{\nu_{r+1}} f(x)\log f_0(x)dx &= \int_{\nu_1}^{\nu_2} f(x)\log f_0(x)dx + \cdots + \int_{\nu_r}^{\nu_{r+1}} f(x)\log\ f_0(x)dx \\ &= \sum_{i=1}^{r}\underbrace{\int_{\nu_i}^{\nu_{i+1}} f(x)\log f_0(x)dx}_{(\star\star)}\end{aligned} \tag{9}$$

Substituting (1) and (7) in the i-th term of equation (9), we get:

$$\begin{aligned}(\star\star) &= \frac{2mn^{-1}}{x_{(i+m)} - x_{(i-m)}}\int_{\nu_i}^{\nu_{i+1}} \log\ f_0\left(x\right)dx \\ &= \frac{2mn^{-1}}{x_{(i+m)} - x_{(i-m)}}\int_{\nu_i}^{\nu_{i+1}}\left\{-\log\theta - \frac{x-\xi}{\theta} - \exp\left(-\frac{x-\xi}{\theta}\right)\right\}dx \\ &= \frac{2mn^{-1}}{x_{(i+m)} - x_{(i-m)}}\left[-\log\theta x - \frac{1}{\theta}\left(\frac{x^2}{2} - \xi x\right) + \theta\exp\left(-\frac{x-\xi}{\theta}\right)\right]\Bigg|_{\nu_i}^{\nu_{i+1}}\end{aligned} \tag{10}$$

The estimator of $F(c)$ in (6) can be obtained using (7), and it is given by r/n (Lim & Park 2007). Finally, the estimator of the incomplete Kullback-Leibler information for type II right censored data $KL^*(f, f_0 : c)$, denoted as $KL^*(m,n,r)$, is obtained by substituting (8), (9), (10) and the Gumbel distribution function in (6):

$$\begin{aligned}KL^*\left(m,n,r\right) = -H(m,n,r) &+ \exp\left\{-\exp\left(-\frac{\nu_{r+1}-\widehat{\xi}}{\widehat{\theta}}\right)\right\} \\ &- \frac{r}{n} - \sum_{i=1}^{r}\frac{2mn^{-1}}{x_{(i+m)} - x_{(i-m)}}\left[-\log\widehat{\theta}x - \frac{1}{\widehat{\theta}}\left(\frac{x^2}{2} - x\right)\right]\Bigg|_{\nu_i}^{\nu_{i+1}} \\ &- \sum_{i=1}^{r}\frac{2mn^{-1}}{x_{(i+m)} - x_{(i-m)}}\left[\widehat{\theta}\exp\left(-\frac{x-\widehat{\xi}}{\widehat{\theta}}\right)\right]\Bigg|_{\nu_i}^{\nu_{i+1}}\end{aligned} \tag{11}$$

where $\widehat{\xi}$ and $\widehat{\theta}$ are Maximum Likelihood Estimators (MLE) of ξ and θ, respectively. In the context of censored data, the estimators of $\Theta = (\xi,\theta)'$ are obtained by numerically maximizing the following likelihood function:

$$L(\Theta)=\prod_{i=1}^{n}\left\{f_0(x_i;\Theta)\right\}^{\delta_i}\left\{1-F_0(x_i;\Theta)\right\}^{1-\delta_i}$$

where $\delta_i=0$ if the i-th observation is censored and $\delta_i=1$ otherwise. We used the Nelder & Mead (1965) algorithm included in **optim** routine available in R (R Core Team 2012) to maximize this likelihood.

2.2. Decision Rule

Notice that under H_0 the values of the test statistic should be close to 0, therefore H_0 is rejected at the significance level α if and only if $KL^*(m,n,r)\geq K_{m,n,r}\left(\alpha\right)$, where the critical value $K_{m,n,r}\left(\alpha\right)$ is the $(1-\alpha)\times 100\%$ quantile of the distribution of $KL^*(m,n,r)$ under the null hypothesis, which fulfills the following condition:

$$\begin{aligned}\alpha&=P\left(\text{Reject } H_0 \mid H_0\right)\\&=P[KL^*(m,n,r)\geq K_{m,n,r}(\alpha)\mid H_0]\end{aligned}$$

2.3. Distribution of the Test Statistic and Critical Values

The distribution of the test statistic under the null hypothesis is hard to obtain analytically, since it depends on the unknown value of m and on non trivial transformations of certain random variables, and of course it also depends on the degree of censorship. Monte Carlo simulation was used to overcome these difficulties. The distribution of $KL^*(m,n,r)$ can be obtained using the following procedure.

1. Fix r, n, ξ, θ, m.
2. Generate a type II right censored sample of the Gumbel distribution, $(x_{(1)}, \ldots,x_{(n)}),(\delta_1,\ldots,\delta_n)$.
3. Obtain the maximum likelihood estimators of ξ and θ.
4. Calculate $KL^*(m,n,r)$ using (11).
5. Repeat steps 2, 3 and 4, B times, where B is the number of Monte Carlo samples hereafter.

Figure 1 shows the distribution of the test statistic $KL^*(m,n,r)$ for $m=3$, $n=50$, $r=45$, $B=10,000$, and for different values of parameters ξ and θ. This figure deserves at least two comments. First of all, the distribution has a big mass of probability close to 0 as expected under H_0. Second, the distribution of $KL^*(m,n,r)$ is location and scale invariant under H_0, that is, this distribution does not depend on ξ, neither on θ, so the critical values can be obtained by setting $\xi=0$ and $\theta=1$ or any other pair of possible values.

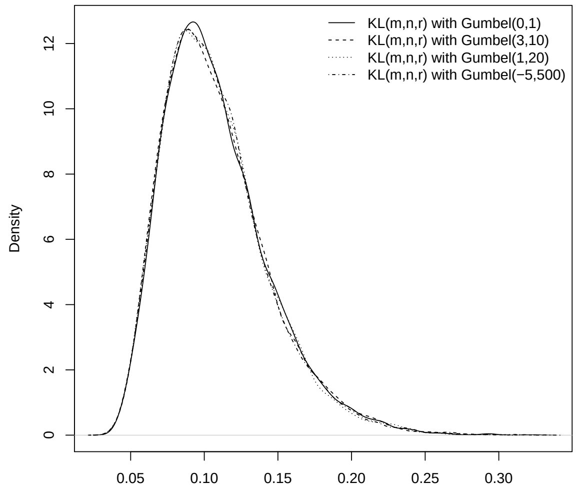


Figure 1: Estimated empirical distributions of $KL^*(m = 3, n = 50, r = 45)$ generated with $B = 10,000$ samples from the Gumbel distribution for the parameters specified in the legend.

The critical values $K_{m,n,r}\,(\alpha)$ were obtained by Monte Carlo Simulation. The used significance levels were $\alpha = 0.01$, 0.02, 0.05, 0.10 and 0.15. Random samples of the standard Gumbel distribution were generated for $n \leq 200$, $r/n = 0.5$, 0.6, 0.7, 0.8, 0.9, and $B = 10,000$. The value of $KL^*(m, n, r)$ was calculated for each $m < n/2$. For each m, n and r, the critical values were obtained with the $(1 - \alpha) \times 100\%$ quantiles of the empirical distribution function of $KL^*(m, n, r)$. For fixed values of n and r, the m value that minimizes $K_{m,n,r}\,(\alpha)$ was taken. Figure 2 plots the critical values $K_{m,n,r}(\alpha)$ for $n = 50$, $r = 40$ and $\alpha = 0.05$, corresponding to several values of m. The value of m that minimizes $K_{m,n,r}\,(\alpha)$ in this case is $m = 6$. More details about how to fix m and get the critical values can be found in Song (2002) and in Pérez-Rodríguez et al. (2009), among others.

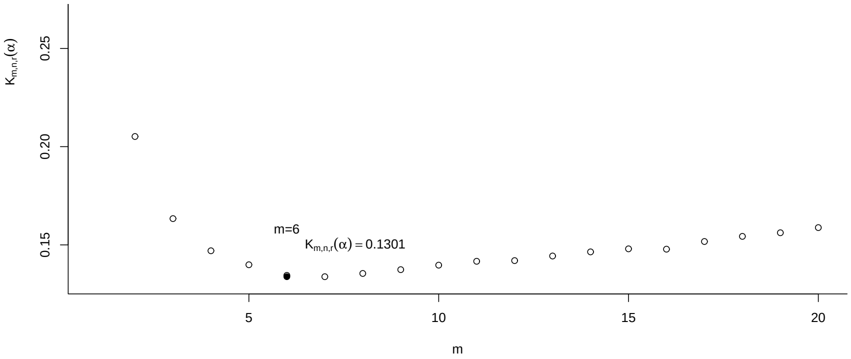


Figure 2: Critical values $K_{m,n,r}$ for $n = 50$, $r = 40$ and $\alpha = 0.05$.

Table 1 shows the critical values obtained by the simulation process described above. An R program (R Core Team 2012) to get the critical values for any sample size and percentage of censored observations is available upon request from the first author.

Table 1: Critical values $K_{m,n,r}\,(\alpha)$ of $KL^*(m,n,r)$ test obtained by Monte Carlo simulation.

α		0.01		0.02		0.05		0.10		0.15	
n	r	m	$K_{m,n,r}$	m	$K_{m,n,r}$	m	$K_{m,n,r}$	m	$K_{m,n,r}$	m	$K_{m,n,r}$
20	8	6	0.1497	5	0.1393	5	0.1242	5	0.1145	4	0.1079
	10	5	0.1662	6	0.1560	6	0.1361	5	0.1300	5	0.1237
	12	8	0.1741	8	0.1688	7	0.1560	5	0.1460	5	0.1403
	14	9	0.1912	8	0.1835	6	0.1724	6	0.1604	6	0.1566
	16	7	0.2119	10	0.2061	9	0.1919	7	0.1845	7	0.1747
	18	11	0.2470	11	0.2400	6	0.2225	5	0.2114	4	0.1990
30	15	10	0.1435	6	0.1379	6	0.1238	7	0.1122	6	0.1084
	18	8	0.1592	7	0.1477	8	0.1381	7	0.1290	7	0.1220
	21	10	0.1688	9	0.1609	8	0.1543	8	0.1424	5	0.1495
	24	11	0.1865	10	0.1779	9	0.1708	8	0.1591	4	0.1342
	27	14	0.2230	11	0.2075	6	0.1864	7	0.1732	4	0.1586
40	20	10	0.1302	10	0.1216	8	0.1105	8	0.1005	5	0.0981
	24	10	0.1405	10	0.1337	12	0.1248	9	0.1152	6	0.1092
	28	14	0.1540	11	0.1461	6	0.1385	8	0.1289	5	0.1155
	32	13	0.1704	12	0.1640	10	0.1540	4	0.1371	6	0.1247
	36	6	0.1989	7	0.1817	7	0.1604	7	0.1445	6	0.1318
50	25	11	0.1180	9	0.1107	10	0.1015	9	0.0954	7	0.0887
	30	11	0.1273	12	0.1201	8	0.1148	7	0.1040	6	0.0956
	35	12	0.1432	12	0.1342	6	0.1248	8	0.1103	5	0.1031
	40	7	0.1597	7	0.1464	6	0.1301	6	0.1166	6	0.1102
	45	6	0.1697	8	0.1559	7	0.1361	7	0.1274	6	0.1153
60	30	12	0.1084	12	0.1043	9	0.0949	5	0.0852	6	0.0784
	36	12	0.1201	13	0.1144	6	0.1040	7	0.0920	6	0.0861
	42	14	0.1342	11	0.1269	9	0.1116	7	0.0981	7	0.0900
	48	6	0.1422	9	0.1325	7	0.1177	7	0.1069	8	0.0987
	54	8	0.1499	8	0.1410	6	0.1248	7	0.1128	8	0.1043
70	35	12	0.1000	12	0.0953	5	0.0878	8	0.0787	5	0.0698
	42	11	0.1129	9	0.1052	10	0.0974	6	0.0847	6	0.0797
	49	8	0.1226	9	0.1116	6	0.1008	9	0.0922	7	0.0839
	56	8	0.1289	9	0.1181	7	0.1084	7	0.0968	8	0.0893
	63	7	0.1322	6	0.1282	9	0.1168	9	0.1027	7	0.0960
80	40	14	0.0949	7	0.0909	8	0.0827	6	0.0732	7	0.0670
	48	11	0.1080	8	0.0988	7	0.0891	8	0.0793	7	0.0736
	56	9	0.1051	9	0.1044	9	0.0940	8	0.0842	6	0.0779
	64	10	0.1218	9	0.1114	7	0.0991	8	0.0884	8	0.0813
	72	9	0.1251	6	0.1186	10	0.1028	9	0.0942	8	0.0873
90	45	10	0.0899	9	0.0854	7	0.0762	8	0.0680	9	0.0641
	54	10	0.0949	10	0.0930	8	0.0804	8	0.0748	8	0.0700
	63	10	0.1023	7	0.0954	9	0.0860	7	0.0785	10	0.0733
	72	9	0.1115	10	0.1028	8	0.0933	10	0.0831	8	0.0767
	81	9	0.1149	10	0.1085	8	0.0965	9	0.0866	8	0.0824
	50	7	0.0877	8	0.0801	7	0.0709	7	0.0650	8	0.0596
	60	7	0.0907	8	0.0849	9	0.0770	7	0.0691	9	0.0648

Table 1. (Continuation)

α		0.01		0.02		0.05		0.10		0.15	
n	r	m	$K_{m,n,r}$	m	$K_{m,n,r}$	m	$K_{m,n,r}$	m	$K_{m,n,r}$	m	$K_{m,n,r}$
100	70	8	0.0981	9	0.0901	7	0.0817	7	0.0725	7	0.0694
	80	8	0.0981	12	0.0948	8	0.0857	10	0.0773	9	0.0723
	90	9	0.1077	11	0.1000	9	0.0899	8	0.0834	9	0.0772
120	60	8	0.0754	11	0.0711	8	0.0644	8	0.0573	8	0.0536
	72	10	0.0791	9	0.0744	10	0.0696	8	0.0630	8	0.0586
	84	7	0.0860	10	0.0809	8	0.0745	9	0.0659	8	0.0628
	96	9	0.0869	10	0.0841	9	0.0771	10	0.0714	10	0.0659
	108	12	0.0943	10	0.0903	9	0.0810	10	0.0742	8	0.0682
140	70	8	0.0692	11	0.0652	9	0.0572	8	0.0534	9	0.0491
	84	10	0.0746	11	0.0695	8	0.0632	8	0.0574	8	0.0524
	98	10	0.0767	10	0.0737	8	0.0660	9	0.0599	9	0.0566
	112	10	0.0812	14	0.0791	11	0.0701	11	0.0636	11	0.0602
	126	12	0.0871	11	0.0807	10	0.0734	10	0.0659	9	0.0643
160	80	11	0.0638	11	0.0606	12	0.0542	10	0.0481	9	0.0452
	96	8	0.0675	11	0.0622	10	0.0577	9	0.0530	9	0.0488
	112	13	0.0731	9	0.0674	11	0.0601	10	0.0564	10	0.0529
	128	11	0.0750	11	0.0704	10	0.0635	9	0.0594	12	0.0561
	144	12	0.0762	11	0.0755	11	0.0675	11	0.0615	11	0.0575
180	90	8	0.0592	11	0.0561	9	0.0504	8	0.0448	9	0.0432
	108	14	0.0628	10	0.0578	10	0.0536	12	0.0483	10	0.0454
	126	10	0.0652	13	0.0623	9	0.0565	9	0.0530	12	0.0486
	144	12	0.0709	14	0.0671	12	0.0600	11	0.0550	10	0.0523
	162	12	0.0723	13	0.0688	11	0.0628	10	0.0576	12	0.0555
200	100	12	0.0548	10	0.0522	12	0.0466	9	0.0423	10	0.0403
	120	14	0.0582	12	0.0564	12	0.0506	11	0.0459	9	0.0436
	140	13	0.0620	10	0.0591	12	0.0531	11	0.0488	13	0.0462
	160	10	0.0665	13	0.0623	12	0.0564	12	0.0523	13	0.0491
	180	13	0.0680	13	0.0631	14	0.0594	11	0.0541	13	0.0516

3. Correlation Tests

In this section we derive two tests based on the correlation coefficient for the Gumbel distribution for type II right censored data. The proposed tests will allow us to test the set of hypotheses given in (2) and (3) with unknown parameters ξ and θ. The first test is based on Kaplan & Meier (1958) estimator for the survival function, and the second test is based on Nelson (1972) and Aalen (1978) estimator for the cumulative risk function. A similar test was proposed by Saldaña-Zepeda, Vaquera-Huerta & Arnold (2010) for assessing the goodness of fit of the Pareto distribution for type II right censored random samples.

Note that the survival function for the Gumbel distribution is:

$$S(x) = 1 - F_0(x) = 1 - \exp\left\{-\exp\left\{-\frac{x-\xi}{\theta}\right\}\right\}$$

Then

$$1 - S\left(x\right) = \exp\left\{-\exp\left\{-\frac{x-\xi}{\theta}\right\}\right\}$$

Thus, taking logarithms twice on both sides of the last expression, we have

$$y = \log\left\{-\log\left\{1 - S\left(x\right)\right\}\right\} = \frac{x - \xi}{\theta} \tag{12}$$

Equation (12) indicates that, under H_0, there is a linear relationship between y and x. Once a type II right censored random sample of size n is observed, it is possible to obtain an estimation of $S(x)$ using the Kaplan-Meier estimator:

$$\widehat{S}\left(x\right) = \prod_{x_{(i)} \leq x} \left(\frac{n - i}{n - i + 1}\right)^{\delta_i} \tag{13}$$

where $\delta_i = 0$ if the $i-$th observation is censored and $\delta_i = 1$ otherwise.

It is well known that the survival function can also be obtained from the cumulative risk function $H(x)$ since $S(x) = \exp(-H(x))$. The function $H(x)$ can be estimated using Nelson (1972) and Aalen (1978) estimator, which for a type II right censored random sample of size n from a continuous population, can be calculated as follows:

$$\widetilde{H}(x_{(i)}) = \sum_{j=1}^{i} \frac{1}{n - j + 1} \tag{14}$$

Substituting $S(x) = \exp(-H(x))$ into equation (12) we have:

$$z = \log\left\{-\log\left\{1 - \exp(-H\left(x\right))\right\}\right\} = \frac{x - \xi}{\theta} \tag{15}$$

Equation (15) indicates that, under H_0, there is a linear relationship between z and x.

The sample correlation coefficient is used for measuring the degree of linear association between x and y (x and z), which is given by:

$$R = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\bar{y} = \sum_{i=1}^{n} y_i/n$.

Let R_{K-M} and R_{N-A} denote the sample correlation coefficient based on Kaplan-Meier and Nelson-Aalen estimators, respectively. Notice that, under H_0, the values of R_{K-M} and R_{N-A} are expected to be close to one. Therefore, the decision rules for the tests based on R_{K-M} and R_{N-A} are:

- Reject H_0 at a significance level α if $R_{K-M} \leq K_{K-M}(\alpha)$, where $\alpha = P(R_{K-M} \leq K_{K-M}(\alpha)|H_0)$.
- Reject H_0 at a significance level α if $R_{N-A} \leq K_{N-A}(\alpha)$, where $\alpha = P(R_{N-A} \leq K_{N-A}(\alpha)|H_0)$.

The critical values $K_{K-M}(\alpha)$ and $K_{N-A}(\alpha)$ are the $100\alpha\%$ quantiles of the null distributions of R_{K-M} and R_{N-A} respectively. These values can be obtained by Monte Carlo simulation using the following algorithm:

1. Fix n, r, $\xi = 0$, $\theta = 1$.
2. Generate a type II right censored random sample from the Gumbel distribution, $\left(x_{(1)}, \ldots, x_{(n)}\right)$, $(\delta_1, \ldots, \delta_n)$.
3. Compute $\widehat{S}\,(x)$ and $\widetilde{H}\,(x)$ using expressions (13) and (14).
4. Calculate y and z using expressions (12) and (15).
5. Calculate R_{K-M} and R_{N-A}.
6. Repeat steps 2 to 5 B times.
7. Take $K_{K-M}(\alpha)$ and $K_{N-A}(\alpha)$ equal to the αB-th order statistic of the simulated values of R_{K-M} and R_{N-A}, respectively.

Figure 3 shows the null distributions of R_{K-M} and R_{N-A} for $n = 100$, $r = 80$ and several values for the location and scale parameters, which were obtained using $B = 10,000$ Monte Carlo samples. Observe that the null distributions of R_{K-M} and R_{N-A} are quite similar. Also notice that the mass of probability is concentrated close to one, as expected. This Figure provides an empirical confirmation of the well known fact that the sample correlation coefficient is location-scale invariant.

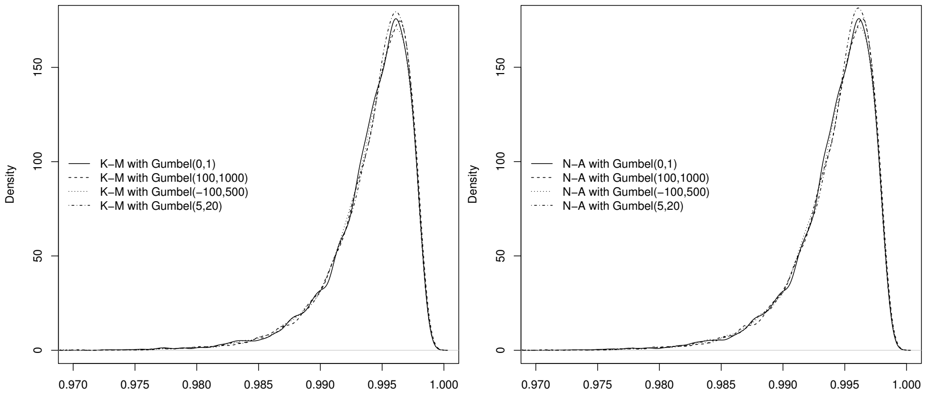


FIGURE 3: Null distribution of R_{K-M} (left) and R_{N-A} (right) for $B = 10,000$, $n = 100$, $r = 80$ and different values of the location and scale parameters.

Tables 2 and 3 contain the critical values for R_{K-M} and R_{N-A} tests corresponding to $n \leq 100$, % of censorship = 10(10)80 and $\alpha = 0.05$[1]. Notice that for

[1] An R program (R Core Team 2012) to get the critical values of R_{K-M} and R_{N-A} tests for any sample size, percentage of censorship and test size is available from the first author.

every fixed value of n, the critical values decrease as the percentage of censored observations increases. For a fixed percentage of censorship, the critical values decrease as the sample size increases, since the sample correlation coefficient is a consistent estimator.

TABLE 2: Critical values $K_{K-M}(\alpha)$ for R_{K-M} test obtained with 10,000 Monte Carlo samples.

n	**% Censored**							
	10	20	30	40	50	60	70	80
10	0.9013	0.9017	0.8948	0.8871	0.8754	0.8629	0.8686	–
20	0.9459	0.9424	0.9385	0.9296	0.9169	0.9048	0.8852	0.8686
30	0.9626	0.9619	0.9564	0.9483	0.9386	0.9271	0.9071	0.8859
40	0.9715	0.9707	0.9672	0.9608	0.9521	0.9414	0.9261	0.9006
50	0.9771	0.9757	0.9725	0.9685	0.9600	0.9507	0.9375	0.9135
60	0.9811	0.9799	0.9766	0.9722	0.9664	0.9576	0.9444	0.9238
70	0.9838	0.9824	0.9795	0.9763	0.9708	0.9632	0.9504	0.9337
80	0.9857	0.9846	0.9824	0.9789	0.9740	0.9670	0.9561	0.9398
90	0.9871	0.9863	0.9842	0.9806	0.9768	0.9703	0.9605	0.9428
100	0.9887	0.9878	0.9861	0.9830	0.9793	0.9729	0.9628	0.9460

TABLE 3: Critical values $K_{N-A}(\alpha)$ for R_{N-A} test obtained with 10,000 Monte Carlo samples.

n	**% Censored**							
	10	20	30	40	50	60	70	80
10	0.9097	0.9030	0.8960	0.8839	0.8779	0.8658	0.8671	–
20	0.9484	0.9441	0.9383	0.9302	0.9188	0.9036	0.8866	0.8679
30	0.9642	0.9618	0.9568	0.9492	0.9408	0.9260	0.9084	0.8851
40	0.9724	0.9703	0.9666	0.9612	0.9539	0.9416	0.9246	0.8997
50	0.9778	0.9762	0.9727	0.9681	0.9608	0.9508	0.9351	0.9148
60	0.9818	0.9796	0.9765	0.9726	0.9664	0.9572	0.9441	0.9239
70	0.9839	0.9831	0.9806	0.9761	0.9712	0.9631	0.9514	0.9314
80	0.9862	0.9851	0.9826	0.9786	0.9740	0.9676	0.9557	0.9380
90	0.9875	0.9864	0.9841	0.9810	0.9762	0.9698	0.9608	0.9423
100	0.9887	0.9877	0.9857	0.9832	0.9790	0.9727	0.9630	0.9471

4. Power and Size of the Tests

A Monte Carlo simulation experiment was conducted in order to study the actual level and power of the Kullback-Leibler test (KL) and the correlation tests based on Kaplan-Meier and Nelson-Aalen estimators (R_{K-M} and R_{N-A}).

Table 4 presents the actual levels of tests for several test sizes ($\alpha = 0.01$, 0.02, 0.05, 0.10 and 0.15). Observe that the estimated test size is close to the nominal test size in almost all cases.

Table 5 shows the estimated powers of KL, R_{K-M} and R_{N-A} tests against the following alternative distributions: $Weibull(3, 1)$, $Weibull(0.5, 1)$, $Gamma(3, 1)$, $Gamma(0.8, 1)$, $Log-normal(1, 1)$ and $Log-normal(5, 3)$. These alternatives in-

Table 4: Estimated test size of the KL, R_{K-M} and R_{N-A} tests.

n	**% Censored**	α	R_{K-M}	R_{N-A}	KL
20	50	0.01	0.011	0.007	0.011
		0.02	0.019	0.016	0.019
		0.05	0.055	0.050	0.058
		0.10	0.099	0.103	0.113
		0.15	0.150	0.146	0.148
50	20	0.01	0.017	0.012	0.014
		0.02	0.018	0.019	0.025
		0.05	0.047	0.050	0.053
		0.10	0.097	0.101	0.107
		0.15	0.150	0.146	0.145

clude monotone increasing, monotone decreasing and non-monotone hazard functions, just as in Saldaña-Zepeda et al. (2010). Every entry of this table was calculated using $B = 10,000$ Monte Carlo samples at a significance level $\alpha = 0.05$.

The main observations that can be made from this table are the following:

- The powers of the tests increase as the sample size increases.
- Under every considered alternative distribution, the tests lose power as the percentage of censorship gets larger for a fixed sample size.
- The KL test is in general more powerful than the correlation tests. R_{N-A} is slightly more powerful than R_{K-M}.
- The tests R_{N-A} and R_{K-M} have little power against $Gamma(3,1)$ alternatives.
- The three tests have no power against $Weibull(3,1)$ alternatives.

Table 5: Estimated power of the KL, R_{K-M} and R_{N-A} tests under several alternatives, for a significance level $\alpha = 0.05$.

Alternative	n	(%) Censored	R_{K-M}	R_{N-A}	KL
$Weibull(3,1)$	20	20	0.0950	0.0860	0.0701
		50	0.0547	0.0541	0.0493
	50	20	0.1636	0.1640	0.1264
		50	0.0559	0.0543	0.0526
	100	20	0.2989	0.2806	0.2023
		50	0.0693	0.0623	0.0907
$Weibull(0.5,1)$	20	20	0.8095	0.8445	0.9642
		50	0.5890	0.6177	0.8151
	50	20	0.9998	0.9995	1.0000
		50	0.9844	0.9850	0.9996
	100	20	1.0000	1.0000	1.0000
		50	1.0000	1.0000	1.0000
$Gamma(3,1)$	20	20	0.0330	0.0372	0.0913
		50	0.0425	0.0444	0.1090
	50	20	0.0390	0.0472	0.1344

Table 5. (Continuation)

Alternative	n	(%) Censored	R_{K-M}	R_{N-A}	KL
		50	0.0368	0.0420	0.1342
	100	20	0.0704	0.0697	0.1959
		50	0.0527	0.0530	0.1504
$Gamma(0.8, 1)$	20	20	0.2613	0.3034	0.6168
		50	0.2091	0.2277	0.4588
	50	20	0.7775	0.8081	0.9762
		50	0.6054	0.6239	0.9321
	100	20	0.9957	0.9955	0.9998
		50	0.9608	0.9632	0.9967
$Log-normal(1, 1)$	20	20	0.2180	0.2666	0.4917
		50	0.0964	0.1053	0.2864
	50	20	0.6337	0.6641	0.8254
		50	0.2242	0.2434	0.5691
	100	20	0.9543	0.9559	0.9887
		50	0.5671	0.5569	0.7433
$Log-normal(5, 2)$	20	20	0.7914	0.8280	0.9466
		50	0.4237	0.4416	0.6815
	50	20	0.9990	0.9997	1.0000
		50	0.9059	0.9043	0.9929
	100	20	1.0000	1.0000	1.0000
		50	0.9989	0.9991	1.0000

5. Application Examples

In this section, two application examples are presented, in which the hypotheses stated in equation (2) and (3) will be proven. This will allow us to carry out the goodness of fit test of the Gumbel distribution, using the Kullback-Leibler, Kaplan-Meier, and Nelson-Aalen test statistics.

Example 1. The data used in this example are from a life expectancy experiment reported by Balakrishnan & Chen (1999). Twenty three ball bearings were placed in the experiment. The data corresponds to the millions of revolutions before failure for each of the bearings. The experiment was terminated once the twentieth ball failed. The data are shown in Table 6.

Table 6: Millions of revolutions before failure for the ball bearing experiment.

x_i	δ_i	x_i	δ_i	x_i	δ_i	x_i	δ_i	x_i	δ_i
17.88	1	45.60	1	55.56	1	84.12	1	105.84	0
28.92	1	48.48	1	67.80	1	93.12	1	105.84	0
33.00	1	51.84	1	68.64	1	96.64	1	105.84	0
41.52	1	51.96	1	68.65	1	105.12	1		
42.12	1	54.12	1	68.88	1	105.84	1		

The MLE for the location and scale parameters are $\widehat{\xi} = 55.1535$ and $\widehat{\theta} = 26.8124$. The critical values for $n = 23$ and $r = 20$ can be obtained from Tables 1, 2 and 3 using interpolation. Table 7, shows the critical values for $\alpha = 0.05$, the value of the statistics $KL^*(m,n,r)$, R_{K-M} and R_{N-A}. The conclusion is that we do not have enough evidence to reject H_0 indicating that the data adjust well to a Gumbel model.

Table 7: Test comparison for example 1.

Test	Critical value	Value of the test statistic	Decision
KL	$KL_{7,23,20}(0.05) = 0.2037$	$KL^*_{m,n,r} = 0.1373$	Not reject H_0
KM	$K_{K-M}(0.05) = 0.9501$	$R_{K-M} = 0.9885$	Not reject H_0
NA	$K_{N-A}(0.05) = 0.9520$	$R_{N-A} = 0.9880$	Not reject H_0

Example 2. The data used in this example were originally presented by Xia, Yu, Cheng, Liu & Wang (2009) and then were analyzed by Saraçoğlu, Kinaci & Kundu (2012) under different censoring schemas. The data corresponds to breaking strengths of jute fiber for different gauge lengths. For illustrative purposes, we assume that only the 24/30 smallest breaking strengths for 20 mm gauge length were observed. The data are shown in Table 8. It is known that this dataset can be modeled by using an exponential distribution, so we expect to reject the null hypothesis given in (2) when applying the goodness of fit tests previously discussed.

Table 8: Breaking strength of jute fiber of gauge length 20 mm.

x_i	δ_i	x_i	δ_i	x_i	δ_i	x_i	δ_i	x_i	δ_i
36.75	1	113.85	1	187.85	1	419.02	1	585.57	0
45.58	1	116.99	1	200.16	1	456.60	1	585.57	0
48.01	1	119.86	1	244.53	1	547.44	1	585.57	0
71.46	1	145.96	1	284.64	1	578.62	1	585.57	0
83.55	1	166.49	1	350.70	1	581.60	1	585.57	0
99.72	1	187.13	1	375.81	1	585.57	1	585.57	0

The maximum likelihood estimators for the location and scale parameters are $\widehat{\xi} = 232.0995$ and $\widehat{\theta} = 210.0513$, respectively. Table 9 shows the critical values for $\alpha = 0.05$ (from Tables 1, 2 and 3) and the values of the test statistics for the data previously discussed. The three statistics reject the null hypothesis, so there is evidence that shows that the data can not be modeled by using a Gumbel distribution.

Table 9: Test comparison for example 2.

Test	Critical value	Value of the test statistic	Decision
KL	$KL_{9,30,24}(0.05) = 0.1708$	$KL^*_{m,n,r} = 0.2274$	Reject H_0
KM	$K_{K-M}(0.05) = 0.9618$	$R_{K-M} = 0.9595$	Reject H_0
NA	$K_{N-A}(0.05) = 0.9619$	$R_{N-A} = 0.9577$	Reject H_0

6. Concluding Remarks

The simulation results indicate that the proposed tests $KL^*(m,n,r)$, R_{K-M} have a good control of the type I error probability, while the R_{N-A} test under-estimate this level. The test based on the Kullback-Leibler information is better in terms of power than the tests based on the sample correlation coefficient under the considered alternative distributions. In future work, it would be interesting to derive the null distribution of the test statistics for finite samples as well as for the limit case.

Acknowledgments

The authors wish to express their thanks to three anonymous referees and the General Editor for their constructive comments which helped to greatly improve the presentation of the original version of this paper. The authors were partially funded by Subdirección de Investigación: Línea 15, Colegio de Postgraduados, México.

[Recibido: septiembre de 2011 — Aceptado: septiembre de 2012]

References

Aalen, O. (1978), 'Nonparametric inference for a family of counting processes', *Annals of Statistics* **6**(4), 701–726.

Balakrishnan, N. & Chen, W. (1999), *Handbook of Tables for Order Statistics from Lognormal Distributions with Applications*, Springer.
*http://books.google.com/books?id=x1862WoJL2EC

Castro-Kuriss, C. (2011), 'On a goodness-of-fit test for censored data from a location-scale distribution with applications', *Chilean Journal of Statistics* **2**, 115–136.

Ebrahimi, N., Habibullah, M. & Soofi, E. (1992), 'Testing exponentiality based on Kullback-Leibler information', *Journal of the Royal Statistical Society* **54**, 739–748.

Kaplan, E. L. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association* **53**(282), 457–481.
*http://dx.doi.org/10.2307/2281868

Lim, J. & Park, S. (2007), 'Censored Kullback-Leibler information and goodness-of-fit test with type II censored data', *Journal of Applied Statistics* **34**(9), 1051–1064.

Lin, C.-T., Huang, Y.-L. & Balakrishnan, N. (2008), 'A new method for goodness-of-fit testing based on type-II right censored samples', *IEEE Transactions on Reliability* **57**, 633–642.

Nelder, J. A. & Mead, R. (1965), 'A simplex algorithm for function minimization', *Computer Journal* **7**, 308–313.

Nelson, W. (1972), 'Theory and applications of hazard plotting for censored failure data', *Technometrics* **14**, 945–965.

Park, S. & Park, D. (2003), 'Correcting moments for goodness of fit tests based on two entropy estimates', *Journal of Statistical Computation and Simulation* **73**, 685–694.

Pérez-Rodríguez, P., Vaquera-Huerta, H. & Villaseñor-Alva, J. A. (2009), 'A goodness-of-fit test for the Gumbel distribution based on Kullback-Leibler information', *Communications in Statistics: Theory and Methods* **38**, 842–855.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org/

Saldaña-Zepeda, D., Vaquera-Huerta, H. & Arnold, B. C. (2010), 'A goodness of fit test for the Pareto distribution in the presence of type II censoring, based on the cumulative hazard function', *Computational Statistics and Data Analysis* **54**(4), 833–842.

Saraçoğlu, B., Kinaci, I. & Kundu, D. (2012), 'On estimation of $R = P(Y < X)$ for exponential distribution under progressive type-II censoring', *Journal of Statistical Computation and Simulation* **82**(5), 729–744.
*http://www.tandfonline.com/doi/abs/10.1080/00949655.2010.551772

Song, K. S. (2002), 'Goodness-of-fit-tests based on Kullback-Leibler discrimination information', *IEEE Transactions On Information Theory* **48**, 1103–1117.

Stephens, M. A. (1986), Tests based on edf statistics, *in* R. D'Agostino & M. Stephens, eds, 'Goodness-of-Fit Techniques', New York.

Xia, Z. P., Yu, J. Y., Cheng, L. D., Liu, L. F. & Wang, W. M. (2009), 'Study on the breaking strength of jute fibres using modified Weibull distribution', *Composites Part A: Applied Science and Manufacturing* **40**(1), 54 – 59.
*http://www.sciencedirect.com/science/article/pii/S1359835X08002595

Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 425 a 442

On the Entropy of Written Spanish

Sobre la entropía del español escrito

Fabio G. Guerrero[a]

Escuela de Ingeniería Eléctrica y Electrónica, Facultad de Ingeniería,
Universidad del Valle, Cali, Colombia

Abstract

A discussion on the entropy of the Spanish language by means of a practical method for calculating the entropy of a text by direct computer processing is presented. As an example of application, thirty samples of Spanish text are analyzed, totaling 22.8 million characters. Symbol lengths from $n = 1$ to 500 were considered for both words and characters. Both direct computer processing and the probability law of large numbers were employed for calculating the probability distribution of the symbols. An empirical relation on entropy involving the length of the text (in characters) and the number of different words in the text is presented. Statistical properties of the Spanish language when viewed as produced by a stochastic source, (such as origin shift invariance, ergodicity and asymptotic equipartition property) are also analyzed.

***Key words*:** Law of large numbers, Shannon entropy, Stochastic process, Zipf's law.

Resumen

Se presenta una discusión sobre la entropía de la lengua española por medio de un método práctico para el cálculo de la entropía de un texto mediante procesamiento informático directo. Como un ejemplo de aplicación, se analizan treinta muestras de texto español, sumando un total de 22,8 millones de caracteres. Longitudes de símbolos desde $n = 1$ hasta 500 fueron consideradas tanto para palabras como caracteres. Para el cálculo de la distribución de probabilidad de los símbolos se emplearon procesamiento directo por computador y la ley de probabilidad de los grandes números. Se presenta una relación empírica de la entropía con la longitud del texto (en caracteres) y el número de palabras diferentes en el texto. Se analizan también propiedades estadísticas de la lengua española cuando se considera como producida por una fuente estocástica, tales como la invarianza al desplazamiento del origen, ergodicidad y la propiedad de equipartición asintótica.

***Palabras clave*:** entropía de Shannon, ley de grandes números, ley de Zipf, procesos estocásticos.

[a] Assistant Professor. E-mail: fabio.guerrero@correounivalle.edu.co

1. Introduction

Spanish is a language which is used by more than four hundred million people in more than twenty countries, and it has been making its presence increasingly felt on the Internet (Marchesi 2007). Yet this language has not been as extensively researched at entropy level. The very few calculations which have been reported have been obtained, as for most languages, by indirect methods, due in part to the complexity of the problem. Having accurate entropy calculations for the Spanish language can thus be considered a pending task. Knowing the value of H, in general for any language, is useful for source coding, cryptography, language space dimension analysis, plagiarism detection, and so on. Entropy calculation is at the lowest level of language analysis because it only takes into account source symbol statistics and their statistical dependence, without any further consideration of more intelligent aspects of language such as grammar, semantics, punctuation marks (which can considerably change the meaning of a sentence), word clustering, and so on.

Several approaches have been devised for several decades for finding the entropy of a language. Shannon (1948) initially showed that one possible way to calculate the entropy of a language, H, would be through the limit $H = \lim_{n\to\infty} -\frac{1}{n}H(B_i)$, where B_i is a sequence of n symbols. Finding H using methods such as the one suggested by this approach is difficult since it assumes that the probability of the sequences, $p(Bi)$, is an asymptotically increasing function of n, as n tends to infinity. Another difficulty posed by this approach is that an extremely large sample of text would be required, one that considered all possible uses of the language. Another suggested way to calculate H is by taking $H = \lim_{n\to\infty} F_n$, where $F_n = H(j|B_i) = H(B_i j) - H(B_i)$. B_i is a block of n-1 symbols, j is the symbol next to B_i, $H(j|B_i)$ is the conditional entropy of symbol j given block B_i. In this approach, the series of approximations $F_1, F_2, \ldots$ provides progressive values of conditional entropy. F_n, in bits/symbol, measures the amount of information in a symbol considering the previous $n-1$ consecutive symbols, due to the statistics of the language. The difficulty of using these previous methods in practice was put under evidence when in his pioneering work Shannon (1951) used instead a human prediction approach for estimating the entropy of English, getting 0.6 and 1.3 bits/letter as bounds for printed English, considering 100-letter sequences. Gambling estimations have also been used, providing an entropy estimation of 1.25 bits per character for English (Cover & King 1978). The entropy rate of a language could also be estimated using ideal source coders since, by definition, this kind of coder should compress to the entropy limit. A value of 1.46 bits per character has been reported for the entropy of English by means of data compression (Teahan & Cleary 1996). The entropy of the fruit fly genetic code has been estimated using universal data compression algorithms (Wyner, Jacob & Wyner 1998). As for the Spanish language, values of 4.70, 4.015, and 1.97 bits/letter for F_0, F_1, and F_W respectively were reported (Barnard III 1955) using an extrapolation technique on frequency data obtained from a sample of 6,513 different words. F_W is the entropy, in bits/letter, based on single-word frequency.

Another venue for finding H has been based on a purely mathematical framework derived from stochastic theory, such as the one proposed by Crutchfield & Feldman (2003). Unfortunately, as the same authors recognize it, it has lead, in practice, to very limited results for finding the entropy of a language. In general, as all these results suggest, finding the entropy of a language by classic methods has proved to be a challenging task. Despite some remarkable findings in the past decades, the search for a unified mathematical model continues to be an open problem (Debowski 2011).

In the past it was implicitly believed that attempting to find the average uncertainty content of a language by direct analysis of sufficiently long samples could be a very difficult task to accomplish. Fortunately, computer processing capacity available at present has made feasible tackling some computing intensive problems such as the search in large geometric spaces employed in this work. Michel, Shen, Aiden, Veres, Gray, Team, Pickett, Hoiberg, Clancy, Norvig, Orwant, Pinker, Nowak & Aiden (2011) discuss, as an example of this trend, the use of huge computational resources to research the relationship between linguistics and cultural phenomena. This paper is organized as follows: In Section 2 the methodology used to obtain all the values reported is discussed; in Section 3 the results of the observations are presented; Section 4 presents a discussion and analysis of the most relevant results and, finally, in Section 5 the main conclusions of this work are summarized. All the samples and support material used in this work are publicly available at http://sistel-uv.univalle.edu.co/EWS.html. Aspects such as the analysis of grammar, semantics, and compression theory are beyond the scope of this paper.

2. Methodology

Thirty samples of literature available in Spanish were chosen for this study. Tables 1 and 2 show the details of the samples and its basic statistics. The works used in this paper as samples of written Spanish were obtained from public libraries available on the Internet such as librodot [1] and the virtual library Miguel de Cervantes [2]. The selection of the samples was done without any particular consideration of publication period, author's country of origin, and suchlike. A file of news provided to the author by the Spanish press agency EFE was also included in the samples for analysis. The selected material was processed using a T3500 Dell workstation with 4 GB RAM. The software used to do the all the calculations presented in this work was written in Mathematica® 8.0. For simplicity, a slight preprocessing was done on each sample, leaving only printable characters. Strings of several spaces were reduced to one character and the line feed control character (carry return) was replaced by a space character, allowing for fairer comparisons between samples. The samples were character encoded using the ISO 8859-1 standard (8-bit single-byte coded graphic character sets - Part 1: Latin alphabet No. 1) which has 191 characters from the Latin script, providing a full set of charac-

[1] http://www.librodot.com

[2] http://www.cervantesvirtual.com

ters for the Spanish language. For instance, the $\tilde{n}$ letter corresponds to 0xf1, etc. The total amount of characters of the thirty samples in table 1 is 22,882,449 and the total amount of words is 4,024,911. The rounded average for the number of different one-character symbols (uppercase, lowercase, and punctuation marks) for the thirty samples was 93. The reason we consider the distinction between uppercase and lowercase symbols is that when characterizing an information source at entropy level, lowercase and uppercase symbols produce different message vectors from the transmission point of view (e.g. the word *HELLO* produces a completely different message vector than the word *hello*).

Table 1: Set of Text Samples

Sample	Name	Author
1	La Biblia	Several authors
2	efe-B2	EFE Press agency
3	Amalia	José Mármol
4	Crimen y Castigo	Fyodor Dostoevsky
5	Rayuela	Julio Cortázar
6	Doña Urraca de Castilla	F. Navarro Villoslada
7	El Corán	Prophet Muhammad
8	Cien Años de Soledad	Gabriel García Márquez
9	La Araucana	Alonso de Ercilla
10	El Papa Verde	Miguel Angel Asturias
11	América	Franz Kafka
12	La Altísima	Felipe Trigo
13	Al Primer Vuelo	José María de Pereda
14	Harry Potter y la Cámara Secreta	J.K. Rowling
15	María	Jorge Isaacs
16	Adiós a las Armas	Ernest Hemingway
17	Colmillo Blanco	Jack London
18	El Alférez Real	Eustaquio Palacios
19	Cañas y Barro	Vicente Blasco Ibáñez
20	Aurora Roja	Pío Baroja
21	El Comendador Mendoza	Juan C. Valera
22	El Archipiélago en Llamas	Jules Verne
23	Doña Luz	Juan Valera
24	El Cisne de Vilamorta	Emilia Pardo Bazán
25	Cuarto Menguante	Enrique Cerdán Tato
26	Las Cerezas del Cementerio	Gabriel Miró
27	Tristana	Benito Pérez Galdós
28	Historia de la Vida del Buscón	Francisco de Quevedo
29	El Caudillo	Armando José del Valle
30	Creció Espesa la Yerba	Carmen Conde

In Table 2 the parameter α is the average word length, given by $\sum L_i p_i$, where L_i and p_i are the length in characters and the probability of the i-th word respectively. The weighted average of α for the whole set of samples is 4.491 letters per word. The word dispersion ratio, WDR, is the percentage of different words over the total number of words.

The values of entropy were calculated using the entropy formula $\sum p_i \log_2 p_i$. The frequency of the different symbols (n-character or n-word symbols) and the law of large numbers were used to find the symbol probabilities as $p_i \approx n_i/n_{total}$. First, we started considering word symbols, since words are the constituent elements of the language. However, a more refined analysis based on characters was also carried out. Entropy values for both n-character and n-word symbols from n=1 to 500 were calculated. Considering symbols up to a length of five hundred was a suitable number for practical proposes, this will be discussed in the next section.

TABLE 2: Sample Details

Sample	Number of Characters	Alphabet Size (A_S)	Number of Words	Different Words	WDR(%)	α
1	5722041	100	1049511	40806	3.89	4.27
2	1669584	110	279917	27780	9.92	4.80
3	1327689	88	231860	18871	8.14	4.51
4	1215215	91	207444	17687	8.53	4.63
5	984129	117	172754	22412	12.97	4.50
6	939952	84	161828	17487	10.81	4.58
7	884841	93	160583	12236	7.62	4.32
8	805614	84	137783	15970	11.59	4.73
9	751698	82	129888	15128	11.65	4.63
10	676121	93	118343	16731	14.14	4.45
11	594392	88	101904	11219	11.01	4.66
12	573399	89	98577	14645	14.86	4.53
13	563060	82	100797	13163	13.06	4.35
14	528706	89	91384	10884	11.91	4.60
15	499131	87	88376	12680	14.35	4.45
16	471391	91	81803	10069	12.31	4.49
17	465032	91	81223	10027	12.35	4.58
18	462326	89	82552	10699	12.96	4.43
19	436444	79	75008	10741	14.32	4.66
20	393920	90	68729	10598	15.42	4.47
21	387617	86	69549	10289	14.79	4.38
22	363171	88	61384	8472	13.80	4.73
23	331921	83	59486	9779	16.44	4.41
24	312174	77	53035	11857	22.36	4.65
25	304837	87	49835	12945	25.98	4.95
26	302100	75	51544	10210	19.81	4.64
27	299951	82	52571	10580	20.13	4.48
28	232236	74	42956	7660	17.83	4.23
29	224382	83	36474	7470	20.48	5.00
30	159375	81	27813	6087	21.89	4.48

One worthwhile question at this point is "does entropy change when changing the origin point in the sample?". For this purpose, we calculated entropy values considering symbols for different shifts from the origin for non overlapping symbols, as illustrated by figure 1, for the case of trigrams.

It can easily be seen that, for symbols of length n, symbols start repeating (i.e., symbols are the same as for shift=0, except for the first one) after n shifts. As a result, the number of individual entropy calculations when analyzing symbols from length $n = 1$ up to k was $\frac{k(k+1)}{2}$. For the $k = 500$ case used in this work, this gives 125,250 individual entropy calculations for every sample analyzed. The

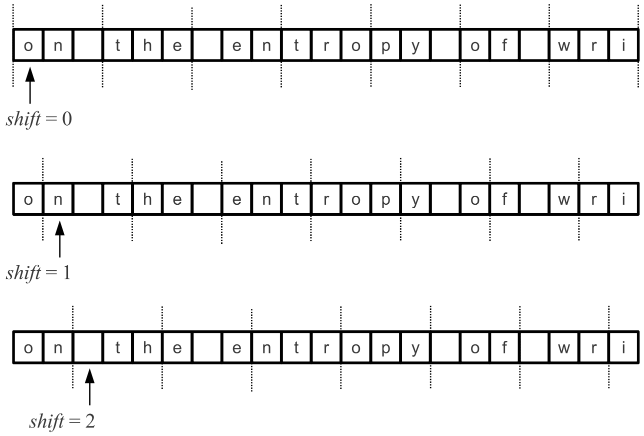


Figure 1: Origin invariance analysis.

individual shift entropies so obtained were then averaged for every n. Values of n for which the maximum value of entropy was produced were identified, as well as values of n from which all symbols present in the text become equiprobable with reasonable certainty, i.e., none of them repeat more than once in the sample.

3. Results

3.1. Entropy Values Considering Words

Figure 2 shows the values of average entropy for n-word symbols. For ease of display, only data for samples 1, 2, 12 and 30 and $n = 1$ to 20 are shown. The rest of the literary works exhibited the same curve shapes with values in between. All the analyzed samples exhibited invariance to origin shift. For example, for sample 8 (*Cien Años de Soledad*) the values for $n = 4$ were: 15.024492 ($shift = 0$), 15.028578 ($shift = 1$), 15.025693 ($shift = 2$), 15.027212 ($shift = 3$). This means that $P(w_1, ..w_L) = P(w_{1+s}, ..w_{L+s})$ for any integer s, where $\{w_1, ..w_L\}$ is a L-word sequence. This is a very useful property to quickly find the entropy of a text it because it makes necessary to compute values for just one shift thus reducing the process to a few seconds for practical purposes.

Also since the weighted value for 1-word entropy for the set analyzed was 10.0064 bits/character, the weighted value of F_W is therefore 2.23 bits/character.

3.2. Entropy Values Considering n-Character Symbols

Figure 3 shows the averaged entropy values for n-character symbols. Again for ease of display, only data for samples 1, 2, 12 and 30 and $n = 1$ to 100 are shown. All samples also exhibited the origin shift invariance property. For example, for sample 8 (*Cien Años de Soledad*), the values of entropy for $n = 4$ characters were:

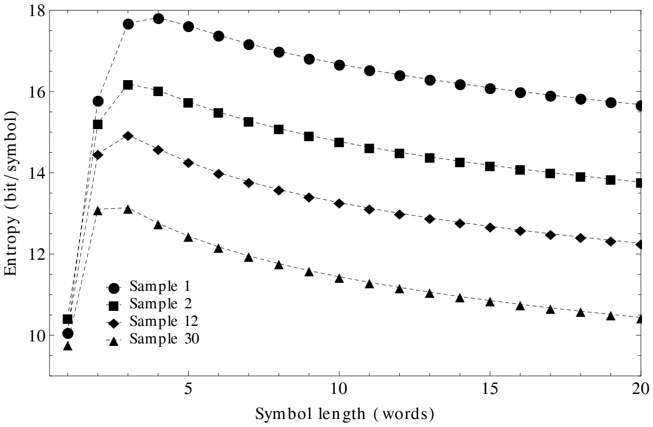


FIGURE 2: Entropy for n-word symbols for samples 1, 2, 12 and 30.

12.267881 ($shift = 0$), 12.264343 ($shift = 1$), 12.268751 ($shift = 2$), 12.269691 ($shift = 3$). Therefore, $P(c_1, .., c_L) = P(c_{1+s}, .., c_{L+s})$ for any integer s. As in the case of words, the rest of literary works exhibited the same curve shapes with values in between.

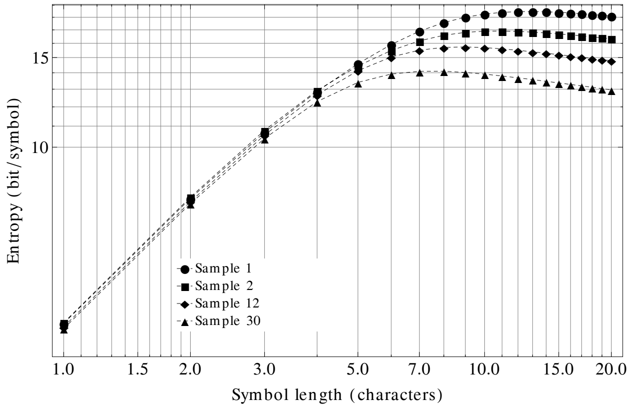


FIGURE 3: Log Plot of entropy for n-character symbols for samples 1, 2, 12 and 30.

3.3. Processing Time

Figure 4 shows the processing time of every sample for both words and characters for *all* shifts of n ($1 \leq n \leq 500$), that is, 125,250 entropy calculations for each sample. Due to the origin shift invariance property, only calculations for one shift (for instance $shift = 0$) are strictly required thus reducing the time substantially. For example, the processing time of sample 1 for only one shift was 433 seconds while the processing time for sample 30 was just nine seconds. Analysis for all shifts of n were done in this work in order to see if entropy varied when changing the point of origin in the text. A carefully designed algorithm based on Mathematica's sorting functions was employed to obtain the probability of symbols, however, a discussion on the optimality of this processing is beyond the scope of this paper.

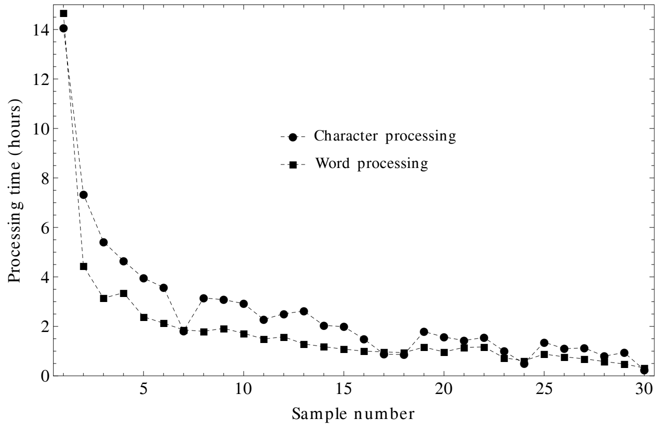


Figure 4: Processing time considering *all* shifts of n (125,250 entropy calculations)

3.4. Reverse Entropy

If we take the text in reverse order, for instance "yportne eht no" instead of "on the entropy", it is possible to evaluate the reverse conditional entropy, that is, the effect of knowing how much information can be gained about a previous character when later characters are known. It was observed that entropy of the reverse text carried out for the same set of samples produced exactly the same values as for the forward entropy case. This was first observed by Shannon for the case of the English language in his classical work (Shannon 1951) on English prediction.

4. Discussion

4.1. Frequency of Symbols and Entropy

Figure 5 shows a plot of the fundamental measure function of information, $p_i \log_2 p_i$, which is at the core of the entropy formula. This function has its maximum, 0.530738, at $p_i = 0.36788$. Therefore, infrequent symbols, as well as very frequent symbols, add very little to the total entropy. This should not be confused with the value of $p_i = \frac{1}{n}$ that produces the maximum amount of entropy for a probability space with n possible outcomes. The entropy model certainly has some limitations because entropy calculation is based solely on probability distribution. In fact, two different texts with very different location of words can have the same entropy, yet one of them can lead to a very much more efficient source encoding than the other.

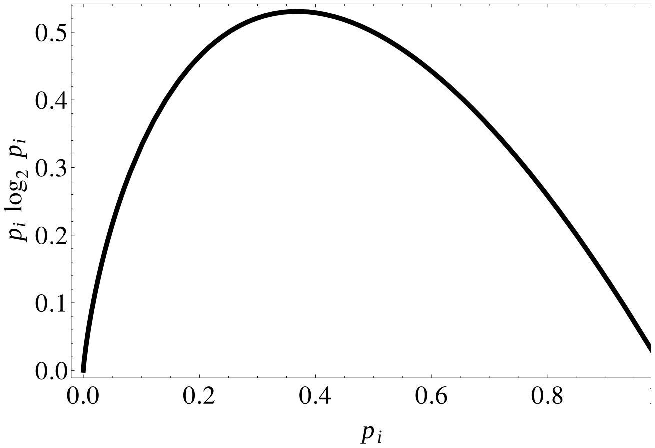


FIGURE 5: Fundamental function of information.

4.2. Log-log Plots and the Spanish Language Constant

The fact that the basic statistical properties of entropy are essentially the same for short length symbols regardless of the sample (and the entropy is similar for any shift of the origin) means it is possible to use a sufficiently long sample, for instance sample 2, to study the Spanish language constant. Figure 6 shows the log-log plot for sample 2 which contained 82,656 different 3-word symbols, 79,704 different 2-word symbols, and 27,780 different 1-word symbols. Log-log plots for the rest of samples were found to be similar to those of figure 6, at least for 2-word and 1-word symbols.

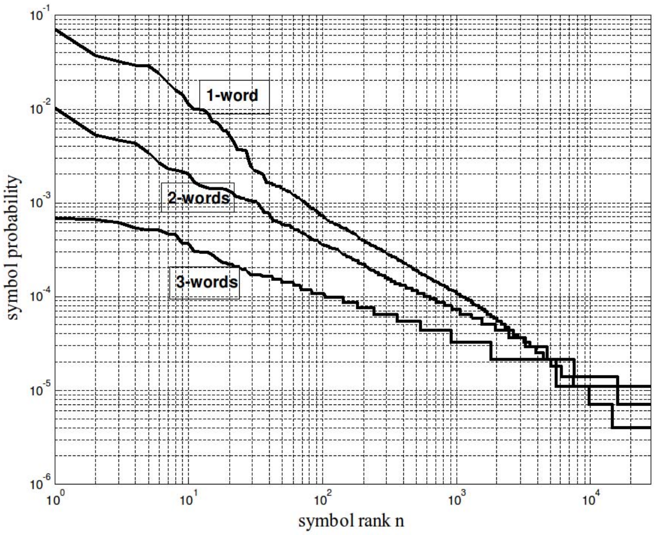


FIGURE 6: Symbol rank versus n-word probability in Sample 2.

Smoothing the 1-word curve in figure 6, the probability of the r-th most frequent 1-word symbol is close to $0.08/r$, assuming r is not too large. This behavior corresponds to the celebrated Zipf law first presented in 1939 (Zipf 1965) which nowadays some authors also call the Zipf-Mandelbrot law (Debowski 2011). Figure 7 shows the log-log plot for for sample 2 which contained 14,693 different trigrams, 2,512 different digrams, and 110 different characters; all values considered for *shift* $= 0$. Log-log plots for the rest of the samples were found to be similar to those of figure 7. Even when a distinction between upper case and lower case symbols is made in this work, no significant difference was found with the constant obtained when analyzing the database of the 81,323 most frequent words (which makes no distinction between upper case and lower case symbols). This database was compiled by Alameda & Cuetos (1995) from a corpus of 1,950,375 words of written Spanish.

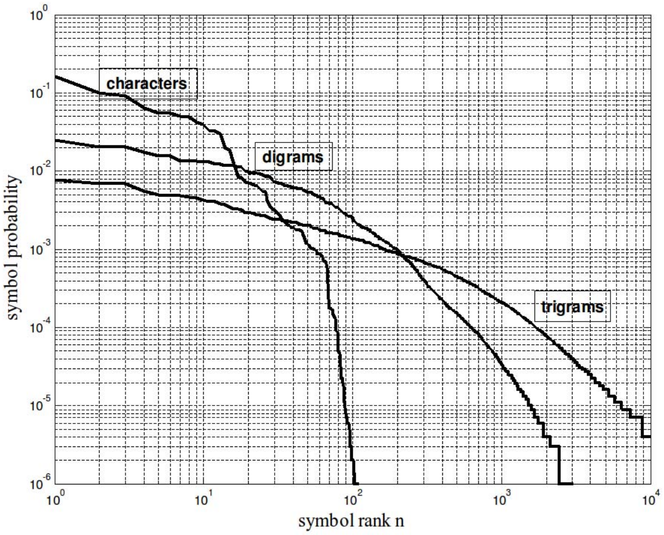


FIGURE 7: Symbol rank versus n-character probability for Sample 2.

4.3. Conditional Entropy

We now evaluate the uncertainty content of a character given some previous text. Initially F_0, in bits per character, is given by $\log_2(A_S)$, where A_S is the alphabet size. F_1 takes into account single-character frequencies and it is given by $F_1 = \sum_i p_i \log_2 p_i$. F_2 considers the uncertainty content of a character given the previous one:

$$F_2 = -\sum_{i,j} p(i,j) \log_2 p(j|i) = -\sum_{i,j} p(i,j) \log_2 p(i,j) + \sum_i p_i \log_2 p_i \qquad (1)$$

Similarly, F_3 gives the entropy of a character given the previous two characters (digram):

$$F_3 = -\sum_{i,j,k} p(i,j,k)\log_2 p(k|ij) = -\sum_{i,j,k} p(i,j,k)\log_2 p(i,j,k) + \sum_{i,j} p_{i,j}\log_2 p_{i,j} \tag{2}$$

and so on. Table 3 shows, for simplicity, values for F_n from F_1 to F_{15} only, rounded to two significant digits.

TABLE 3: Conditional Entropy F_n

								n							
S_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4.51	3.43	2.76	2.18	1.72	1.33	0.98	0.68	0.43	0.26	0.14	0.06	0.01	-0.02	-0.04
2	4.52	3.46	2.82	2.13	1.52	1.05	0.69	0.42	0.22	0.09	0.01	-0.03	-0.06	-0.06	-0.07
3	4.39	3.34	2.73	2.11	1.57	1.11	0.72	0.42	0.21	0.08	-0.01	-0.05	-0.07	-0.08	-0.08
4	4.43	3.39	2.74	2.09	1.54	1.07	0.67	0.37	0.18	0.06	-0.01	-0.05	-0.06	-0.07	-0.08
5	4.40	3.41	2.81	2.16	1.55	1.02	0.59	0.30	0.12	0.01	-0.05	-0.08	-0.08	-0.09	-0.09
6	4.39	3.35	2.74	2.13	1.56	1.05	0.62	0.32	0.13	0.02	-0.04	-0.07	-0.08	-0.09	-0.09
7	4.46	3.31	2.57	1.93	1.40	0.96	0.61	0.36	0.20	0.09	0.03	-0.02	-0.04	-0.06	-0.06
8	4.27	3.27	2.67	2.06	1.50	1.02	0.63	0.34	0.16	0.05	-0.02	-0.06	-0.07	-0.08	-0.08
9	4.32	3.28	2.70	2.11	1.58	1.06	0.61	0.29	0.10	-0.01	-0.06	-0.09	-0.10	-0.10	-0.09
10	4.40	3.36	2.78	2.16	1.51	0.95	0.51	0.22	0.05	-0.04	-0.08	-0.09	-0.10	-0.10	-0.09
11	4.33	3.32	2.66	2.00	1.43	0.93	0.54	0.28	0.11	0.01	-0.04	-0.07	-0.08	-0.09	-0.09
12	4.44	3.38	2.74	2.11	1.46	0.90	0.47	0.20	0.03	-0.05	-0.09	-0.10	-0.10	-0.10	-0.10
13	4.36	3.31	2.71	2.07	1.47	0.93	0.52	0.23	0.07	-0.03	-0.07	-0.09	-0.09	-0.10	-0.09
14	4.44	3.40	2.69	1.98	1.35	0.84	0.47	0.22	0.08	-0.01	-0.05	-0.07	-0.08	-0.09	-0.09
15	4.38	3.33	2.72	2.07	1.43	0.87	0.46	0.20	0.05	-0.04	-0.08	-0.09	-0.10	-0.10	-0.09
16	4.46	3.35	2.69	2.00	1.37	0.83	0.44	0.19	0.05	-0.03	-0.07	-0.08	-0.09	-0.09	-0.09
17	4.32	3.30	2.63	1.98	1.39	0.89	0.50	0.24	0.07	-0.01	-0.06	-0.08	-0.09	-0.09	-0.09
18	4.35	3.33	2.71	2.05	1.41	0.86	0.45	0.19	0.04	-0.04	-0.08	-0.09	-0.10	-0.09	-0.09
19	4.29	3.29	2.64	1.98	1.37	0.87	0.49	0.23	0.08	-0.01	-0.06	-0.08	-0.09	-0.09	-0.09
20	4.44	3.37	2.73	2.03	1.34	0.78	0.37	0.14	0.01	-0.06	-0.08	-0.10	-0.10	-0.10	-0.09
21	4.37	3.33	2.71	2.04	1.37	0.79	0.39	0.13	0.00	-0.05	-0.09	-0.09	-0.11	-0.06	-0.12
22	4.38	3.34	2.65	1.91	1.26	0.75	0.40	0.17	0.05	-0.03	-0.05	-0.08	-0.08	-0.08	-0.08
23	4.34	3.30	2.67	2.00	1.35	0.78	0.38	0.13	0.01	-0.06	-0.09	-0.10	-0.10	-0.10	-0.09
24	4.38	3.36	2.78	2.08	1.34	0.71	0.30	0.06	-0.05	-0.09	-0.11	-0.11	-0.11	-0.10	-0.10
25	4.32	3.37	2.80	2.09	1.32	0.69	0.29	0.06	-0.05	-0.09	-0.10	-0.11	-0.11	-0.10	-0.10
26	4.42	3.35	2.71	2.01	1.28	0.71	0.32	0.10	-0.03	-0.07	-0.10	-0.11	-0.10	-0.10	-0.10
27	4.37	3.34	2.74	2.06	1.33	0.72	0.31	0.08	-0.04	-0.09	-0.11	-0.11	-0.11	-0.10	-0.10
28	4.33	3.25	2.63	1.94	1.26	0.71	0.32	0.10	-0.03	-0.09	-0.10	-0.11	-0.11	-0.10	-0.10
29	4.28	3.28	2.62	1.89	1.21	0.68	0.32	0.11	0.00	-0.07	-0.09	-0.10	-0.10	-0.09	-0.09
30	4.40	3.35	2.66	1.89	1.11	0.52	0.17	0.01	-0.08	-0.11	-0.12	-0.11	-0.11	-0.11	-0.10

We observe in table 3 that, at some point, conditional entropies become negative. Although $H(X,Y)$ should always be greater or equal to $H(Y)$, the estimation on conditional entropy in this study becomes negative because the length of the text is not sufficiently long, in contrast to the required condition of the theoretical model $n \to \infty$. This behavior has also been observed in the context of bioinformatics and linguistics (Kaltchenko & Laurier 2004). The following example should help to clarify the explanation. Let's consider first the following text in Spanish which has 1000 characters: ⟨⟨*Yo, señora, soy de Segovia. Mi padre se llamó Clemente Pablo, natural del mismo pueblo; Dios le tenga en el cielo. Fue, tal como todos dicen, de oficio barbero, aunque eran tan altos sus pensamientos que se corría de que le llamasen así, diciendo que él era tundidor de mejillas y sastre de barbas. Dicen que era de muy buena cepa, y según él bebía es cosa para creer. Estuvo casado con Aldonza de San Pedro, hija de Diego de San Juan y nieta de Andrés de San Cristóbal. Sospechábase en el pueblo que no era cristiana vieja, aun viéndola con canas y rota, aunque ella, por los nombres y sobrenombres de sus pasados, quiso esforzar que era descendiente de la gloria. Tuvo muy buen parecer para letrado; mujer de amigas y cuadrilla, y de pocos enemigos, porque hasta los tres del alma no los tuvo por tales; persona de valor y* *conocida por quien era. Padeció grandes trabajos recién casada, y aun después, porque malas lenguas daban en decir que mi padre metía el dos de bastos para sacar el as de oros*⟩⟩. This text has 250 four-character symbols (e.g. {Yo, },{seño},{ra, }) with 227 of them being different. Factorizing common probability terms we find: $H_{4-char} = 209(\frac{1}{250}\log_2\frac{1}{250})+14(\frac{1}{125}\log_2\frac{1}{125})+3(\frac{3}{250}\log_2\frac{3}{250})+\frac{2}{125}\log_2\frac{2}{125} = 7.76$ bits/symbol. This text has 200 five-character symbols (e.g.{Yo, s},{eñora},{, soy}) with 192 being different five-character symbols. Factorizing common probability terms we find: $H_{5-char} = 184(\frac{1}{200}\log_2\frac{1}{200}) + 8(\frac{1}{100}\log_2\frac{1}{100}) = 7.56$ bits/symbol. Thus the entropy of a character given the previous four characters are know and would be $H(X|Y) = H_{5-char} - H_{4-char} = -0.20$ bits/character. For sample 1 (which has 5,722,040 characters) a similar behavior is observed: The greatest number of different symbols (418,993) occurs for n=10 (572,204 total 10-character symbols) for which H=18.26 bits/symbol. The highest entropy, 18.47 bits/symbol, is produced by 13-character symbols (there are 440,156 total 13-character symbols, and 395,104 different 13-character symbols). For 14-character symbols (408,717 total; 378,750 different) the entropy is 18.45 bits/symbol. Then the entropy of a character given the previous thirteen characters are know, in this case, is $18.45 - 18.47 = -0.02$ bits/character. With increasing n, the probability distribution tends to become uniform and H starts decreasing monotonically with n, as shown in figure 3 of the paper. When the symbols in the sample become equiprobable the value of H is given by $\log_2\lfloor\frac{\text{total number of characters}}{n}\rfloor$. Again, these seemingly paradoxical values are explained by the differences between mathematical models and real world, as well as the assumptions on which they are based[3].

4.4. Entropy Rate and Redundancy

To estimate the entropy rate, a polynomial interpolation of third degree is first applied to the values of F_n. As an example, figure 8 shows the interpolated curves for samples one and thirty.

Figure 8 shows that F_n becomes negative after crossing by zero, and from this point asymptotically approaches zero as $n \to \infty$. Therefore,

$$\lim_{n\to\infty} F_n = \lim_{n\to N_Z} F_n \tag{3}$$

In equation 3, N_Z is the root of the interpolated function F_n. The n-character entropy values of figure 3 are also interpolated to find H_{NZ}, the entropy value corresponding to N_Z. The redundancy is given by $R = \frac{H_L}{H_{max}}$, where H_L is the source's entropy rate, and $H_{max} = \log_2(A_S)$. Finally, the value of H_L is calculated as $H_L \approx \frac{H_{NZ}}{N_Z}$. Table 4 summarizes the values of N_Z, H_{NZ}, H_L, and R. It should be clear that the previous interpolation process is used to get a finer approximation to the value of entropy. Just as in thermodynamics a system in equilibrium state produces maximum entropy, equation 3 captures the symbol distribution that produces the highest level of entropy (or amount of information) in the text.

[3] An insightful dissertation on real world and models is presented in Slepian (1976).

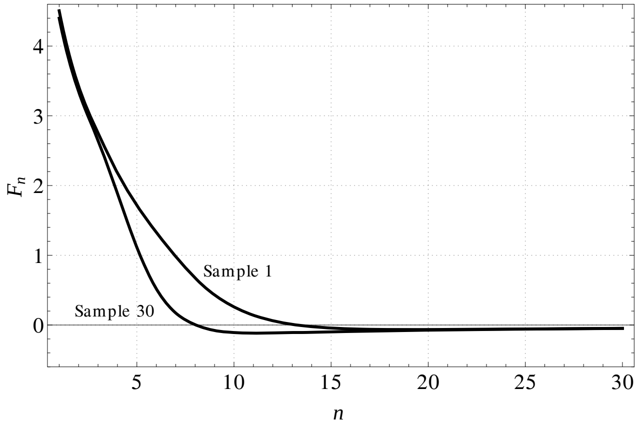


FIGURE 8: Interpolated curves of conditional entropy (bits/character) for samples 1 and 30.

In Table 4, the weighted average of H_L is 1.54 bits/character. Since the weighted average of the alphabet size in Table 1 is 92.98 characters the average redundancy, R, for the analyzed sample set, comprising nearly 23 million characters, is:

$$R = 1 - \frac{1.54}{\log_2 92.98} \approx 76.486\%$$

Taking $H(X)$ equal to 1.54 bits/character, for a text of Spanish of 140 characters, there would exist $2^{nH(X)} \approx 7.98 \times 10^{64}$ typical sequences. Because the roots of F_n occur at small values of n and, as it has been observed this method permits to find the value of entropy in a very short time (analysis for only one shift, for instance *shift*=0, is required). As it can be observed in Table 4, in general, a sample with lower WDR has more redundancy, the opposite also being true. In general, and as a consequence of Zipf's, law the greater the size of a sample, the smaller its WDR. An interesting empirical relation found in this work involving H_L, the length of the text (in characters) L, and the number of different words (V) in the text is:

$$H_L \approx \frac{2.192}{\log_V L} \tag{4}$$

Equation 4 indicates that texts with small word dictionaries (compared to the length of the text in characters) have smaller H_L because there is higher redundancy. This corroborates the well known fact that larger documents are more compressible than smaller ones. The compression factor[4] using bzip compression for samples 1 and 30 is 0.25 and 0.33 respectively, which is in total agreement with sample 1 having more redundancy than sample 30. Equation 4 is a reasonable approximation considering that in this work L takes into consideration punctuation

[4] The compression factor is defined in this work as the size after compression over the size before compression.

Table 4: Entropy Rate and Redundancy

Sample	N_Z	H_{NZ}	H_L	$R(\%)$
1	13.23	18.47	1.40	78.99
2	11.23	16.91	1.51	77.80
3	10.92	16.67	1.53	76.38
4	10.82	16.53	1.53	76.54
5	10.10	16.36	1.62	76.43
6	10.22	16.28	1.59	75.08
7	11.54	15.91	1.38	78.92
8	10.60	15.95	1.50	76.46
9	9.90	16.05	1.62	74.49
10	9.48	15.93	1.68	74.31
11	10.14	15.61	1.54	76.16
12	9.31	15.73	1.69	73.92
13	9.64	15.64	1.62	74.47
14	9.92	15.46	1.56	75.94
15	9.49	15.50	1.63	74.64
16	9.55	15.36	1.61	75.28
17	9.79	15.30	1.56	75.98
18	9.44	15.37	1.63	74.86
19	9.81	15.23	1.55	75.36
20	9.11	15.19	1.67	74.32
21	9.01	15.14	1.68	73.85
22	9.60	14.90	1.55	75.98
23	9.06	14.96	1.65	74.12
24	8.47	14.99	1.77	71.77
25	8.48	14.94	1.76	72.66
26	8.72	14.89	1.71	72.58
27	8.58	14.93	1.74	72.61
28	8.71	14.53	1.67	73.14
29	9.01	14.40	1.60	74.93
30	8.05	14.10	1.75	72.38

marks. Figure 9 is intended to illustrate that as a sample has a higher WDR, there is a tendency to the equipartition of the sample space, increasing thus H_L.

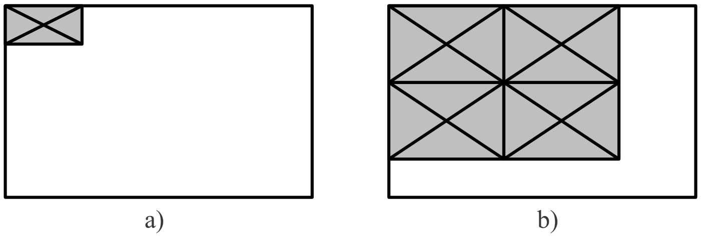


Figure 9: Illustration of Word Dispersion Ratio over word space: a) Lower WDR. b) Higher WDR.

The twenty-second version of the Dictionary of the Royal Academy of the Spanish Language (DRAS) has 88,431 lemmas (entries) with 161,962 definitions (i.e., meanings for the words according to the context in which they appear). If compared to the total number of lemmas of the DRAS, the works analyzed in this work use a relatively small number of words. For instance, literature's Nobel Prize winner Gabriel García Márquez in his masterpiece, *Cien Años de Soledad*, used around sixteen thousand different words. Because the vocabulary at the end is finite, the WDR for larger texts has to be, in general, smaller.

Finally, when concatenating the whole set of thirty samples to form one larger sample (22.9 million characters) the results were: $\alpha = 4.491$ letter/word, $H_L = 1.496$ bits/character, and $R = 78.92\%$. The computing time ($shift = 0$) was thirty four minutes.

Many other samples of Spanish can be analyzed (for instance, science, sports, etc.) but Table 4 should give a good indication of what to expect in terms of the entropy for ordinary samples of written Spanish. However, as Table 4 also shows, finding an exact value for the entropy of Spanish is an elusive goal. We can only make estimations of entropy for particular text samples. The usefulness of the method presented here lies on its ability to provide a direct entropy estimation of a particular text sample.

4.5. Character Equiprobability Distance

We define the character equiprobability distance of a text sample, n_{aep}, as the value of n such that for any $n \geq n_{aep}$, all n-length symbols in the sample become equiprobable for all shifts of n. This means,

$$H = \log_2 \left\lfloor \frac{(\text{Total number of characters}) - shift}{n} \right\rfloor$$

for all $n \geq n_{aep}$. This definition demands symbol equiprobability for all shifts for every $n \geq n_{aep}$, in other words, every substring of length $n \geq n_{aep}$ only appears once, not matter its position in the text.

Table 5 shows the values of n_{aep} evaluated from $n = 1$ to 500 characters and $2^{n_{aep} H_L}$, the number of typical sequences of length n_{aep} characters. Plagiarism detection tools should take into account the value of n_{aep}, because for sequences shorter than n_{aep} characters, it is more likely to find similar substrings of text due to the natural restriction imposed by the statistical structure of the language. Large values of n_{aep} in Table 5 were found to be related to some text reuse such as, for instance, sample 2 where some partial news are repeated as part of a larger updated news report. As it is observed, the number of typical sequences is of considerable size despite the apparently small number of characters involved.

5. Conclusions

The evidence analyzed in this work shows that the joint probability distribution of Spanish does not change with position in time (origin shift invariance). Due to this property the method for finding the entropy of a sample of Spanish presented in this work is simple and computing time efficient. Both, a redundancy of 76.5%

Table 5: Equiprobable Distance ($1 \leq n \leq 500$)

Sample	n_{aep}	$2^{n_{aep}H_L}$
1	412	4.31E+173
2	452	2.88E+205
3	93	6.82E+042
4	53	2.57E+024
5	356	4.07E+173
6	124	2.24E+059
7	101	9.07E+041
8	76	2.08E+034
9	36	3.60E+017
10	116	4.62E+058
11	39	1.20E+018
12	255	5.36E+129
13	189	1.48E+092
14	84	2.80E+039
15	50	3.42E+024
16	61	3.67E+029
17	118	2.59E+055
18	208	1.15E+102
19	37	1.84E+017
20	43	4.14E+021
21	453	1.25E+229
22	69	1.57E+032
23	43	2.28E+021
24	29	2.83E+015
25	55	1.38E+029
26	38	3.64E+019
27	27	1.39E+014
28	32	1.22E+016
29	50	1.21E+024
30	43	4.49E+022

and a rate entropy of 1.54 bits/character were found for the sample set analyzed. A value of 2.23 bits/character was found for F_W. In general, lower values of WDR were observed for longer samples leading to higher values of redundancy, just in accordance with Zipf's law. Evidence also shows that, for every day texts of the Spanish language, $p(B_i)$ is not an asymptotically increasing function of n and the highest moment of uncertainty in a sample occurs for a relatively small value of n. Considering n-word symbols, $H_{\max}$ was found at a value of four or less words. When considering n-character symbols, $_{H\max}$ was found at a value of fourteen or less characters. An averaged value of n_{aep} close to 125 characters can be a good indication of how constrained we are by the statistical structure of the language. The probability of the r-th most frequent word in Spanish is approximately $0.08/r$. If compared to the constant of English, $0.1/r$, it can be concluded that the total probability of words in Spanish is spread among more words than in English. There is a clear indication of the relation between a text's dictionary size (number of different words) and H_L. In general, a text with a larger dictionary size causes H_L to increase. Texts with small word dictionaries compared to the length of the text in characters have smaller H_L and thus should be more compressible. Since reverse entropy analysis produced exactly the same values as forward entropy, for prediction purposes the amount of uncertainty when predicting a text backwards is, despite being apparently more difficult, the same as predicting the text forwards. Finally, despite the fact that the basic statistical properties are similar regardless of the text sample analyzed, since entropy depends solely on probability distribution, every text of Spanish will exhibit its own value of entropy, thus making it difficult to talk about *the* entropy of Spanish.

Acknowledgment

The author would like to thank Ms. Ana Mengotti, edition board director of the EFE press agency in Bogota (Colombia), for the news archive provided for this research. Also thanks to the anonymous reviewers for their helpful comments.

[Recibido: noviembre de 2011 — Aceptado: septiembre de 2012]

References

Alameda, J. & Cuetos, F. (1995), 'Diccionario de las unidades lingüísticas del castellano, Volumen II: Orden por frecuencias'.
*http://www.uhu.es/jose.alameda

Barnard III, G. (1955), 'Statistical calculation of word entropies for four Western languages', *IEEE Transactions on Information Theory* **1**, 49–53.

Cover, T. & King, R. (1978), 'A convergent gambling estimate of the entropy of English', *IEEE Transactions on Information Theory* **IT-24**(6), 413–421.

Crutchfield, J. & Feldman, D. (2003), 'Regularities unseen, randomness observed: Levels of entropy convergence', *Chaos* **13**(6), 25–54.

Debowski, L. (2011), 'Excess entropy in natural language: Present state and perspectives', *Chaos* **21**(3).

Kaltchenko, A. & Laurier, W. (2004), 'Algorithms for estimating information distance with applications to bioinformatics and linguistics', *Canadian Conference Electrical and Computer Engineering* .

Marchesi, A. (2007), 'Spanish language, science and diplomacy (In Spanish)'. International Congress of the Spanish Language, Cartagena.
*http://corpus.canterbury.ac.nz

Michel, J., Shen, Y. K., Aiden, A., Veres, A., Gray, M., Team, T. G. B., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. & Aiden, E. (2011), 'Quantitative analysis of culture using millions of digitized books', *Science* **331**, 176–182.

Shannon, C. E. (1948), 'A mathematical theory of communication', *Bell System Technical Journal* **27**, 379–423.

Shannon, C. E. (1951), 'Prediction and entropy of printed English', *Bell System Technical Journal* **30**, 47–51.

Slepian, D. (1976), 'On bandwidth', *Proceedings of the IEEE* **34**(3).

Teahan, W. & Cleary, J. (1996), 'The entropy of English using PPM-based models', *Data Compression Conference* pp. 53–62.

Wyner, A., Jacob, Z. & Wyner, A. (1998), 'On the role of pattern matching in information theory', *IEEE Transactions on Information Theory* **44**(6), 2045–2056.

Zipf, G. K. (1965), *The Psycho-Biology of Language: An Introduction to Dynamic Philology, Second Edition*, The MIT Press.

Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 443 a 455

On the Use of Ranked Set Samples in Entropy Based Test of Fit for the Laplace Distribution

Uso de muestras de rango ordenado en una prueba de ajuste basada en entropía para la distribución Laplace

Mahdi Mahdizadeh[a]

Department of Statistics, Hakim Sabzevari University, Sabzevar, Iran

Abstract

Statistical methods based on ranked set sampling (RSS) often lead to marked improvement over analogous methods based on simple random sampling (SRS). Entropy has been influential in the development of measures of fit of parametric models to the data. This article develops goodness-of-fit tests of the Laplace distribution based on sample entropy when data are collected according to some RSS-based schemes. For each design, critical values of the corresponding test statistic are estimated, by means of simulation, for some sample sizes. A Monte Carlo study on the power of the new tests is performed for several alternative distributions and sample sizes in order to compare our proposal with available method in SRS. Simulation results show that RSS and its variations lead to tests giving higher power than the test based on SRS.

Key words**:** Entropy estimation, Goodness-of-fit test, Ranked set sampling.

Resumen

Los métodos estadísticos basados en muestreo de rango ordenado a menudo son una considerable mejora que el muestreo aleatorio simple. La medida de entropía ha sido influencial en el desarrollo de medidas de ajuste de modelos paramétricos. Este artículo propone pruebas de bondad de ajuste de la distribución Laplace basada en la entropía muestral cuando se usan estructuras basadas en muestras de rango ordenado. Para cada diseño, los valores críticos del correspondiente estadístico de prueba son estimados por medio de simulaciones para diferentes tamaños de muestra. Un estudio de Monte Carlo de la potencia de los nuevos tests es implementado para diferentes distribuciones alternas y tamaños de muestra con el fin de comparar el método propuesto con otros disponibles. La simulación muestra que el muestreo de rango ordenado y sus variaciones brindan mayor potencia que los métodos basados en muestreo aleatorio simple.

Palabras clave**:** entropía, muestreo rango ordenado, prueba de bondad de ajuste.

[a] Assistant Professor. E-mail: mahdizadeh.m@live.com, noniid@yahoo.com

1. Introduction

The ranked set sampling (RSS) was introduced by McIntyre (1952) who built on the sample mean to obtain a more precise estimator of the population mean. In this design, the experimenter exploits inexpensive additional information about the characteristic of interest for ranking randomly drawn sampling units and then quantifies a selected subset of them. Auxiliary information may be provided by, for example, visual inspection, concomitant variables, expert opinion, etc., or some combinations of these methods. This flexibility in the choice of ranking mechanism is an appealing feature which makes RSS a cost-efficient sampling technique potentially applicable in fields such as agriculture, biology, ecology, forestry, etc. As an example, consider the following situation mentioned by Takahasi & Wakimoto (1968). Suppose that the quantity of interest is the height of trees in a orchard. While the actual measurement is going to be laborious, a simple glance can help us to rank a handful of trees locating close to each other.

The RSS method can be summarized as follows:

1. Draw k random samples, each of size k, from the target population.
2. Apply judgement ordering, by any cheap method, on the elements of the ith ($i = 1, \ldots, k$) sample and identify the ith smallest unit.
3. Actually measure the k identified units in step 2.
4. Repeat steps 1-3, h times (cycles), if needed, to obtain a ranked set sample of size $n = hk$.

The set of n measured observations are said to constitute the ranked set sample denoted by $\{X_{[i]j} : i = 1, \ldots, k\,; j = 1, \ldots, h\}$, where $X_{[i]j}$ is the ith judgement order statistic from the jth cycle. Current literature on RSS reports many statistical procedures, in both parametric and nonparametric settings, which are superior to their counterparts in simple random sampling (SRS). For an excellent review of most previous works on RSS, see the recent book by Chen, Bai & Sinha (2004). The success of RSS can be traced to the fact that a ranked set sample consists of independent order statistics and contains more information than a simple random sample of the same size, whose ordered values are correlated.

A basic version of RSS has been extensively modified to come up with schemes resulting in more accurate estimators of the population attributes. Multistage ranked set sampling (MSRSS) introduced by Al-Saleh & Al-Omari (2002) is such a variation surpassing RSS. The MSRSS scheme can be described as follows:

1. Randomly identify k^{r+1} units from the population of interest, where r is the number of stages.
2. Allocate the k^{r+1} units randomly into k^{r-1} sets of k^2 units each.
3. For each set in step 2, apply 1-2 of RSS procedure explained above, to get a (judgement) ranked set of size k. This step gives k^{r-1} (judgement) ranked sets, each of size k.

4. Without actual measuring of the ranked sets, apply step 3 on the k^{r-1} ranked set to gain k^{r-2} second stage (judgement) ranked sets, of size k each.

5. Repeat step 3, without any actual measurement, until an rth stage (judgement) ranked set of size k is acquired.

6. Actually measure the k identified units in step 5.

7. Repeat steps 1-6, h times, if needed, to obtain an rth stage ranked set sample of size $n = hk$.

Similarly, the rth stage ranked set sample will be denoted by $\{X^{(r)}_{[i]j} : i = 1, \dots, k;$ $j = 1, \dots, h\}$. It is to be noted that special case of MSRSS with $r = 2$ is known as double ranked set sampling (DRSS) (Al-Saleh & Al-Kadiri 2000). Clearly, the case $r = 1$ corresponds to RSS.

While testing hypotheses on the parameters of the normal, exponential and uniform distributions under RSS and its variations have been widely investigated, little effort has been made for developing a test of fit based on RSS. Stokes & Sager (1988) characterized a ranked set sample as a sample from a conditional distribution, conditioning on a multinomial random vector, and applied RSS to the estimation of the cumulative distribution function. They proposed the Kolmogorov-Smirnov test in RSS setup and derived the null distribution of the test statistic.

Entropy of a distribution was proposed by Shannon (1948) as a measure of uncertainty in information theory. He found that the entropy of the normal distribution is maximum among all distributions with fixed variance. Based on this result, Vasicek (1976) developed a test for normality and, indeed, introduced a new approach for constructing test of fit. Similar tests have been suggested for other distributions based on their entropy characterization results. See Dudewicz & van der Meulen (1981), Gokhale (1983), Grzegorzewski & Wieczorkowski (1999), Mudholkar & Tian (2002), and Choi & Kim (2006).

The classical Laplace distribution introduced by Laplace in 1774 is one of the basic symmetric distributions often used for modeling phenomena with heavier than the normal tails. It has been applied in steam generator inspection, navigation, reliability, generalized linear regression and Bayesian analysis. For more recent applications refer to Kotz, Kozubowski & Podgórski (2001). In this work, we deal with the problem of developing a goodness-of-fit test for the Laplace distribution when the researcher obtains data using RSS and MSRSS. Mahdizadeh & Arghami (2010) suggested similar procedures for the inverse Gaussian law.

The layout of this article is as follows: In Section 2, entropy estimation is extended to RSS and MSRSS, goodness-of-fit tests for the Laplace distribution based on these designs are introduced. Section 3 contains the results of simulation studies carried out to expose the power properties of the new tests. Section 4 is given to the effect of entropy estimator used in the test statistics on power properties. Some brief conclusions are provided in Section 5.

2. Proposed Tests

To put the procedure into perspective, we first review some concepts from information theory. Suppose that a continuous random variable X has distribution function F_X with density function f_X. Shannon's entropy of f_X is given by

$$H(f_X) = -\int_{-\infty}^{\infty} f_X(x) \log f_X(x)\, dx \tag{1}$$

It is easy to show that using the quantile function $F_X^{-1}(u) = \inf\{x : F_X(x) \geq u\}$, (1) can be written as

$$H(f_X) = \int_0^1 \log\big(\frac{d}{du} F_X^{-1}(u)\big)\, du \tag{2}$$

This entropy representation was used by Vasicek (1976) to define the sample entropy in terms of order statistics as follows: Let $X_{(1)}, \ldots, X_{(n)}$ be the ordered values of a random sample of size n from F_X. At each sample point $(X_{(i)}, \frac{i}{n})$, the derivative in (2) is estimated by

$$s_i(m,n) = \frac{X_{(i+m)} - X_{(i-m)}}{2\,m/n} \tag{3}$$

where $m \in \{1, \ldots, \frac{n}{2}\}$ is a window size to be determined. Vasicek's entropy estimator is the mean of logarithm of d_i's defined in the above, i.e.,

$$V_{m,n}(f_X) = \frac{1}{n}\sum_{i=1}^{n} \log\Big(\frac{n}{2m}(X_{(i+m)} - X_{(i-m)})\Big) \tag{4}$$

where $X_{(i-m)} = X_{(1)}$ for $i \leq m$ and $X_{(i+m)} = X_{(n)}$ for $i \geq n-m$.

Since the entropy estimator (4) is based on spacings, we would need ordered values of the ranked set sample to estimate entropy in RSS. Proceeding as in the SRS case, we first pool the units in all cycles and then form the estimator based on the ordered pooled sample. The MSRSS analogue of $V_{m,n}(f_X)$ becomes

$$V_{m,n}^{(r)}(f_X) = \frac{1}{n}\sum_{i=1}^{n} \log\Big(\frac{n}{2m}(X_{(i+m)}^{(r)} - X_{(i-m)}^{(r)})\Big) \tag{5}$$

where $X_{(a)}^{(r)}$ is the ath $(a = 1, \ldots, n)$ order statistic of the rth stage ranked set sample. The reference to subscript k is not made here for conciseness in notation. From now on, we use $V_{m,n}^{(0)}(f_X)$ to denote the estimator (4). So $X_{(a)}^{(0)}$ represents ath order statistic of a simple random sample of size n. In fact, $\{V_{m,n}^{(r)}(.)\}$ is a sequence of entropy estimators indexed by the stage number in MSRSS.

A Monte Carlo experiment was conducted to compare the proposed estimators of entropy when the underlying distribution is the standard Laplace with mean 0 and variance 2. Generation of random samples is easily done based on a result from distribution theory; difference a of two standard exponential random variables has the standard Laplace distribution. Figure 1 displays simulated biases and root mean square errors (RMSEs) of $V_{m,n}^{(r)}$ for $r = 0, 1, 2$ based on 50,000 samples with $n = 10, 20, 30$, and $k = 5$ in MSRSS design (this setup will be used in the rest of the paper). An empty circle is used as the plotting symbol, and points corresponding to SRS, RSS and DRSS are connected by solid, dashed and dotted lines, respectively. It is seen that given a sample size, MSRSS improves entropy estimation with respect to SRS. Moreover, the larger stage number r the smaller absolute value of bias, and RMSE of the corresponding estimator. This property is helpful in distinguishing between the results of different designs when the types of connecting lines are not visible because of compactness in Figure 1.

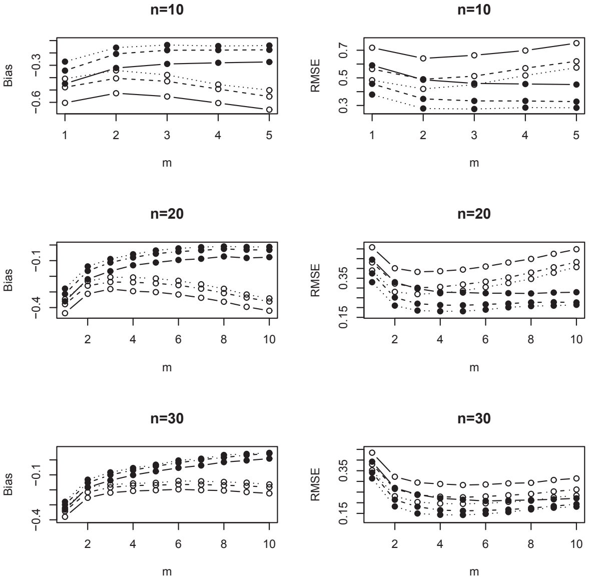


FIGURE 1: Bias and RMSE comparison for the entropy estimators $V_{m,n}$ and $E_{m,n}^1$ for the standard Laplace distribution with $H(f) = 1.6931$.

Choi & Kim (2006) presented an entropy characterization of the Laplace distribution and used the following result (Corollary 2) to establish an entropy based test of fit for the Laplace distribution.

Corollary 1. *(Choi & Kim 2006). Suppose X has a Laplace distribution La(μ,θ) with density function*

$$f_X(x; \mu, \theta) = \frac{1}{2\theta} \exp(-|x - \mu|/\theta) \quad \mu \in R\,,\, \theta > 0$$

Then the entropy of f_X is given by

$$H(f_X) = \log(2\theta) + 1$$

Corollary 2. *(Choi & Kim 2006). Let X be a random variable with density function $f_X(x)$ satisfying the restriction*

$$E_{f_X}(|X|) = \int_{-\infty}^{\infty} |x| \, f_X(x) \, dx \equiv \theta$$

Under this restriction, the distribution of X maximizing Shannon's entropy is La(0,θ).

Consider a random sample $X_1, \dots, X_n$ from a population with density function f and suppose it is of interest to test $H_0 : f_X \in \mathcal{L} = \{La(\mu, \theta) : \mu \in R, \theta > 0\}$ against the general alternative $H_1 : f_X \notin \mathcal{L}$. Choi & Kim (2006) proposed rejecting the null hypothesis if

$$T_{m,n}(g_Y) = \exp\left(V_{m,n}(g_Y)\right) / \,\widehat{\theta} \leq T^*_{m,n,\alpha}(g_Y) \tag{6}$$

where

$$V_{m,n}(g_Y) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{n}{2m} (\widetilde{Y}_{(i+m)} - \widetilde{Y}_{(i-m)}) \right)$$

is the estimate of the entropy of $\widetilde{Y} = X - \mu$ based on $\widetilde{Y}_{(i)} = X_{(i)} - \widehat{\mu}$ $(i = 1, \dots, n)$ with $\widehat{\mu}$ being the median of X_i's, $\widehat{\theta} = \sum_{i=1}^{n} |\widetilde{Y}_i|/n$, and $T^*_{m,n,\alpha}(g_Y)$ is the 100α percentile of the null distribution of $T_{m,n}(g_Y)$.

In order to obtain the percentiles of the null distribution, $T_{m,n}(g_Y)$ was calculated using the estimators $V^{(r)}_{m,n}(g_Y)$ for $r = 0, 1, 2$ based on 50,000 samples of size n generated from the La(0,1) distribution. The values were then used to determine $T^*_{m,n,0.05}(g_Y)$ in different designs and for different sample sizes. To estimate μ and θ in MSRSS, we simply plug the data into the formulae available in SRS. Tables of 0.05 critical points for the tests could be requested from the author. They are not reported here.

To implement the tests, we must first select the window size m associated with a given sample size. In general, there is no unanimous rule to choose the optimal m for each n. Previous studies, however, suggest to of use the window size which leads to the least conservative test. Thus, using the window size giving the largest critical value is advised to achieve higher power. The optimal window size, denoted by m^*, for sample sizes 10, 20 and 30 are approximately 3, 3 and 4, respectively.

3. Simulation Study

In this section, we shall use the Monte Carlo approach to evaluate the entropy tests in terms of power. The distributions considered in the simulation study are as follows: (A) normal(0,1), (B) t(10), (C) logistic(0,1), (D) uniform(0,1), (E) Beta(2,2), (F) chi-square(4), (G) lognormal(0,0.5) and (H) Gamma(1.5,1). We note that (A)-(E) are symmetric and (F)-(H) are asymmetric.

Under each design, 50,000 samples of sizes $n = 10, 20, 30$ were generated from each alternative distribution and the power of the tests were estimated by the fraction of the samples falling into the corresponding critical region. Figures 2-7 depict the estimated power of the tests in which the same plotting symbol and connecting lines of Figure 1 are employed.

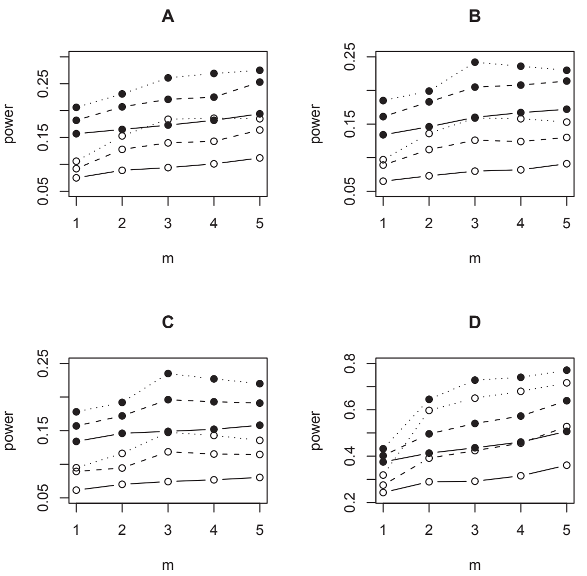


FIGURE 2: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives A-D when $n = 10$.

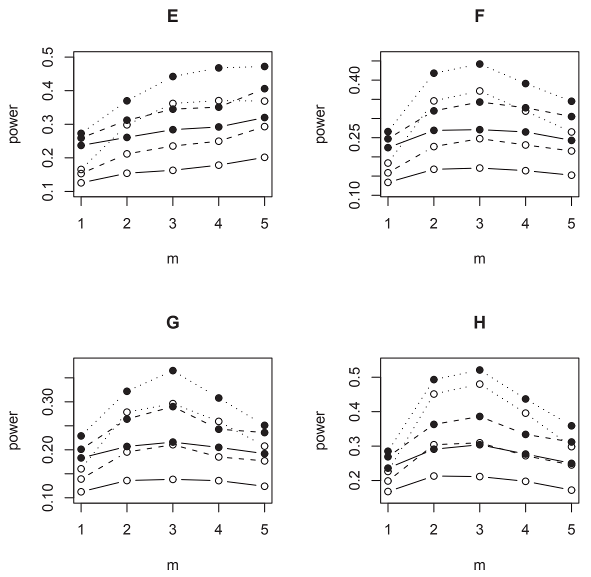


FIGURE 3: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives E-H when $n = 10$.

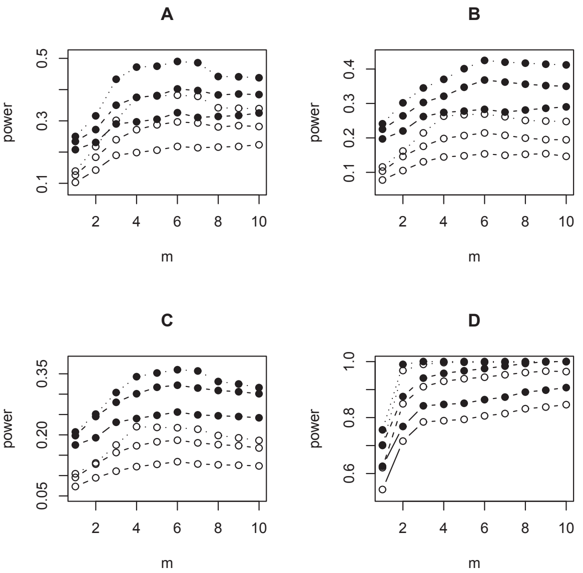


FIGURE 4: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives A-D when $n = 20$.

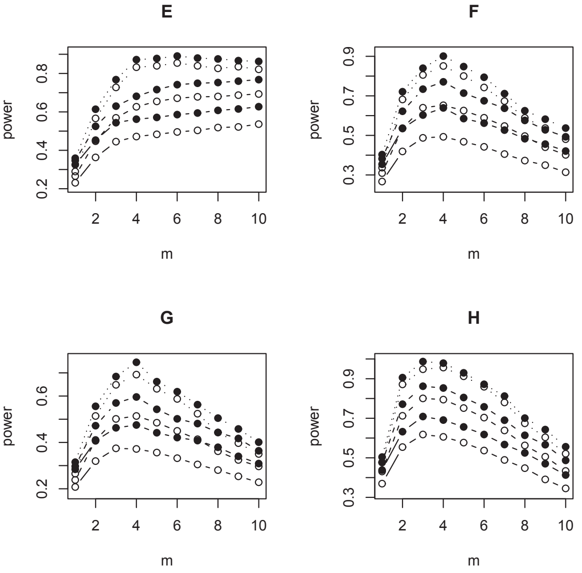


FIGURE 5: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives E-H when $n = 20$.

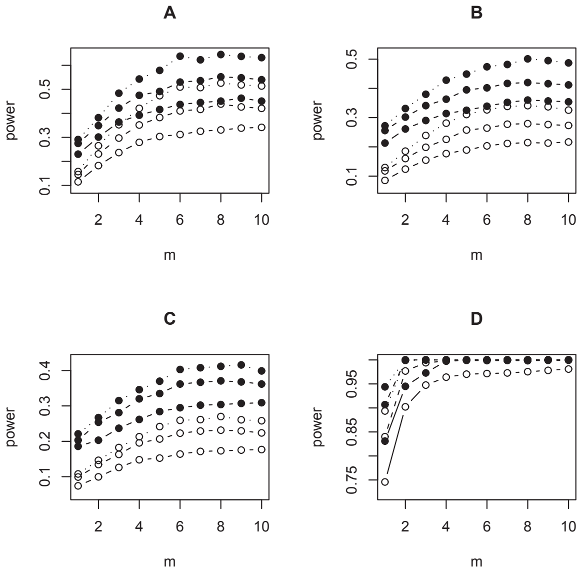


Figure 6: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives A-D when $n = 30$.

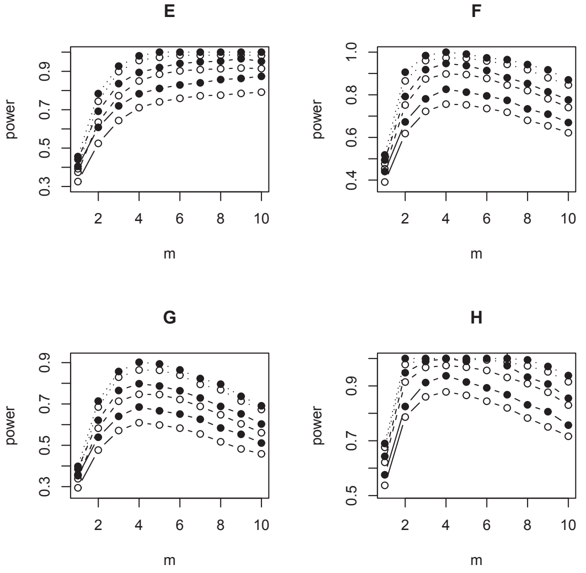


Figure 7: Power comparison for the entropy tests of size 0.05 based on $V_{m,n}$ and $E^1_{m,n}$ against alternatives E-H when $n = 30$.

It is observed that given a sample size, the entropy tests based on RSS and DRSS are more powerful than that based on SRS regardless of the alternative distribution. More interestingly, the higher sampling effort the more powerful resulting test would be. That is DRSS has the best performance among three considered designs. Remember that a similar trait was reported earlier in the context of entropy estimation. This is fairly expected because the test statistic in each design is constructed based on the corresponding entropy estimator. It should be mentioned that against asymmetric distributions and for each n, maximum power is gained at optimal m or at one of its neighboring values. This trend, however, does not hold for symmetric distributions where maximum power occurs in $m \approx \frac{n}{2}$. Since the best m associated with a sample size varies according to the alternative, we may use a data histogram to decide on the best window size for applying the tests.

It is interesting to examine whether a further increase in power is possible by increasing the number of stages in MSRSS. To this end, testing procedures under MSRSS with $r = 3, 4$ were developed. Figure 8 displays the power of the tests, where alternatives A-H are denoted by integers 1-8 on the X axis, and points corresponding to $r = 2, 3, 4$ are connected by solid, dashed and dotted lines, respectively. Results of DRSS design were included to facilitate comparison. For a given n, the result are provided only for optimal m, not for all $m \in \{1, \dots, \frac{n}{2}\}$, to save space. From Figure 8, we can see as r increases, some improvement in power happens. Since the differences in results for $r = 2$ and $r = 3, 4$ are not marked for (A)-(C), we may confine ourselves to DRSS against these alternatives.

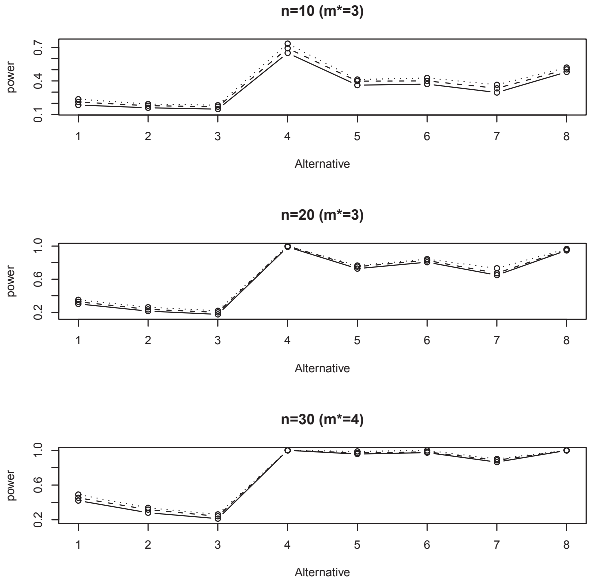


Figure 8: Power comparison for the entropy tests of size 0.05 against alternatives A-H under MSRSS designs.

4. Effect of Entropy Estimator

As mentioned before, Vasicek's estimator has been widely used for developing entropy based test of fit. Many authors have modified this test to come up with more efficient estimators. In this section, power behavior of the tests employing such estimators are investigated. To this end, we consider two entropy estimators proposed by Ebrahimi, Pflughoeft & Soofi (1994).

The first estimator which modifies the denominator of (3) is defined as follows

$$E^1_{m,n} = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{X_{(i+m)} - X_{(i-m)}}{c_i m/n}\right) \tag{7}$$

where

$$c_i = \begin{cases} 1 + \frac{i-1}{m} & 1 \le i \le m, \\ 2 & m+1 \le i \le n-m, \\ 1 + \frac{n-i}{m} & n-m+1 \le i \le n \end{cases}$$

The second estimator, obtained by modifying both the numerator and denominator of (3), is given by

$$E^2_{m,n} = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{Z_{(i+m)} - Z_{(i-m)}}{d_i m/n}\right) \tag{8}$$

where

$$d_i = \begin{cases} 1 + \frac{i+1}{m} - \frac{i}{m^2} & 1 \le i \le m, \\ 2 & m+1 \le i \le n-m-1, \\ 1 + \frac{n-i}{m+1} & n-m \le i \le n, \end{cases}$$

the $Z_{(i)}$'s are

$$Z_{(i)} = \begin{cases} a + \frac{i-1}{m}(X_{(1)} - a) & 1 \le i \le m, \\ X_{(i)} & m+1 \le i \le n-m-1, \\ b - \frac{n-i}{m}(b - X_{(n)}) & n-m \le i \le n, \end{cases}$$

and a and b are constants to be determined such that $P(a \le X \le b) \approx 1$. For example, when F has a bounded support, a and b are lower and upper bound, respectively (for uniform(0,1) distribution, $a = 0$ and $b = 1$); if F is bounded below (above), then $a(b)$ is lower (upper) support, $a = \overline{x} - ks$ $(b = \overline{x} + ks)$, where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

and k is a suitable number say 3 to 5 (for exponential distribution, $a = 0$ and $b = \overline{x} + ks$); in the case that F has no bound on its support, a and b may be chosen as $a = \overline{x} - ks$ and $b = \overline{x} + ks$.

Simulation results show that both estimators have less bias and less RMSE than Vasicek's estimator (uniformly). Since $E^1_{m,n}$ has simpler form, we focus on that in the sequel. Simulated biases and RMSEs of $E^{1(r)}_{m,n}$ (The MSRSS analogue of $E^{1}_{m,n}$) for $r = 0, 1, 2$ are given in Figure 1, where a filled circle is used as the plotting symbol, and points corresponding to SRS, RSS and DRSS are connected by solid, dashed and dotted lines, respectively. Again, it is evident that as r increases, $E^{1(r)}_{m,n}$ becomes more efficient. Also, the estimated power of the tests developed using $E^{1(r)}_{m,n}$ for $r = 0, 1, 2$ appear in Figures 2-7 with the same display conventions used for bias and RMSE of the corresponding entropy estimator. In each design, tests based on the new estimator is more powerful than those based on the original estimator for all sample sizes and alternatives.

5. Conclusion

The aim of this paper was to develop goodness-of-fit tests for the Laplace distribution under RSS and MSRSS designs. Motivated by the entropy based test of fit in SRS, we employed the sample entropy based on aforesaid designs to construct the corresponding tests of fit. An extensive simulation study was conducted to provide insight into the finite sample power behavior of the proposed tests. The results indicate that using (multistage) ranked set samples in entropy based test of fit for the Laplace distribution result in higher power as compared with simple random samples. We have developed analogous tests for the uniform, normal, exponential, Weibull and some other distributions using improved entropy estimators whose results will be reported in future articles. Tables of critical points and power of the tests in different designs along with the corresponding computer codes are available on request from the author.

Acknowledgements

We thank the reviewers and the Editor for comments which led to presentational improvements.

[Recibido: noviembre de 2010 — Aceptado: octubre de 2012]

References

Al-Saleh, M. F. & Al-Kadiri, M. (2000), ‘Double ranked set sampling’, *Statistics & Probability Letters* **48**, 205–212.

Al-Saleh, M. F. & Al-Omari, A. I. (2002), ‘Multistage ranked set sampling’, *Journal of Statistical Planning and Inference* **102**, 273–286.

Chen, Z., Bai, Z. & Sinha, B. K. (2004), *Ranked set sampling: Theory and Applications*, Springer, New York.

Choi, B. & Kim, K. (2006), ‘Testing goodness-of-fit for laplace distribution based on maximum entropy’, *Statistics* **40**, 517–531.

Dudewicz, E. J. & van der Meulen, E. C. (1981), 'Entropy-based tests of uniformity', *Journal of the American Statistical Association* **76**, 967–974.

Ebrahimi, N., Pflughoeft, K. & Soofi, E. S. (1994), 'Two measures of sample entropy', *Statistics & Probability Letters* **20**, 225–234.

Gokhale, D. V. (1983), 'On the entropy-based goodness-of-fit tests', *Computational Statistics & Data Analysis* **1**, 157–165.

Grzegorzewski, P. & Wieczorkowski, R. (1999), 'Entropy based goodness-of-fit test for exponentiality', *Communications in Statistics: Theory and Methods* **28**, 1183–1202.

Kotz, S., Kozubowski, T. J. & Podgórski, K. (2001), *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*, Birkhäuser, Boston, USA.

Mahdizadeh, M. & Arghami, N. R. (2010), 'Efficiency of ranked set sampling in entropy estimation and goodness-of-fit testing for the inverse gaussian law', *Journal of Statistical Computation and Simulation* **80**, 761–774.

McIntyre, G. A. (1952), 'A method of unbiased selective sampling using ranked sets', *Australian Journal of Agricultural Research* **3**, 385–390.

Mudholkar, G. S. & Tian, L. (2002), 'An entropy characterization of the inverse gaussian distribution and related goodness-of-fit test', *Journal of Statistical Planning and Inference* **102**, 211–221.

Shannon, C. E. (1948), 'A mathematical theory of communications', *Bell System Technical Journal* **27**, 379–423.

Stokes, S. L. & Sager, T. W. (1988), 'Characterization of a ranked-set sample with application to estimating distribution function', *Journal of the American Statistical Association* **83**, 374–381.

Takahasi, K. & Wakimoto, K. (1968), 'On unbiased estimates of the population mean based on the sample stratified by means of ordering', *Annals of the Institute of Statistical Mathematics* **21**, 249–255.

Vasicek, O. (1976), 'A test of normality based on sample entropy', *Journal of the Royal Statistical Society Series B* **38**, 54–59.

Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 457 a 478

An Empirical Comparison of EM Initialization Methods and Model Choice Criteria for Mixtures of Skew-Normal Distributions

Una comparación empírica de algunos métodos de inicialización EM y criterios de selección de modelos para mezclas de distribuciones normales asimetricas

José R. Pereira[a], Leyne A. Marques[b], José M. da Costa[c]

Departamento de Estatística, Instituto de Ciências Exatas, Universidade Federal do Amazonas, Manaus, Brasil

Abstract

We investigate, via simulation study, the performance of the EM algorithm for maximum likelihood estimation in finite mixtures of skew-normal distributions with component specific parameters. The study takes into account the initialization method, the number of iterations needed to attain a fixed stopping rule and the ability of some classical model choice criteria to estimate the correct number of mixture components. The results show that the algorithm produces quite reasonable estimates when using the method of moments to obtain the starting points and that, combining them with the AIC, BIC, ICL or EDC criteria, represents a good alternative to estimate the number of components of the mixture. Exceptions occur in the estimation of the skewness parameters, notably when the sample size is relatively small, and in some classical problematic cases, as when the mixture components are poorly separated.

Key words: EM algorithm, Mixture of distributions, Skewed distributions.

Resumen

El presente artículo muestra un estudio de simulación que evalúa el desempeño del algoritmo EM utilizado para determinar estimaciones por máxima verosimilitud de los parámetros de la mezcla finita de distribuciones normales asimétricas. Diferentes métodos de inicialización, así como el número de interacciones necesarias para establecer una regla de parada especificada y algunos criterios de selección del modelo para permitir estimar el número

[a] Associate professor. E-mail: jrpereira@ufam.edu.br
[b] Assistant professor. E-mail: leyneabuim@gmail.com
[c] Assistant professor. E-mail: zemirufam@gmail.com

apropiado de componentes de la mezcla han sido considerados. Los resultados indican que el algoritmo genera estimaciones razonables cuando los valores iniciales son obtenidos mediante el método de momentos, que junto con los criterios AIC, BIC, ICL o EDC constituyen una eficaz alternativa en la estimación del número de componentes de la mezcla. Resultados insatisfactorios se verificaron al estimar los parámetros de simetría, principalmente seleccionando un tamaño pequeño para la muestra, y en los casos conocidamente problemáticos en los cuales los componentes de la mezcla están suficientemente separados.

***Palabras clave*:** algoritmo EM, distribuciones asimétricas, mezcla de distribuciones.

1. Introduction

Finite mixtures have been widely used as a powerful tool to model heterogeneous data and to approximate complicated probability densities, presenting multimodality, skewness and heavy tails. These models have been applied in several areas like genetics, image processing, medicine and economics. For comprehensive surveys, see McLachlan & Peel (2000) and Frühwirth-Schnatter (2006).

Maximum likelihood estimation in finite mixtures is a research area with several challenging aspects. There are nontrivial issues, such as lack of identifiability and saddle regions surrounding the possible local maxima of the likelihood. Another problem is that the likelihood is possibly unbounded, which happens when the components are normal densities.

There is a lot of literature involving mixtures of normal distributions, some references can be found in the above-mentioned books. In this work we consider mixtures of *skew-normal (SN) distributions*, as defined by Azzalini (1985). This distribution is an extension of the normal distribution that accommodates asymmetry.

The standard algorithm for maximum likelihood estimation in finite mixtures is the *Expectation Maximization* (EM) of Dempster, Laird & Rubin (1977), see also McLachlan & Krishnan (2008) and Ho, Pyne & Lin (2012). It is well known that it has slow convergence and that its performance is strongly dependent on the stopping rule and starting points. For normal mixtures, several authors have computationally investigated the performance of the EM algorithm by taking into account initial values (Karlis & Xekalaki (2003); Biernacki, Celeux & Govaert (2003)), asymptotic properties (Nityasuddhi & Böhning 2003) and comparisons of the standard EM with other algorithms (Dias & Wedel 2004).

Although there are some purposes to overcome the unboundedness problem in the normal mixture case, involving constrained optimization and alternative algorithms (see Hathaway (1985), Ingrassia (2004), and Yao (2010)), it is interesting to investigate the performance of the (unrestricted) EM algorithm in the presence of skewness in the component distributions, since algorithms of this kind have been presented in recent works as Lin, Lee & Hsieh (2007), Lin, Lee & Yen (2007), Lin

(2009), Lin (2010) and Lin & Lin (2010). Here, we employ the algorithm presented in Basso, Lachos, Cabral & Ghosh (2010).

The goal of this work is to study the performance of the estimates produced by the EM algorithm, taking into account the method of moments and a random initialization method to obtain initial values, the number of iterations needed to attain a fixed stopping rule and the ability of some classical model choice criteria (AIC, BIC, ICL and EDC) to estimate the correct number of mixture components. We also investigated the density estimation issue by analyzing the estimates of the log-likelihood function at the true values of the parameters. The work is restricted to the univariate case.

The rest of the paper is organized as follows. In Sections 2 and 3, for the sake of completeness, we give a brief sketch of the skew-normal mixture model and of estimation via the EM algorithm, respectively. In Section 4, the simulation study about the initialization methods, the number of iterations and density estimation are presented. The study concerning model choice criteria is presented in Section 5. Finally, in Section 6 the conclusions of our study are draw and additional comments are given.

2. The Finite Mixture of SN Distributions Model

2.1. The Skew-Normal (SN) Distribution

The skew-normal distribution, introduced by (Azzalini 1985), is given by the density

$$\mathrm{SN}(y|\mu,\sigma^2,\lambda) = 2\mathrm{N}(y|\mu,\sigma^2)\Phi\left(\lambda\frac{y-\mu}{\sigma}\right)$$

where $\mathrm{N}(\cdot|\mu,\sigma^2)$ denotes the univariate normal density with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ and $\Phi(\cdot)$ is the distribution function of the standard normal distribution. In this definition, μ, $\lambda \in \mathbb{R}$ and σ^2 are parameters regulating location, skewness and scale, respectively. For a random variable Y with this distribution, we use the notation $Y \sim \mathrm{SN}(\mu,\sigma^2,\lambda)$.

To simulate realizations of Y and to implement the EM-type algorithm a convenient stochastic representation is given by

$$Y = \mu + \sigma\delta T + \sigma(1-\delta^2)^{1/2}T_1 \tag{1}$$

where $\delta = \lambda/\sqrt{1+\lambda^2}$, $T = |T_0|$, T_0 and T_1 are independent standard normal random variables and $|\cdot|$ denotes absolute value. (for proof see Henze (1986)). To reduce computational difficulties related to the implementation of the algorithms used for estimation, we use the parametrization

$$\Gamma = (1-\delta^2)\sigma^2 \qquad \text{and} \qquad \Delta = \sigma\delta$$

which was first suggested by Bayes & Branco (2007). Note that $(\lambda,\sigma^2) \to (\Delta,\Gamma)$ is a one to one mapping. To recover λ and σ^2, we use

$$\lambda = \Delta/\sqrt{\Gamma} \quad \text{and} \quad \sigma^2 = \Delta^2 + \Gamma$$

Then, it follows easily from (1) that

$$Y|T = t \sim N(\mu + \Delta t, \Gamma) \quad \text{and} \quad T \sim HN(0,1) \tag{2}$$

where $HN(0,1)$ denotes the half-normal distribution with parameters 0 and 1.

The expectation, variance and skewness coefficient of $Y \sim \mathrm{SN}(\mu, \sigma^2, \lambda)$ are respectively given by

$$E(Y) = \mu + \sigma\Delta\sqrt{2/\pi}, \quad Var[Y] = \sigma^2\left(1 - \frac{2}{\pi}\delta^2\right), \quad \gamma(Y) = \frac{\kappa\delta^3}{(1 - \frac{2}{\pi}\delta^2)^{3/2}} \tag{3}$$

where $\kappa = \frac{4-\pi}{2}(\frac{2}{\pi})^{3/2}$ (see Azzalini (2005, Lemma 2)).

2.2. Finite Mixture of SN Distributions

The finite mixture of SN distributions model, hereafter FM-SN model, is defined by considering a random sample $\mathbf{y} = (y_1, \ldots, y_n)^\top$ from a mixture of SN densities given by

$$g(y_j|\Theta) = \sum_{i=1}^{k} p_i \mathrm{SN}(y_j|\theta_i), \qquad j = 1, \ldots, n \tag{4}$$

where $p_i \geq 0$, $i = 1, \ldots, k$ are the mixing probabilities, $\sum_{i=1}^{k} p_i = 1$, $\theta_i = (\mu_i, \sigma_i^2, \lambda_i)^\top$ is the specific vector of parameters for the component i and $\Theta = ((p_1, \ldots, p_k)^\top, \theta_1^\top, \ldots, \theta_k^\top)^\top$ is the vector with all parameters.

For each j consider a latent classification random variable Z_j taking values in $\{1, \ldots, k\}$, such that

$$y_j|Z_j = i \sim \mathrm{SN}(\theta_i), \quad P(Z_j = i) = p_i, \quad i = 1, \ldots, k; \quad j = 1, \ldots, n.$$

Then it is straightforward to prove, integrating out Z_j, that y_j has density (4). If we combine this result with (2), we have the following stochastic representation for the FM-SN model

$$\begin{aligned} &y_j|T_j = t_j, Z_j = i \sim \mathrm{N}(\mu_i + \Delta_i t_j, \Gamma_i), \\ &T_j \sim \mathrm{HN}(0,1), \\ &P(Z_j = i) = p_i, \quad i = 1, \ldots, k; \quad j = 1, \ldots, n \end{aligned}$$

where

$$\Gamma_i = (1 - \delta_i^2)\sigma_i^2, \qquad \Delta_i = \sigma_i\delta_i, \qquad \delta_i = \lambda_i/\sqrt{1 + \lambda_i^2}, \qquad i = 1, \ldots, k \tag{5}$$

More details can be found in Basso et al. (2010) and references herein.

3. Estimation

3.1. An EM-type Algorithm

In this section we present an EM-type algorithm for estimation of the parameters of a FM-SN distribution. This algorithm was presented before in Basso et al. (2010) and we emphasize that, in order to do this, the representation (5) is crucial. The estimates are obtained using a faster extension of EM called the *Expectation-Conditional Maximization* (ECM) algorithm (Meng & Rubin 1993). When applying it to the FM-SN model, we obtain a simple set of closed form expressions to update a current estimate of the vector Θ, as we will see below. It is important to emphasize that this procedure differs from the algorithm presented by Lin, Lee & Yen (2007), because in the former case the updating equations for the component skewness parameter have a closed form. In what follows we consider the parametrization (5), and still use Θ to denote the vector with all parameters.

Let $\hat{\Theta}^{(m)} = ((\hat{p}_1^{(m)}, \dots, \hat{p}_k^{(m)})^\top, (\hat{\theta}_1^{(m)})^\top, \dots, (\hat{\theta}_k^{(m)})^\top)^\top$ be the current estimate (at the mth iteration of the algorithm) of Θ, where $\hat{\theta}_i^{(m)} = (\hat{\mu}_i^{(m)}, \hat{\Delta}_i^{(m)}, \hat{\Gamma}_i^{(m)})^\top$. The E-step of the algorithm is to evaluate the expected value of the complete data function, known as the *Q−function* and defined as

$$Q(\Theta|\hat{\Theta}^{(m)}) = E[\ell_c(\Theta)|\mathbf{y}, \hat{\Theta}^{(m)}]$$

where $\ell_c(\Theta)$ is the *complete-data log-likelihood function*, given by

$$\ell_c(\Theta) = c + \sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij}\left(\log p_i - \frac{1}{2}\log\Gamma_i - \frac{1}{2\Gamma_i}(y_j - \mu_i - \Delta_i t_j)^2\right)$$

where z_{ij} is the indicator function of the set $(Z_j = i)$ and c is a constant that is independent of Θ. The M-step consists in maximizing the Q-function over Θ. As the M-step turns out to be analytically intractable, we use, alternatively, the ECM algorithm, which is an extension that essentially replaces it with a sequence of conditional maximization (CM) steps. The following scheme is used to obtain an updated value $\hat{\Theta}^{(m+1)}$. We can find more details about the conditional expectations involved in the computation of the Q-function and the related maximization steps in Basso et al. (2010). Here, ϕ denotes the standard normal density and we employ the following notations

$$\hat{z}_{ij} = E[Z_{ij}|y_j; \hat{\Theta}], \quad \hat{s}_{1ij} = E[Z_{ij}T_j|y_j; \hat{\Theta}] \quad \text{and} \quad \hat{s}_{2ij} = E[Z_{ij}T_j^2|y_j; \hat{\Theta}]$$

E-step: Given a current estimate $\hat{\Theta}^{(m)}$, compute $\hat{z}_{ij}$, $\hat{s}_{1ij}$ and $\hat{s}_{2ij}$, for $j = 1, \dots, n$ and $i = 1, \dots, k$, where:

$$\begin{aligned}
\hat{z}_{ij}^{(m)} &= \frac{\hat{p}_i^{(m)} \mathrm{SN}(y_j|\hat{\theta}_i^{(m)})}{\sum_{i=1}^{k} \hat{p}_i^{(m)} \mathrm{SN}(y_j|\hat{\theta}_i^{(m)})} \qquad (6)\\
\hat{s}_{1ij}^{(m)} &= \hat{z}_{ij}^{(m)} \left[\hat{\mu}_{T_{ij}}^{(m)} + \frac{\phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)}{\Phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)} \hat{\sigma}_{T_i}^{(m)} \right]\\
\hat{s}_{2ij}^{(m)} &= \hat{z}_{ij}^{(m)} \left[(\hat{\mu}_{T_{ij}}^{(m)})^2 + (\hat{\sigma}_{T_i}^{(m)})^2 + \frac{\phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)}{\Phi\left(\hat{\mu}_{T_{ij}}^{(m)}/\hat{\sigma}_{T_i}^{(m)}\right)} \, \hat{\mu}_{T_{ij}}^{(m)} \, \hat{\sigma}_{T_i}^{(m)} \right]\\
\hat{\mu}_{T_{ij}}^{(m)} &= \frac{\hat{\Delta}_i^{(m)}}{\hat{\Gamma}_i^{(m)} + (\hat{\Delta}_i^{(m)})^2} (y_j - \hat{\mu}_i^{(m)}),\\
\hat{\sigma}_{T_i}^{(m)} &= \left(\frac{\hat{\Gamma}_i^{(m)}}{\hat{\Gamma}_i^{(m)} + (\hat{\Delta}_i^{(m)})^2} \right)^{1/2}
\end{aligned}$$

CM-steps: Update $\hat{\Theta}^{(m)}$ by maximizing $Q(\Theta|\hat{\Theta}^{(m)})$ over Θ, which leads to the following closed form expressions:

$$\begin{aligned}
\hat{p}_i^{(m+1)} &= n^{-1} \sum_{j=1}^{n} \hat{z}_{ij}^{(m)}\\
\hat{\mu_i}^{(m+1)} &= \frac{\sum_{j=1}^{n} (y_j \hat{z}_{ij}^{(m)} - \hat{\Delta}_i^{(m)} \hat{s}_{1ij}^{(m)})}{\sum_{j=1}^{n} \hat{z}_{ij}^{(m)}}\\
\hat{\Gamma}_i^{(m+1)} &= \frac{\sum_{j=1}^{n} (\hat{z}_{ij}^{(m)} (y_j - \hat{\mu}_i^{(m+1)})^2 - 2(y_j - \hat{\mu}_i^{(m+1)}) \hat{\Delta}_i^{(m)} \hat{s}_{1ij}^{(m)} + (\hat{\Delta}_i^{(m)})^2 \hat{s}_{2ij}^{(m)})}{\sum_{j=1}^{n} \hat{z}_{ij}^{(m)}}\\
\hat{\Delta}_i^{(m+1)} &= \frac{\sum_{j=1}^{n} (y_j - \hat{\mu}_i^{(m+1)}) \hat{s}_{1ij}^{(m)}}{\sum_{j=1}^{n} \hat{s}_{2ij}^{(m)}}
\end{aligned}$$

The algorithm iterates between the E and CM steps until a suitable convergence rule is satisfied and several rules are proposed in the literature (see e.g., McLachlan & Krishnan (2008)). In this work our rule is to stop the process at stage m when $|\ell(\hat{\Theta}^{(m+1)})/\ell(\hat{\Theta}^{(m)}) - 1|$ is small enough.

3.2. Some Problems with Estimation in Finite Mixtures

It is well known that the likelihood of normal mixtures can be unbounded (see e.g., Frühwirth-Schnatter 2006, Chapter 6) and it is not difficult to verify that the FM-SN models also have this feature. One way to circumvent the unboundedness problem is the constrained optimization of the likelihood, imposing conditions on the component variances in order to obtain global maximization (see e.g., Hathaway 1985, Ingrassia 2004, Ingrassia & Rocci 2007, Greselin & Ingrassia 2010). Thus, following Nityasuddhi & Böhning (2003), we investigate only the performance of the EM algorithm when considered component specific parameters (that is, unrestricted) of the mixture and we mention the estimates produced by the algorithm of section 3.1 as "EM estimates", that is, some sort of solution of the score equation, instead of "maximum likelihood estimates".

Another nontrivial issue is the lack of identifiability. Strictly speaking, finite mixtures are always non-identifiable because an arbitrary permutation of the labels of the component parameters lead to the same finite mixture distribution. In the finite mixture context, a more flexible concept of identifiability is used (see, Titterington, Smith & Makov 1985, Chapter 3 for details). The normal mixture model identifiability was first verified by Yakowitz & Spragins (1968), but it is interesting to note that subsequent discussions in the related literature concerning mixtures of Student-t distributions (see e.g., Peel & McLachlan 2000, Shoham 2002, Shoham, Fellows & Normann 2003, Lin, Lee & Ni 2004) do not present a formal proof of its identifiability. It is important to mention that the non-identifiability problem is not a major one if we are interested only in the likelihood values, which are robust to label switching. This is the case, for example, when density estimation is the main goal.

4. A Simulation Study of Initial Values

4.1. Description of the Experiment

It is well known that the performance of the EM algorithm is strongly dependent on the choice of the criterion of convergence and starting points. In this work we do not consider the stopping rule issue, we adopt a fixed rule to stop the process at stage m when

$$\left| \frac{\ell(\hat{\Theta}^{(m+1)})}{\ell(\hat{\Theta}^{(m)})} - 1 \right| < 10^{-6}$$

because we believe that this tolerance for the change in $\ell(\hat{\Theta})$ is quite reasonable in the applications where the primary interest is on the sequence of the log-likelihood values rather than the sequence of parameter estimates (McLachlan & Peel 2000, Section 2.11).

In the mixture context, the choice of starting values for the EM algorithm is crucial because, as noted by Dias & Wedel (2004), there are various saddle regions surrounding the possible local maximum of the likelihood function, and the EM algorithm can be trapped in some of these subsets of the parameter space.

In this work, we make a simulation study in order to compare some methods to obtain starting points for the algorithm proposed in section 3.1, where an interesting question is to investigate the performance of the EM algorithm with respect to the skewness parameter estimation for each component density in the FM-SN model. We consider the following methods to obtain initial values:

The Random Values Method (RVM): we first divide the generated random sample into k sub-samples employing the *k-means* method. The initialization of *k-means* algorithm is random, being recommended to adopt many different choices and we employ five random initializations (see Hastie, Tibshirani & Friedman 2009, Section 14.3). Let φ_i be the sub-sample i. Consider the following points artificially generated from uniform distributions over the specified intervals

$$\begin{aligned} \hat{\xi}_i^{(0)} &\sim U(\min\{\varphi_i\}, \max\{\varphi_i\}) \\ \hat{\omega}_i^{(0)} &\sim U(0, var\{\varphi_i\})\ , \\ \hat{\gamma}_i^{(0)} &\sim sgn(sc\{\varphi_i\}) \times |U(-0.9953, 0.9953)| \end{aligned} \tag{7}$$

where $\min\{\varphi_i\}$, $\max\{\varphi_i\}$, $var\{\varphi_i\}$ and $sc\{\varphi_i\}$ denote, respectively, the minimum, the maximum, the sample variance and the sample skewness coefficient of φ_i, $i = 1, .., k$, also $|\cdot|$ denotes absolute value. These quantities are taken as rough estimates for the mean, variance and skewness coefficient associated to subpopulation i, respectively. The suggested form for $\hat{\gamma}_i^{(0)}$ is due to the fact that the range for the skewness coefficient in SN models is $(-0.9953, 0.9953)$ and to maintain the sign of the sample skewness coefficient.

The starting points for the specific component locations, scale and skewness parameters are given respectively by

$$\begin{aligned} \hat{\mu}_i^{(0)} &= \hat{\xi}_i^{(0)} - \sqrt{2/\pi}\,\delta_{(\hat{\lambda}_i^{(0)})}\hat{\sigma}_i^{(0)} \\ \hat{\sigma}_i^{(0)} &= \sqrt{\frac{\hat{\omega}_i^{(0)}}{1 - \frac{2}{\pi}\delta^2_{(\hat{\lambda}_i^{(0)})}}} \\ \hat{\lambda}_i^{(0)} &= \pm\sqrt{\frac{\pi(\hat{\gamma}_i^{(0)})^{2/3}}{2^{1/3}(4-\pi)^{2/3} - (\pi-2)(\hat{\gamma}_i^{(0)})^{2/3}}} \end{aligned} \tag{8}$$

where $\delta_{(\hat{\lambda}_i^{(0)})} = \hat{\lambda}_i^{(0)}/\sqrt{1 + (\hat{\lambda}_i^{(0)})^2}$, $i = 1, .., k$ and the sign of $\hat{\lambda}_i^{(0)}$ is the same of $\hat{\gamma}_i^{(0)}$. They are obtained by replacing $E(Y)$, $Var(Y)$ and $\gamma(Y)$ in (3) with their respective estimators in (7) and solving the resulting equations in μ_i, σ_i and λ_i. The initial values for the weights p_i are obtained as

$$(\hat{p}_1^{(0)}, \dots, \hat{p}_k^{(0)}) \sim \text{Dirichlet}(1, \dots, 1)$$

a Dirichlet distribution with all parameters equal to 1, namely, a uniform distribution over the unit simplex $\Big\{(p_1, \dots, p_k);\ p_i \geq 0,\ \sum_{i=1}^k p_i = 1\Big\}$.

Method of Moments (MM): the initial values are obtained using equations (8), but replacing $\hat{\xi}_i^{(0)}$, $\hat{\omega}_i^{(0)}$ and $\hat{\gamma}_i^{(0)}$ with the mean, variance and skewness coefficient of sub-sample i, $i = 1, ..k$, with the k sub-samples obtained by the *k-means* method with five random initializations . Let n be the sample size and n_i be the size of sub-sample i. The initial values for the weights are given by

$$\hat{p}_i^{(0)} = \frac{n_i}{n}, \qquad i = 1, .., k.$$

We generated samples from the FM-SN model with $k = 2$ and $k = 3$ components, with sizes fixed as $n = 500$; 1000; 5000 and 1,0000. In addition, we consider different degree of heterogeneity of the components, for $k = 2$ the "moderately separated" ($2MS$), "well separated" ($2WS$) and "poorly separated" ($2PS$) cases and for $k = 3$ the "two poorly separated and one well separated" ($3PWS$) and the "three well separated" ($3WWS$) cases. These degrees of heterogeneity were obtained informally, based on the location parameter values and the reason to consider them as an factor to our study is that the convergence of the EM algorithm is typically affected when the components overlap largely (see Park & Ozeki (2009) and the references herein). In Table 1 the parameters values used in the study are presented and the figures 1 and 2 show some histograms exemplifying these degrees of heterogeneity.

Table 1: Parameters values for FM-SN models

Case	p_1	μ_1	σ_1^2	λ_1	p_2	μ_2	σ_2^2	λ_2	p_3	μ_3	σ_3^2	λ_3
$2MS$	0.6	5	9	6	0.4	20	16	-4				
$2WS$	0.6	5	9	6	0.4	40	16	-4				
$2PS$	0.6	5	9	6	0.4	15	16	-4				
$3PWS$	0.4	5	9	6	0.3	20	16	-4	0.3	28	16	4
$3WWS$	0.4	5	9	6	0.3	30	16	-4	0.3	38	16	4

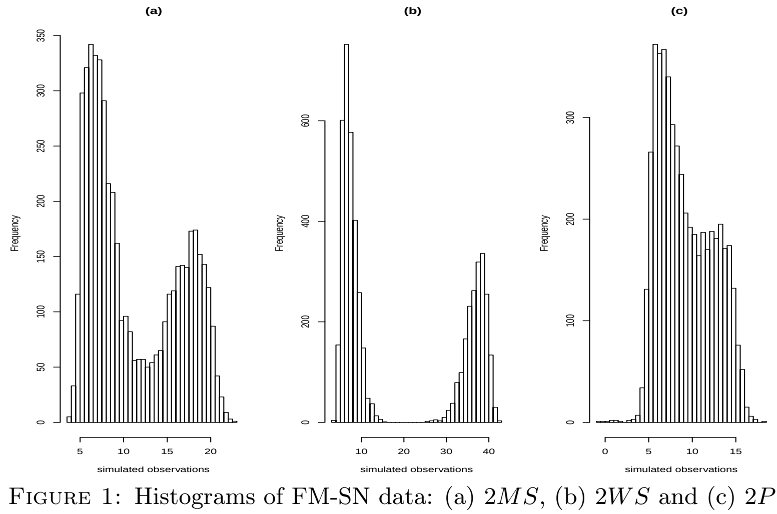


Figure 1: Histograms of FM-SN data: (a) $2MS$, (b) $2WS$ and (c) $2PS$.

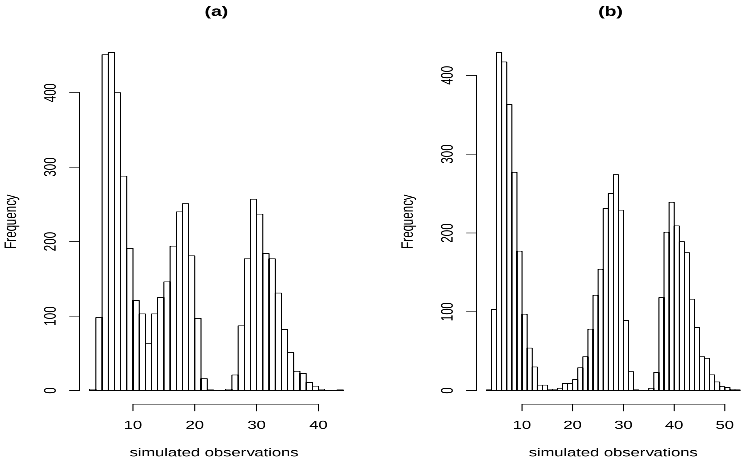


Figure 2: Histograms of FM-SN data: (a) $3PWS$ and (b) $3WWS$.

For each combination of parameters and sample size, samples from the FM-SN model were artificially generated and we obtained estimates of the parameters using the algorithm presented in section 3.1 initialized by each method proposed. This procedure was repeated 5,000 times and we computed the bias and mean squared error (MSE) over all samples, which for μ_i are defined as

$$\text{bias} = \frac{1}{5,000}\sum_{j=1}^{5,000}\hat{\mu}_i^{(j)} - \mu_i \quad \text{and} \quad \text{MSE} = \frac{1}{5,000}\sum_{j=1}^{5,000}(\hat{\mu}_i^{(j)} - \mu_i)^2,$$

respectively, where $\hat{\mu}_i^{(j)}$ is the estimate of μ_i when the data is sample j. Definitions for the other parameters are obtained by analogy. All the computations were made using the R system (R Development Core Team 2009) and the implementation of the EM algorithm was computed by employing the R package `mixsmsn` (Cabral, Lachos & Prates 2012), available on CRAN.

As a note about implementation, an expected consequence of the non-identifiability cited in section 3.2 is the permutation of the component labels when using the *k-means* method to perform an initial clustering of the data. This label-witching problem seriously affects the determination of the MSE and consequently the evaluation of the consistency of the estimates (on this issue see e.g Stephens 2000). To overcome this problem we adopted an order restriction on the initial values of the location parameters and estimates for all parameters were sorted according to their true values before computing the bias and MSE. We emphasize that we employ this order restriction order to ensure the determination of the MSE, impartially, to compare the initialization methods.

4.2. Bias and Mean Squared Error (MSE)

Tables 2 and 3 present, respectively, bias and MSE of the estimates in the $2MS$ case. From these tables, we can see that, with both methods, the convergence of the estimates is evidenced, as we can conclude observing the decreasing values of bias and MSE when the sample size increases. They also show that the estimates of the weights p_i and of the location parameters μ_i have lower bias and MSE. On the other side, investigating the MSE values, we can note a different pattern of (slower) convergence to zero for the skewness parameters estimates. It is possibly due to well known inferential problems related to the skewness parameter (DiCiccio & Monti 2004), suggesting the use of larger samples in order to attain the consistency property.

When we analyze the initialization methods performances, we can see that the MM showed better performance than the RVM, for all sample sizes and parameters. When using the RVM, in general, the absolute value of the bias and MSE of the estimates of σ_i^2 and λ_i are very large compared with that obtained using the MM. In general, according to our criteria, we can conclude that the MM method presented a satisfactory performance in all situations.

Table 2: Bias of the estimates - two moderately separated ($2MS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	RVM	0.14309	−0.39956	−0.45779	1.30675	1.32753	−1.10056	0.01927	−0.01927
	MM	−0.01842	−0.02842	0.39275	−0.12159	1.05604	−0.37835	0.00159	0.00159
1000	RVM	0.10815	−0.33814	−0.26339	1.09304	0.61695	−0.60402	0.01599	−0.01599
	MM	−0.02194	−0.01214	0.39285	−0.25737	0.58058	−0.10558	0.00212	−0.00212
5000	RVM	0.10776	−0.26897	−0.21237	0.68574	0.27081	−0.10679	0.01412	−0.01412
	MM	−0.02641	−0.01109	0.35592	−0.45453	0.44153	0.04753	0.00283	0.00283
10000	RVM	0.09904	−0.28784	−0.19722	0.67306	0.29762	0.04354	0.01451	−0.01451
	MM	−0.02783	−0.01026	0.35406	−0.44957	0.42318	0.05692	0.00275	0.00275

Table 3: MSE of the estimates - two moderately separated ($2MS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	RVM	0.69489	2.62335	5.70944	74.43050	55.24861	108.27990	0.00797	0.00797
	MM	0.01509	0.09895	1.81950	13.41757	8.11938	5.48844	0.00071	0.00071
1000	RVM	0.49336	2.34634	4.40063	64.12685	13.10896	44.12675	0.00651	0.00651
	MM	0.00732	0.03158	0.99780	6.14854	1.94043	0.72317	0.00035	0.00035
5000	RVM	0.51481	1.91370	2.94570	44.01847	12.20841	7.70399	0.00544	0.00544
	MM	0.00203	0.00611	0.30162	1.24062	0.45855	0.11495	7.41e-05	7.41e-05
10000	RVM	0.48098	2.14718	3.13302	38.22046	10.01400	4.01044	0.00564	0.00564
	MM	0.00141	0.00598	0.22876	0.71719	0.30481	0.06338	3.97e-05	3.97e-05

The bias and MSE of the estimates for the $2WS$ case are presented in tables 4 and 5, respectively. As in the $2MS$ case, their values decrease when the sample size increases. Comparing the initialization methods, we can see again the poor performance of RVM, notably when estimating σ_i^2 and λ_i. The performance of

Table 4: Bias of the estimates - two well separated ($2WS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	RVM	0.12911	−0.43583	−0.29445	7.32555	1.49636	−0.96377	0.01053	−0.01053
	MM	−0.01943	−0.00087	0.12457	0.09558	0.78816	−0.51038	−0.00013	0.00013
1000	RVM	0.11189	−0.39592	−0.24494	6.83576	1.06715	−0.42096	0.00938	−0.00938
	MM	−0.01896	0.00764	0.11564	0.09444	0.50192	−0.2533	0.00031	0.00031
5000	RVM	0.20668	−0.61846	−0.33728	8.68389	0.57418	−0.26060	0.01440	−0.01440
	MM	−0.01863	0.00770	0.10208	0.09145	0.29384	−0.09125	3.87e-05	−3.87e-05
10000	RVM	0.24109	−0.64407	−0.33492	8.68057	0.37535	−0.15859	0.01457	−0.01457
	MM	−0.01791	0.00615	0.10220	0.08006	−0.27967	−0.07295	−8.37e-05	8.37e-05

Table 5: MSE of the estimates - two well separated ($2WS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	RVM	0.95939	6.72054	2.57082	2314.91800	204.77300	121.73630	0.00466	0.00466
	MM	0.01411	0.08918	0.86552	5.30267	4.67931	6.21511	0.00047	0.00047
1000	RVM	0.86198	6.26880	2.01054	2385.87300	157.64460	38.02320	0.00411	0.00411
	MM	0.00705	0.05053	0.45151	2.58499	1.64592	0.88093	0.00024	0.00024
5000	RVM	1.62846	10.58635	2.33173	2366.03700	80.19515	32.81540	0.00577	0.00577
	MM	0.00164	0.03406	0.62874	0.11866	0.30607	0.16967	4.79e-05	4.79e-05
10000	RVM	2.18692	10.97898	2.35874	2324.93000	31.51082	26.92428	0.00587	0.00587
	MM	0.00099	0.03115	0.07153	0.37187	0.18739	0.10765	2.36e-05	2.36e-05

Table 6: Bias of the estimates - two poorly separated ($2PS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	RVM	0.54135	−3.05044	−6.45635	−4.92197	0.54115	3.93878	0.10598	−0.10598
	MM	0.67462	−2.97859	−6.68560	−5.96459	3.41267	4.80134	0.09124	−0.09124
1000	RVM	0.47419	−3.24148	−6.49341	−5.12707	−0.84129	3.81489	0.09821	−0.09821
	MM	0.67263	−2.76804	−6.44084	−5.94106	−1.59226	4.52825	0.09288	−0.09288
5000	RVM	0.11827	−3.38188	−6.31995	−4.61048	−0.32021	4.27876	0.10485	−0.10485
	MM	0.43959	−2.63605	−6.54154	−5.28364	−1.61009	4.58531	0.10837	−0.10837
10000	RVM	−0.01664	−3.32212	−6.26154	−4.56424	0.04022	4.33918	0.10793	−0.10793
	MM	0.34272	−2.58084	−6.36467	−5.15910	−0.35239	4.26839	0.11364	−0.11364

Table 7: MSE of the estimates - two poorly separated ($2PS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}_1$	$\hat{p}_2$
500	RVM	1.53553	12.77114	44.60716	61.74994	196.12690	48.65583	0.02622	0.02622
	MM	1.56413	10.28718	46.61793	49.10880	343.77230	30.24965	0.01746	0.01746
1000	RVM	1.67311	14.41824	44.47786	63.53965	56.47113	51.72255	0.02561	0.02561
	MM	1.39797	8.39757	42.72479	49.20213	25.63282	30.17503	0.01755	0.01755
5000	RVM	1.91642	16.52284	42.21555	60.68243	84.34230	37.77900	0.02987	0.02987
	MM	0.72188	7.45145	43.16745	35.72524	17.52604	25.12349	0.01734	0.01734
10000	RVM	2.18807	16.37261	41.17759	57.61406	117.18120	37.92872	0.02981	0.02981
	MM	0.48657	7.39552	41.31111	32.51687	19.04316	19.12946	0.01783	0.01783

Table 8: Bias of the estimates - two poorly separated and one well separated ($3PWS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$
500	RVM	3.86197	−2.22030	0.34248	0.02341	0.01208	−0.03549
	MM	−0.02634	−0.07339	0.00338	0.49861	−1.95137	1.98042
1000	RVM	0.13962	−0.32733	0.06264	−0.43484	−1.79386	4.25781
	MM	−0.02324	−0.04743	−0.01057	0.35349	−1.50284	1.19124
5000	RVM	0.09945	−0.27154	0.06481	−0.25861	−1.36352	3.16178
	MM	−0.02336	−0.04234	−0.01724	0.43048	−0.91359	0.40114
10000	RVM	0.08625	−0.25897	0.04110	−0.23192	−1.23694	3.24305
	MM	−0.02290	−0.04478	−0.01054	0.42927	−0.74921	0.32357

Table 9: Bias of the estimates - two poorly separated and one well separated ($3PWS$).

n	Method	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	RVM	3.86197	−2.22029	0.34248	0.02341	−0.03549	0.01208
	MM	2.78183	−1.00171	0.37213	0.00378	−0.01641	0.01263
1000	RVM	1.46755	−0.59057	0.17228	0.01745	−0.02687	0.00941
	MM	0.92051	−0.11912	0.22981	0.00218	−0.01139	0.00917
5000	RVM	0.66370	−0.05447	0.02601	0.01554	−0.01996	0.00441
	MM	0.57731	0.10412	0.13295	0.00285	−0.00614	0.00328
10000	RVM	0.53239	0.12565	0.01455	0.01376	−0.01651	0.00274
	MM	0.53269	0.10807	0.11858	0.00261	−0.00474	0.00212

Table 10: MSE of the estimates - two poorly separated and one well separated ($3PWS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$
500	RVM	0.73592	4.19665	1.33364	7.66927	13.88521	443.83960
	MM	0.02326	0.11884	0.10291	2.91056	9.62248	16.26670
1000	RVM	0.63416	3.13945	1.31999	5.46298	9.99991	335.50440
	MM	0.01062	0.12819	0.04066	1.56864	5.58594	6.68605
5000	RVM	0.46937	2.92722	1.35910	4.07581	5.74955	360.56410
	MM	0.00305	0.08728	0.00844	0.68559	1.76444	5.77978
10000	RVM	0.38973	2.83554	1.24776	3.41887	5.16622	349.93790
	MM	0.00204	0.07621	0.00434	0.54306	1.11403	3.81096

Table 11: MSE of the estimates - two poorly separated and one well separated ($3PWS$).

n	Method	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	RVM	401.77671	659.16871	15.66503	0.00691	0.00077	0.00633
	MM	89.47438	57.35417	2.67579	0.00068	0.00043	0.00056
1000	RVM	74.62215	95.01346	8.04131	0.00553	0.00053	0.00506
	MM	3.92305	1.23581	0.88842	0.00033	0.00022	0.00027
5000	RVM	19.78691	120.88661	1.57637	0.00433	0.00021	0.00427
	MM	0.73751	0.30914	0.18520	7.20e-05	3.75e-05	6.93e-05
10000	RVM	18.46118	7.03221	1.60081	0.00417	0.00031	0.00363
	MM	0.50922	0.24319	0.09469	4.23e-05	1.82e-05	4.10e-05

Table 12: Bias of the estimates - three well separated ($3WWS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$
500	RVM	0.24221	−0.48999	0.07992	−0.70713	−1.98125	11.24986
	MM	−0.02717	−0.02233	0.00361	0.19686	−1.50136	1.60453
1000	RVM	0.24327	−0.49722	0.04203	−0.63272	−1.46754	9.55012
	MM	−0.02742	−0.01908	−0.01276	0.10826	−0.97860	1.18908
5000	RVM	0.33293	−0.67092	0.01975	−0.65223	−1.02971	10.20861
	MM	−0.02265	−0.01872	−0.01604	0.11568	−0.39018	0.58417
10000	RVM	0.33457	−0.80115	−0.07909	−0.63707	−0.90179	8.87343
	MM	−0.02328	−0.01509	−0.01704	0.10738	−0.28876	0.43182

Table 13: Bias of the estimates - three well separated ($3WWS$).

n	Method	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	RVM	3.63818	−0.59785	0.16417	0.02216	0.01369	−0.03586
	MM	1.98187	−0.83366	0.31660	−0.00014	0.01385	−0.01371
1000	RVM	1.68173	−0.44696	0.07427	0.02026	0.00987	−0.03014
	MM	0.83325	−0.29503	0.25427	9.95e-05	0.00977	−0.00987
5000	RVM	0.85209	−0.16743	0.11006	0.02449	0.00285	−0.02735
	MM	0.38848	−0.08961	0.12745	−1.00e-05	0.00435	−0.00434
10000	RVM	0.45846	0.06753	0.13875	0.02299	0.00213	−0.02512
	MM	0.34917	−0.04635	0.10236	−1.45-e05	0.00308	−0.00306

Table 14: MSE of the estimates - three well separated ($3WWS$).

n	Method	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$
500	RVM	1.92196	9.35109	2.52362	6.19796	10.97185	1848.86500
	MM	0.02247	0.17468	0.09472	1.43362	6.87787	8.26021
1000	RVM	2.02266	10.57483	2.68414	5.17628	9.12883	1517.61500
	MM	0.01091	0.14971	0.04303	0.79533	3.51761	3.92717
5000	RVM	2.93899	12.82373	3.23907	5.17035	8.56857	1510.38300
	MM	0.00249	0.14126	0.00836	0.21648	0.89455	0.81992
10000	RVM	3.08463	14.30240	2.85946	4.53532	9.51754	1383.25000
	MM	0.00156	0.13665	0.00412	0.17027	0.79811	0.42657

Table 15: MSE of the estimates - three well separated ($3WWS$).

n	Method	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
500	RVM	361.45080	31.09128	9.18895	0.00727	0.00102	0.00712
	MM	34.37280	18.59549	2.08116	0.00047	0.00042	0.00041
1000	RVM	132.29800	28.30730	3.20316	0.00656	0.00089	0.00632
	MM	3.81642	0.92243	1.51732	0.00024	0.00022	0.00021
5000	RVM	41.27475	23.97153	1.03027	0.00845	0.00111	0.00684
	MM	0.50781	0.30959	0.18493	4.86e05	4.15e-05	4.15e-05
10000	RVM	22.70630	10.52715	1.66676	0.00754	0.00091	0.00647
	MM	0.30177	0.29030	0.08916	2.41e-05	2.11e-05	2.07e-05

MM is satisfactory and we can say that, in general, the conclusions made for the $2MS$ case are still valid here.

We present the results for the $2PS$ case in tables 6 and 7. Bias and MSE are larger than in the $2MS$ and $2WS$ cases (for all sample sizes) with both methods. Also, the consistency of the estimates seems to be unsatisfactory, clearly not attained in the σ_i^2 and λ_i cases. According to the related literature, such drawbacks of the algorithm are expected when the population presents a remarkable homogeneity. An exception is made when the initial values are closer to the true parameter values see, for example, McLachlan & Peel (2000) and the abovementioned work of Park & Ozeki (2009).

For the $3PWS$ case the results for the bias are shown in tables 8 and 9 and for the MSE in tables 10 and 11. It seems that consistency is achieved for $\hat{p}_i$, $\hat{\mu}_i$ and $\hat{\sigma}_i^2$, using MM. However, this is not the behavior for $\hat{\lambda}_i$. This instability is common to all initialization methods, according to the MSE criterion. Using the RVM method we obtained, as before, larger values of bias and MSE. These results are similar to that obtained for the FM-SN model with two components

Finally, for the $3PWW$ case, the bias of the estimates is presented in tables 12 and 13 and the EQM are shown in tables 14 and 15. Concerning the estimates $\hat{p}_i$ and $\hat{\mu}_i$, very satisfactory results are obtained, with small values of bias and MSE when using the MM. The values of MSE of $\hat{\sigma}_i^2$ exhibit a decreasing behavior when the sample size increases. On the other side, although we are in the well separated case, the values of bias and MSE of $\hat{\lambda}_i$ are larger, notably when using RVM as the initialization method.

Concluding this section, we can say that, as a general rule, the MM can be seen as a good alternative for real applications. If this condition is maintained, our study suggests that the consistency property holds for all EM estimates (it may be slower for the scale parameter!), except for the skewness parameter, indicating that a sample size larger than 5,000 is necessary to achieve consistency in the case of this parameter. The study also suggests that the degree of heterogeneity of the population has a remarkable influence on the quality of the estimates.

TABLE 16: Means and standard deviations ($\times 10^{-4}$) of d_r.

Method	n	Cases		
		$2MS$	2WS	2PS
RVM	500	1.43 (2.55)	2.51 (7.31)	2.23 (1.31)
	1000	1.05 (2.21)	2.24 (6.48)	0.84 (0.69)
	5000	0.75 (2.24)	2.01 (8.67)	1.02 (0.75)
	10000	0.67 (2.35)	2.14 (8.86)	0.67 (0.69)
MM	500	0.90 (0.66)	1.32 (0.95)	2.20 (1.21)
	1000	0.62 (0.48)	1.23 (0.80)	0.76 (0.57)
	5000	0.33 (0.24)	0.34 (0.26)	0.79 (0.46)
	10000	0.22 (0.16)	0.42 (0.26)	0.44 (0.33)

4.3. Density Estimation Analysis

In this section we consider the density estimation issue, that is, the point estimation of the parameter $\ell(\Theta)$, the log-likelihood evaluated at the true value of the parameter. We considered FM-SN models with two components and restricted ourselves to the cases $2MS$, $2WS$ and $2PS$, with sample sizes $n =$500; 1,000; 5,000 and 10,000. For each combination of parameters and sample size, 5000 samples were generated and the following measure was considered to compare the methods of initialization

$$d_r(M) = \left| \frac{\ell(\Theta) - \ell_{(M)}(\hat{\Theta})}{\ell(\Theta)} \right| \times 100$$

where $\ell_{(M)}(\hat{\Theta})$ is the log-likelihood evaluated at the EM estimate $\hat{\Theta}$, which was obtained using the initialization method M. According to this criterion, an initialization method M is better than M' if $d_r(M) < d_r(M')$. Table 16 presents the means and standard deviations of d_r.

For $2MS$ case, we can see that these values decrease when the sample size increases with both methods and that the MM presented the lowest mean value and standard deviation for all sample sizes. For the $2WS$ case, we do not observe a monotone behavior for d_r, the mean values and the standard errors are larger than that presented in the $2MS$ case, with poor performance of RVM. In this $2PS$ case, although we also do not observe a monotone behavior for d_r, we can see that the MM presented a better performance than the TVM.

The main message is that the MM method seems to be suitable when we are interested in the estimation of the likelihood values, with some caution when the population is highly homogeneous.

4.4. Number of Iterations

It is well known that one of the major drawbacks of the EM algorithm is the slow convergence. The problem becomes more serious when there is a bad choice of the starting values (McLachlan & Krishnan 2008). Consequently, an important issue is the investigation of the number of iterations necessary for the convergence of the algorithm. As in subsection 4.3, here we consider only the $2MS$, $2WS$ and $2PS$ cases, with the same sample sizes and number of replications. For each generated sample, we observed the number of iterations and the means and standard deviations of this quantity were computed. The simulations results are reported in Table 17.

Results suggest that in the three cases, using MM, the mean number of iterations decreases as the sample size increases, but the same is not true when RVM is adopted as the initialization method. For the $2PS$ case, as expected, we have a poor behavior possibly due to the population homogeneity, as we commented before. An interesting fact is that, in the $2PS$ case, the RVM has a smaller mean number of iterations.

Table 17: Means and standard deviations of number of iterations.

Method	n	Cases		
		$2MS$	2WS	2PS
RVM	500	306.14 (387.70)	129.81 (117.01)	337.82 (289.01)
	1,000	283.57 (289.32)	126.87 (130.61)	319.70 (242.08)
	5,000	280.28 (260.36)	128.29 (213.99)	336.85 (206.38)
	10,000	286.31 (271.83)	131.33 (190.77)	353.61 (211.99)
MM	500	147.53 (72.79)	126.62 (26.80)	457.97 (167.28)
	1,000	129.37 (33.34)	119.70 (15.39)	429.42 (119.77)
	5,000	116.35 (11.57)	113.91 (5.90)	372.14 (71.89)
	10,000	115.10 (8.37)	113.29 (4.23)	352.33 (97.54)

5. A Simulation Study of Model Choice

There is a key issue with the use of finite mixtures to estimate the number of components in order to obtain a suitable fit. One possible approach is to use some criteria function and compute

$$\hat{k} = \arg\min_{k}\{C(\hat{\Theta}_{(k)}),\ k \in \{k_{\min}, \ldots, k_{\max}\}\}$$

where $C(\hat{\Theta}_{(k)})$ is the criterion function evaluated at the EM estimate $\hat{\Theta}_{(k)}$, obtained by modeling the data using the FM-SN model with k components, and k_{min} and k_{max} are fixed positive integers (for other approaches see McLachlan & Peel 2000).

Our main purpose in this section is to investigate the ability of some classical criteria to estimate the correct number of mixture components. We consider the Akaike Information Criterion (AIC) (Akaike 1974), the Bayesian Information Criterion (BIC) (Schwarz 1978), the Efficient Determination Criterion (EDC) (Bai, Krishnaiah & Zhao 1989) and the Integrated Completed Likelihood Criterion (ICL) (Biernacki, Celeux & Govaert 2000). The AIC, BIC and EDC criteria have the form

$$-2\,\ell(\hat{\Theta}) + d_k\, c_n$$

where $\ell(\cdot)$ is the actual log-likelihood, d_k is the number of free parameters that have to be estimated under the model with k components and the penalty term c_n is a convenient sequence of positive numbers. We have $c_n = 2$ for AIC and $c_n = \log(n)$ for BIC. For the EDC criterion, c_n is chosen so that it satisfies the conditions

$$\lim(c_n/n) = 0 \quad \text{and} \quad \lim_{n\to\infty}(c_n/(\log\log n)) = \infty$$

Here we compare the following alternatives

$$c_n = 0.2\sqrt{n}, \quad c_n = 0.2\log(n), \quad c_n = 0.2n/\log(n), \quad \text{and} \quad c_n = 0.5\sqrt{n}$$

The ICL is defined as

$$-2\,\ell^{*}(\hat{\Theta}) + d_k\ \log(n),$$

where $\ell^*(\cdot)$ is the integrated log-likelihood of the sample and the indicator latent variables, given by

$$\ell^*(\hat{\Theta}) = \sum_{i=1}^{k} \sum_{j \in C_i} \log(\hat{p}_i \text{SN}(y_j|\hat{\theta}_i))$$

where C_i is a set of indexes defined as: j belongs to C_i if, and only if, the observation y_j is allocated to component i by the following clustering process: after the FM-SN model with k components was fitted using the EM algorithm we obtain the estimate of the posterior probability that an observation y_i belongs to the jth component of the mixture, $\hat{z}_{ij}$ (see equation (6)). If $q = \arg\max_j \{\hat{z}_{ij}\}$ we allocate y_i to the component q.

In this study we simulated samples of the FM-SN model with $k = 3$, $p_1 = p_2 = p_3 = 1/3$, $\mu_1 = 5$, $\mu_2 = 20$, $\mu_3 = 28$, $\sigma_1^2 = 9$, $\sigma_2^2 = 16$, $\sigma_3^2 = 16$, $\lambda_1 = 6$, $\lambda_2 = -4$ and $\lambda_3 = 4$, and considered the sample sizes $n = 200$, 300, 500,1000, 5000. Figure 3 shows a typical sample of size 1000 following this specified set up.

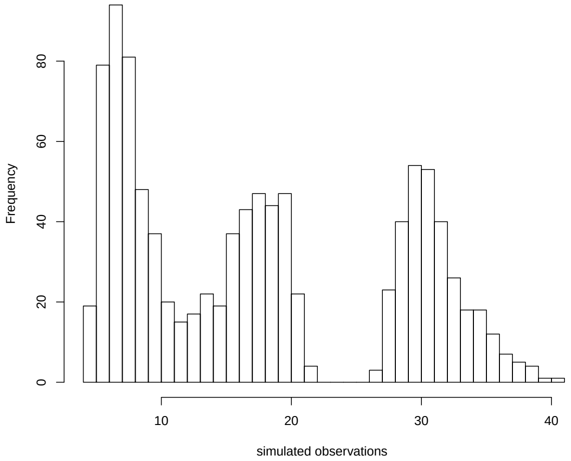


Figure 3: Histogram of a FM-SN sample with $k = 3$ and $n = 1,000$.

For each generated sample (with fixed number of 3 components) we fitted the FM-SN model with $k = 2$, $k = 3$ and $k = 4$, using the EM algorithm initialized by the method of moments. For each fitted model the criteria AIC, BIC, ICL and EDC were computed. We repeated this procedure 500 times and obtained the percentage of times some given criterion chooses the correct number of components. The results are reported in Table 18

We can see that BIC and ICL have a better performance than AIC for all sample sizes. Except for AIC, the rates presented an increasing behavior when the sample size increases. This possible drawback of AIC may be due to the fact that its definition does not take into account the sample size in its penalty term. Results for BIC and ICL were similar, while EDC showed some dependence on the term c_n. In general, we can say that BIC and ICL have equivalent abilities

Table 18: Percentage of times that the criteria chosen the correct model.

n	AIC	BIC	ICL	EDC/ c_n			
				$0.2\log(n)$	$0.2\sqrt{n}$	$0.2n/\log(n)$	$0.5\sqrt{n}$
200	94.2	99.2	99.2	77.8	98.4	99.4	99.4
300	94.0	98.8	98.8	78.2	98.4	98.8	98.8
500	95.8	99.8	99.8	86.4	99.8	99.8	99.8
1000	96.2	100.0	100.0	88.5	100.0	100.0	100.0
5000	95.6	100.0	100.0	92.8	100.0	100.0	100.0

to choose the correct number of components and that, depending on the choice of c_n, ICL can not be as good as AIC or better than ICL and BIC.

6. Final Remarks

In this work we presented a simulation study in order to investigate the performance of the EM algorithm for maximum likelihood estimation in finite mixtures of skew-normal distributions with component specific parameters. The results show that the algorithm produces quite reasonable estimates, in the sense of consistency and the total number of iterations, when using the method of moments to obtain the starting points. The study also suggested that the random initialization method used is not a reasonable procedure. When the EM estimates were used to compute some model choice criteria (AIC, BIC, ICL and EDC), the results suggest that the EDC, with the penalization term appropriate, provides a good alternative to estimate the number of components of the mixture. On the other side, these patterns do not hold when the mixture components are poorly separated, notably for the skewness parameters estimates which, in addition, showed a performance strongly dependent on large samples. Possible extensions of this work include the multivariate case and a wider family of skewed distributions, like the class of skew-normal independent distributions (see Cabral et al. (2012)).

[Recibido: agosto de 2011 — Aceptado: octubre de 2012]

References

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**, 716–723.

Azzalini, A. (1985), 'A class of distributions which includes the normal ones', *Scandinavian Journal of Statistics* **12**, 171–178.

Azzalini, A. (2005), 'The skew-normal distribution and related multivariate families', *Scandinavian Journal of Statistics* **32**, 159–188.

Bai, Z. D., Krishnaiah, P. R. & Zhao, L. C. (1989), 'On rates of convergence of efficient detection criteria in signal processing with white noise', *IEEE Transactions on Information Theory* **35**, 380–388.

Basso, R. M., Lachos, V. H., Cabral, C. R. B. & Ghosh, P. (2010), 'Robust mixture modeling based on scale mixtures of skew-normal distributions', *Computational Statistics and Data Analysis* **54**, 2926–2941.

Bayes, C. L. & Branco, M. D. (2007), 'Bayesian inference for the skewness parameter of the scalar skew-normal distribution', *Brazilian Journal of Probability and Statistics* **21**, 141–163.

Biernacki, C., Celeux, G. & Govaert, G. (2000), 'Assessing a mixture model for clustering with the integrated completed likelihood', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 719–725.

Biernacki, C., Celeux, G. & Govaert, G. (2003), 'Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models.', *Computational Statistics and Data Analysis* **41**, 561–575.

Cabral, C. R. B., Lachos, V. H. & Prates, M. O. (2012), 'Multivariate mixture modeling using skew-normal independent distributions', *Computational Statistics and Data Analysis* **56**, 126–142.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Dias, J. G. & Wedel, M. (2004), 'An empirical comparison of EM, SEM and MCMC performance for problematic gaussian mixture likelihoods', *Statistics and Computing* **14**, 323–332.

DiCiccio, T. J. & Monti, A. C. (2004), 'Inferential aspects of the skew exponential power distribution', *Journal of the American Statistical Association* **99**, 439–450.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer Verlag.

Greselin, F. & Ingrassia, S. (2010), 'Constrained monotone EM algorithms for mixtures of multivariate t distributions', *Statistics and Computing* **20**(1), 9–22.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition*, Springer, USA.

Hathaway, R. J. (1985), 'A constrained formulation of maximum-likelihood estimation for normal mixture models', *The Annals of Statistics* **13**, 795–800.

Henze, N. (1986), 'A probabilistic representation of the skew-normal distribution', *Scandinavian Journal of Statistics* **13**, 271–275.

Ho, H. J., Pyne, S. & Lin, T. I. (2012), 'Maximum likelihood inference for mixtures of skew Student-t-normal distributions through practical EM-type algorithms', *Statistics and Computing* **22**(1), 287–299.

Ingrassia, S. (2004), 'A likelihood-based constrained algorithm for multivariate normal mixture models', *Statistical Methods and Applications* **13**, 151–166.

Ingrassia, S. & Rocci, R. (2007), 'Constrained monotone EM algorithms for finite mixture of multivariate gaussians', *Computational Statistics and Data Analysis* **51**, 5339–5351.

Karlis, D. & Xekalaki, E. (2003), 'Choosing initial values for the EM algorithm for finite mixtures', *Computational Statistics and Data Analysis* **41**, 577–590.

Lin, T. I. (2009), 'Maximum likelihood estimation for multivariate skew normal mixture models', *Journal of Multivariate Analysis* **100**, 257–265.

Lin, T. I. (2010), 'Robust mixture modeling using multivariate skew t distributions', *Statistics and Computing* **20**(3), 343–356.

Lin, T. I., Lee, J. C. & Hsieh, W. J. (2007), 'Robust mixture modelling using the skew t distribution', *Statistics and Computing* **17**, 81–92.

Lin, T. I., Lee, J. C. & Ni, H. F. (2004), 'Bayesian analysis of mixture modelling using the multivariate t distribution', *Statistics and Computing* **14**, 119–130.

Lin, T. I., Lee, J. C. & Yen, S. Y. (2007), 'Finite mixture modelling using the skew normal distribution', *Statistica Sinica* **17**, 909–927.

Lin, T. & Lin, T. (2010), 'Supervised learning of multivariate skew normal mixture models with missing information', *Computational Statistics* **25**(2), 183–201.

McLachlan, G. J. & Krishnan, T. (2008), *The EM Algorithm and Extensions*, 2 edn, John Wiley and Sons.

McLachlan, G. J. & Peel, G. J. (2000), *Finite Mixture Models*, John Wiley and Sons.

Meng, X. L. & Rubin, D. B. (1993), 'Maximum likelihood estimation via the ECM algorithm: A general framework', *Biometrika* **80**, 267–278.

Nityasuddhi, D. & Böhning, D. (2003), 'Asymptotic properties of the EM algorithm estimate for normal mixture models with component specific variances', *Computational Statistics and Data Analysis* **41**, 591–601.

Park, H. & Ozeki, T. (2009), 'Singularity and slow convergence of the EM algorithm for gaussian mixtures', *Neural Process Letters* **29**, 45–59.

Peel, D. & McLachlan, G. J. (2000), 'Robust mixture modelling using the t distribution', *Statistics and Computing* **10**, 339–348.

R Development Core Team (2009), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Schwarz, G. (1978), ‘Estimating the dimension of a model’, *Annals of Statistics* **6**, 461–464.

Shoham, S. (2002), ‘Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions’, *Pattern Recognition* **35**, 1127–1142.

Shoham, S., Fellows, M. R. & Normann, R. A. (2003), ‘Robust, automatic spike sorting using mixtures of multivariate t-distributions’, *Journal of Neuroscience Methods* **127**, 111–122.

Stephens, M. (2000), ‘Dealing with label switching in mixture models’, *Journal of the Royal Statistical Society. Series B* **62**, 795–809.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, John Wiley and Sons.

Yakowitz, S. J. & Spragins, J. D. (1968), ‘On the identifiability of finite mixtures’, *The Annals of Mathematical Statistics* **39**, 209–214.

Yao, W. (2010), ‘A profile likelihood method for normal mixture with unequal variance’, *Journal of Statistical Planning and Inference* **140**, 2089–2098.

Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 479 a 508

An Introductory Review of a Structural VAR-X Estimation and Applications

Una revisión introductoria de la estimación y aplicaciones de un VAR-X estructural

Sergio Ocampo[1,a], Norberto Rodríguez[2,3,b]

[1]Research Department, Inter-American Development Bank, Washington, DC, United States of America

[2]Macroeconomic Modeling Department, Banco de la República, Bogotá, Colombia

[3]Statistics Department, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

This document presents how to estimate and implement a structural VAR-X model under long run and impact identification restrictions. Estimation by Bayesian and classical methods is presented. Applications of the structural VAR-X for impulse response functions to structural shocks, multiplier analysis of the exogenous variables, forecast error variance decomposition and historical decomposition of the endogenous variables are also described, as well as a method for computing higher posterior density regions in a Bayesian context. Some of the concepts are exemplified with an application to US data.

***Key words*:** Econometrics, Bayesian time series, Vector autoregression, Structural model.

Resumen

Este documento cubre la estimación e implementación del modelo VAR-X estructural bajo restricciones de identificación de corto y largo plazo. Se presenta la estimación tanto por métodos clásicos como Bayesianos. También se describen aplicaciones del modelo como impulsos respuesta ante choques estructurales, análisis de multiplicadores de las variables exógenas, descomposición de varianza del error de pronóstico y descomposición histórica de las variables endógenas. Así mismo se presenta un método para calcular regiones de alta densidad posterior en el contexto Bayesiano. Algunos de los conceptos son ejemplificados con una aplicación a datos de los Estados Unidos.

***Palabras clave*:** econometría, modelo estructural, series de tiempo Bayesianas, vector autoregresivo.

[a]Research Fellow. E-mail: socampo@iadb.org

[b]Principal Econometrist and Lecturer. E-mail: nrodrini@banrep.gov.co

1. Introduction

The use of Vector Autoregression with exogenous variables (VAR-X) and structural VAR-X models in econometrics is not new, yet textbooks and articles that use them often fail to provide the reader a concise (and moreover useful) description of how to implement these models (Lütkepohl (2005) constitutes an exception of this statement). The use of Bayesian techniques in the estimation of VAR-X models is also largely neglected from the literature, as is the construction of the historical decomposition of the endogenous variables. This document builds upon the Structural Vector Autoregression (S-VAR) and Bayesian Vector Autoregression (B-VAR) literature and its purpose is to present a review of some of the basic features that accompany the implementation of a structural VAR-X model.

Section 2 presents the notation and general setup to be followed throughout the document. Section 3 discusses the identification of structural shocks in a VAR-X, with both long run restrictions, as in Blanchard & Quah (1989), and impact restrictions, as in Sims (1980, 1986). Section 4 considers the estimation of the parameters by classical and Bayesian methods. In Section 5, four of the possible applications of the model are presented, namely the construction of impulse response functions to structural shocks, multiplier analysis of the exogenous variables, forecast error variance decomposition and historical decomposition of the endogenous variables. Section 6 exemplifies some of the concepts developed in the document using Galí's (1999) structural VAR augmented with oil prices as an exogenous variable. Finally Section 7 concludes.

2. General Setup

In all sections the case of a structural VAR-X whose reduced form is a VAR-X(p, q) will be considered. It is assumed that the system has n endogenous variables ($\mathbf{y}_t$) and m exogenous variables ($\mathbf{x}_t$). The variables in $\mathbf{y}_t$ and $\mathbf{x}_t$ may be in levels or in first differences, this depends on the characteristics of the data, the purpose of the study, and the identification strategy, in all cases no co-integration is assumed. The reduced form of the structural model includes the first p lags of the endogenous variables, the contemporaneous values and first q lags of the exogenous variables and a constant vector.[1] Under this specification it is assumed that the model is stable and presents white-noise Gaussian residuals ($\mathbf{e}_t$), i.e. $\mathbf{e}_t \overset{iid}{\sim} N(\mathbf{0}, \mathbf{\Sigma})$, moreover, $\mathbf{x}_t$ is assumed to be uncorrelated with $\mathbf{e}_t$ for all leads and lags.

The reduced form VAR-X(p, q) can be represented as in equation (1) or equation (2), where $\mathbf{v}$ is a n-vector, $\mathbf{B}_i$ are $n \times n$ matrices, with $i \in \{1, \ldots, p\}$, and $\mathbf{\Theta}_j$ are $n \times m$ matrices, with $j \in \{1, \ldots, q\}$. In equation (2) one has $\mathbf{B}(L) = \mathbf{B}_1 L + \cdots + \mathbf{B}_p L^p$ and $\mathbf{\Theta}(L) = \mathbf{\Theta}_0 + \cdots + \mathbf{\Theta}_q L^q$, both matrices of polynomials in

[1] The lag structure of the exogenous variables may be relaxed allowing different lags for each variable. This complicates the estimation and is not done here for simplicity. Also, the constant vector or intercept may be omitted according to the characteristics of the series used.

the lag operator L.

$$\mathbf{y}_t = \mathbf{v} + \mathbf{B}_1\mathbf{y}_{t-1} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \mathbf{\Theta}_0\mathbf{x}_t + \cdots + \mathbf{\Theta}_q\mathbf{x}_{t-q} + \mathbf{e}_t \tag{1}$$

$$\mathbf{y}_t = \mathbf{v} + \mathbf{B}\left(L\right)\mathbf{y}_t + \mathbf{\Theta}\left(L\right)\mathbf{x}_t + \mathbf{e}_t \tag{2}$$

Defining $\mathbf{\Psi}\left(L\right) = \mathbf{\Psi}_0 + \mathbf{\Psi}_1 L + \ldots = \left[\mathbf{I} - \mathbf{B}\left(L\right)\right]^{-1}$ with $\mathbf{\Psi}_0 = \mathbf{I}$ as an infinite polynomial on the lag operator L, one has the VMA-X representation of the model, equation (3).[2]

$$\mathbf{y}_t = \mathbf{\Psi}\left(1\right)\mathbf{v} + \mathbf{\Psi}\left(L\right)\mathbf{\Theta}\left(L\right)\mathbf{x}_t + \mathbf{\Psi}\left(L\right)\mathbf{e}_t \tag{3}$$

Finally, there is a structural VAR-X model associated with the equations above, most of the applications are obtained from it, for example those covered in Section 5. Instead of the residuals ($\mathbf{e}$), which can be correlated among them, the structural model contains structural disturbances with economic interpretation (ϵ), this is what makes it useful for policy analysis. It will be convenient to represent the model by its Vector Moving Average (VMA-X) form, equation (4),

$$\mathbf{y}_t = \mu + \mathbf{C}\left(L\right)\epsilon_t + \mathbf{\Lambda}\left(L\right)\mathbf{x}_t \tag{4}$$

where the endogenous variables are expressed as a function of a constant n-vector (μ), and the current and past values of the structural shocks (ϵ) and the exogenous variables. It is assumed that ϵ is a vector of white noise Gaussian disturbances with identity covariance matrix, i.e. $\epsilon_t \overset{iid}{\sim} N\left(\mathbf{0}, \mathbf{I}\right)$. Both $\mathbf{C}\left(L\right)$ and $\mathbf{\Lambda}\left(L\right)$ are infinite polynomials in the lag operator L, each matrix of $\mathbf{C}\left(L\right)$ $\left(\mathbf{C}_0, \mathbf{C}_1, \ldots\right)$ is of size $n \times n$, and each matrix of $\mathbf{\Lambda}\left(L\right)$ $\left(\mathbf{\Lambda}_0, \mathbf{\Lambda}_1, \ldots\right)$ is of size $n \times m$.

3. Identification of Structural Shocks in a VAR-X

The identification of structural shocks is understood here as a procedure which enables the econometrician to obtain the parameters of a structural VAR-X from the estimated parameters of the reduced form of the model. As will be clear from the exposition below, the identification in presence of exogenous variables is no different from what is usually done in the S-VAR literature. Equating (3) and (4) one has:

$$\mu + \mathbf{\Lambda}\left(L\right)\mathbf{x}_t + \mathbf{C}\left(L\right)\epsilon_t = \mathbf{\Psi}\left(1\right)\mathbf{v} + \mathbf{\Psi}\left(L\right)\mathbf{\Theta}\left(L\right)\mathbf{x}_t + \mathbf{\Psi}\left(L\right)\mathbf{e}_t$$

then the following equalities can be inferred:

$$\mu = \mathbf{\Psi}\left(1\right)\mathbf{v} \tag{5}$$

$$\mathbf{\Lambda}\left(L\right) = \mathbf{\Psi}\left(L\right)\mathbf{\Theta}\left(L\right) \tag{6}$$

$$\mathbf{C}\left(L\right)\epsilon_t = \mathbf{\Psi}\left(L\right)\mathbf{e}_t \tag{7}$$

[2] The models stability condition implies that $\mathbf{\Psi}\left(1\right) = \left[\mathbf{I} - \sum_{i=1}^{p}\mathbf{B}_i\right]^{-1}$ exists and is finite.

Since the parameters in $\mathbf{v}$, $\mathbf{B}(L)$ and $\mathbf{\Theta}(L)$ can be estimated from the reduced form VAR-X representation, the values of μ and $\mathbf{\Lambda}(L)$ are also known.[3] Only the parameters in $\mathbf{C}(L)$ are left to be identified, the identification depends on the type of restrictions to be imposed. From equations (5), (6) and (7) is clear that the inclusion of exogenous variables in the model has no effect in the identification of the structural shocks. Equation (7) also holds for a structural VAR model.

The identification restrictions to be imposed over $\mathbf{C}(L)$ may take several forms. Since there is nothing different in the identification between the case presented here and the S-VAR literature, we cover only two types of identification procedures, namely: impact and long run restrictions that allow the use of the Cholesky decomposition. It is also possible that the economic theory points at restrictions that make impossible a representation in which the Cholesky decomposition can be used, or that the number of restrictions exceeds what is needed for exact identification. Both cases complicate the estimation of the model, and the second one (over-identification) makes possible to carry out tests over the restrictions imposed. For a more comprehensive treatment of these problems we refer to Amisano & Giannini (1997).

There is another identification strategy that will not be covered in this document, identification by sign restrictions over some of the impulse response functions. This kind of identification allows to avoid some puzzles that commonly arise in the VAR literature. References to this can be found in Uhlig (2005), Mountford & Uhlig (2009), Canova & De Nicolo (2002), Canova & Pappa (2007) and preceding working papers of those articles originally presented in the late 1990's. More recently, the work of Moon, Schorfheide, Granziera & Lee (2011) presents how to conduct inference over impulse response functions with sign restrictions, both by classical and Bayesian methods.

3.1. Identification by Impact Restrictions

In Sims (1980, 1986) the identification by impact restrictions is proposed, the idea behind is that equation (7) is equating two polynomials in the lag operator L, for them to be equal it must be the case that:

$$\begin{aligned} \mathbf{C}_i L^i \epsilon_t &= \mathbf{\Psi}_i L^i \mathbf{e}_t \\ \mathbf{C}_i \epsilon_t &= \mathbf{\Psi}_i \mathbf{e}_t \end{aligned} \tag{8}$$

Equation (8) holds for all i, in particular it holds for $i = 0$. Recalling that $\mathbf{\Psi}_0 = \mathbf{I}$, the following result is obtained:

$$\mathbf{C}_0 \epsilon_t = \mathbf{e}_t \tag{9}$$

then, by taking the variance on both sides one gets:

$$\mathbf{C}_0 \mathbf{C}_0^{'} = \mathbf{\Sigma} \tag{10}$$

[3] Lütkepohl (2005) presents methods for obtaining the matrices in $\mathbf{\Psi}(L)$ and the product $\mathbf{\Psi}(L)\mathbf{\Theta}(L)$ recursively in Sections 2.1.2 and 10.6, respectively. $\mathbf{\Psi}(1)$ is easily computed by taking the inverse on $\mathbf{I} - \mathbf{B}_1 - \ldots - \mathbf{B}_p$.

Algorithm 1 Identification by Impact Restrictions

1. Estimate the reduced form of the VAR-X.
2. Calculate the VMA-X representation of the model (matrices $\mathbf{\Psi}_i$) and the covariance matrix of the reduced form disturbances e (matrix $\mathbf{\Sigma}$).
3. From the Cholesky decomposition of $\mathbf{\Sigma}$ calculate matrix $\mathbf{C}_0$.

$$\mathbf{C}_0 = \text{chol}\,(\mathbf{\Sigma})$$

4. For $i = 1, \ldots, R$, with R given, calculate the matrices $\mathbf{C}_i$ as:

$$\mathbf{C}_i = \mathbf{\Psi}_i \mathbf{C}_0$$

Identification is completed since all matrices of the structural VMA-X are known.

Since $\mathbf{\Sigma}$ is a symmetric, positive definite matrix it is not possible to infer in a unique form the parameters of $\mathbf{C}_0$ from equation (10), restrictions over the parameters of $\mathbf{C}_0$ have to be imposed. Because $\mathbf{C}_0$ measures the impact effect of the structural shocks over the endogenous variables, those restrictions are called here impact restrictions. Following Sims (1980), the restrictions to be imposed ensure that $\mathbf{C}_0$ is a triangular matrix, this allows to use the Cholesky decomposition of $\mathbf{\Sigma}$ to obtain the non-zero elements of $\mathbf{C}_0$. This amount of restrictions account $n \times (n-1)/2$ and make the model just identifiable.

In econometrics the use of the Cholesky decomposition with identifying impact restrictions is also reffered to as recursive identification. This is because the procedure implies a recursive effect of the shocks over the variables, thus making the order in which the variables appear in the model matter for the interpretation of the results. Since the matrix $\mathbf{C}_0$ is restricted to be triangular, e.g. lower triangular, the first variable can only be affected at impact by the first shock (ϵ first element), whereas the second variable can be affected at impact by both the first and second shocks. This is better illustrated in Christiano, Eichenbaum & Evans (1999) where the recursive identification is applied to determine the effects of a monetary policy shock.

Once $\mathbf{C}_0$ is known, equations (8) and (9) can be used to calculate $\mathbf{C}_i$ for all i:

$$\mathbf{C}_i = \mathbf{\Psi}_i \mathbf{C}_0 \tag{11}$$

Identification by impact restrictions is summarized in Algorithm 1.

3.2. Identification by Long Run Restrictions

Another way to identify the matrices of the structural VMA-X is to impose restrictions on the long run impact of the shocks over the endogenous variables. This method is proposed in Blanchard & Quah (1989). For the model under consideration, if the variables in $\mathbf{y}_t$ are in differences, the matrix $\mathbf{C}(1) = \sum_{i=0}^{\infty} \mathbf{C}_i$ measures the long run impact of the structural shocks over the levels of the variables.[4] Matrix $\mathbf{C}(1)$ is obtained by evaluating equation (7) in $L = 1$. As in the case of impact restrictions, the variance of each side of the equation is taken, the result is:

$$\mathbf{C}(1)\,\mathbf{C}^{'}(1) = \boldsymbol{\Psi}(1)\,\boldsymbol{\Sigma}\boldsymbol{\Psi}^{'}(1) \tag{12}$$

Again, since $\boldsymbol{\Psi}(1)\,\boldsymbol{\Sigma}\boldsymbol{\Psi}^{'}(1)$ is a symmetric, positive definite matrix it is not possible to infer the parameters of $\mathbf{C}(1)$ from equation (12), restrictions over the parameters of $\mathbf{C}(1)$ have to be imposed. It is conveniently assumed that those restrictions make $\mathbf{C}(1)$ a triangular matrix, as before, this allows to use the Cholesky decomposition to calculate the non-zero elements of $\mathbf{C}(1)$. Again, this amount of restrictions account $n \times (n-1)/2$ and make the model just identifiable. It is important to note that the ordering of the variables matters as before. If, for example, $\mathbf{C}(1)$ is lower triangular, the first shock will be the only one that can have long run effects over the first variable, whereas the second variable can be affected by both the first and second shock in the long run.

Finally, it is possible to use $\mathbf{C}(1)$ to calculate the parameters in the $\mathbf{C}_0$ matrix, with it, the matrices $\mathbf{C}_i$ for $i > 0$ are obtained as in the identification by impact restrictions. Combining (10) with (7) evaluated in $L = 1$ the following expression for $\mathbf{C}_0$ is derived:

$$\mathbf{C}_0 = [\boldsymbol{\Psi}(1)]^{-1}\,\mathbf{C}(1) \tag{13}$$

Identification by long run restrictions is summarized in Algorithm 2.

4. Estimation

The estimation of the parameters of the VAR-X can be carried out by classical or Bayesian methods, as will become clear it is convenient to write the model in a more compact form. Following Zellner (1996) and Bauwens, Lubrano & Richard (2000), equation (1), for a sample of T observations, plus a fixed presample, can be written as:

$$\mathbf{Y} \quad = \quad \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{E} \tag{14}$$

where $\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^{'} \\ \vdots \\ \mathbf{y}_t^{'} \\ \vdots \\ \mathbf{y}_T^{'} \end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix} 1 & \mathbf{y}_0^{'} & \dots & \mathbf{y}_{1-p}^{'} & \mathbf{x}_1^{'} & \dots & \mathbf{x}_{1-q}^{'} \\ \vdots & & & & & & \\ 1 & \mathbf{y}_{t-1}^{'} & \dots & \mathbf{y}_{t-p}^{'} & \mathbf{x}_t^{'} & \dots & \mathbf{x}_{t-q}^{'} \\ \vdots & & & & & & \\ 1 & \mathbf{y}_{T-1}^{'} & \dots & \mathbf{y}_{T-p}^{'} & \mathbf{x}_T^{'} & \dots & \mathbf{x}_{T-q}^{'} \end{bmatrix}$, $\mathbf{E} = \begin{bmatrix} \mathbf{e}_1^{'} \\ \vdots \\ \mathbf{e}_t^{'} \\ \vdots \\ \mathbf{e}_T^{'} \end{bmatrix}$

and $\boldsymbol{\Gamma} = \begin{bmatrix} \mathbf{v} & \mathbf{B}_1 & \dots & \mathbf{B}_p & \boldsymbol{\Theta}_o & \dots & \boldsymbol{\Theta}_q \end{bmatrix}^{'}$.

[4] Of course, not all the variables of $\mathbf{y}_t$ must be in differences, but the only meaningful restrictions are those imposed over variables that enter the model in that way. We restrict our attention to a case in which there are no variables in levels in $\mathbf{y}_t$.

Algorithm 2 Identification by Long Run Restrictions

1. Estimate the reduced form of the VAR-X.
2. Calculate the VMA-X representation of the model (matrices $\mathbf{\Psi}_i$) and the covariance matrix of the reduced form disturbances $\mathbf{e}$ (matrix $\mathbf{\Sigma}$).
3. From the Cholesky decomposition of $\mathbf{\Psi}(1)\mathbf{\Sigma}\mathbf{\Psi}'(1)$ calculate matrix $\mathbf{C}(1)$.
$$\mathbf{C}(1) = \text{chol}\left(\mathbf{\Psi}(1)\mathbf{\Sigma}\mathbf{\Psi}'(1)\right)$$
4. With the matrices of long run effects of the reduced form, $\mathbf{\Psi}(1)$, and structural shocks, $\mathbf{C}(1)$, calculate the matrix of contemporaneous effects of the structural shocks, $\mathbf{C}_0$.
$$\mathbf{C}_0 = [\mathbf{\Psi}(1)]^{-1}\mathbf{C}(1)$$
5. For $i = 1, \ldots, R$, with R sufficiently large, calculate the matrices $\mathbf{C}_i$ as:
$$\mathbf{C}_i = \mathbf{\Psi}_i\mathbf{C}_0$$

Identification is completed since all matrices of the structural VMA-X are known.

For convenience we define the auxiliary variable $k = (1 + np + m(q+1))$ as the total number of regressors. The matrices sizes are as follow: $\mathbf{Y}$ is a $T \times n$ matrix, $\mathbf{Z}$ a $T \times k$ matrix, $\mathbf{E}$ a $T \times n$ matrix and $\mathbf{\Gamma}$ a $k \times n$ matrix.

Equation (14) is useful because it allows to represent the VAR-X model as a multivariate linear regression model, with it the likelihood function is derived. The parameters can be obtained by maximizing that function or by means of Bayes theorem.

4.1. The Likelihood Function

From equation (14) one derives the likelihood function for the error terms. Since $\mathbf{e}_t \sim N(\mathbf{0}, \mathbf{\Sigma})$, one has: $\mathbf{E} \sim \text{MN}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I})$, a matricvariate normal distribution with $\mathbf{I}$ the identity matrix with dimension $T \times T$. The following box defines the probability density function for the matricvariate normal distribution.

The Matricvariate Normal Distribution. The probability density function of a $(p \times q)$ matrix $\mathbf{X}$ that follows a matricvariate normal distribution with mean $\mathbf{M}_{p\times q}$ and covariance matrix $\mathbf{Q}_{q\times q} \otimes \mathbf{P}_{p\times p}$ $(\mathbf{X} \sim \mathrm{MN}(\mathbf{M}, \mathbf{Q} \otimes \mathbf{P}))$ is:

$$\mathrm{MN}_{\text{pdf}} \propto |\mathbf{Q} \otimes \mathbf{P}|^{-1/2} \exp\left(-1/2 \left[\operatorname{vec}(\mathbf{X}-\mathbf{M})\right]' (\mathbf{Q} \otimes \mathbf{P})^{-1} \left[\operatorname{vec}(\mathbf{X}-\mathbf{M})\right]\right) \tag{15}$$

Following Bauwens et al. (2000), the vec operator can be replaced by a trace operator (tr):

$$\mathrm{MN}_{\text{pdf}} \propto |\mathbf{Q}|^{-p/2} |\mathbf{P}|^{-q/2} \exp\left(-1/2 \operatorname{tr}\left(\mathbf{Q}^{-1} (\mathbf{X}-\mathbf{M})' \mathbf{P}^{-1} (\mathbf{X}-\mathbf{M})\right)\right) \tag{16}$$

Both representations of the matricvariate normal pdf are useful when dealing with the compact representation of the VAR-X model. Note that the equations above are only proportional to the actual probability density function. The missing constant term has no effects in the estimation procedure.

Using the definition in the preceding box and applying it to $\mathbf{E} \sim \mathrm{MN}(\mathbf{O}, \mathbf{\Sigma} \otimes \mathbf{I})$ one gets the likelihood function of the VAR-X model, conditioned to the path of the exogenous variables:

$$\mathcal{L} \propto |\mathbf{\Sigma}|^{-T/2} \exp\left(-1/2 \operatorname{tr}\left(\mathbf{\Sigma}^{-1} \mathbf{E}' \mathbf{E}\right)\right)$$

From (14) one has $\mathbf{E} = \mathbf{Y} - \mathbf{Z\Gamma}$, replacing:

$$\mathcal{L} \propto |\mathbf{\Sigma}|^{-T/2} \exp\left(-1/2 \operatorname{tr}\left(\mathbf{\Sigma}^{-1} (\mathbf{Y}-\mathbf{Z\Gamma})' (\mathbf{Y}-\mathbf{Z\Gamma})\right)\right)$$

Finally, after tedious algebraic manipulation, one gets to the following expression:

$$\begin{aligned}\mathcal{L} \propto & \left[|\mathbf{\Sigma}|^{-(T-k)/2} \exp\left(-1/2 \operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{S}\right)\right)\right] \\ & \left[|\mathbf{\Sigma}|^{-k/2} \exp\left(-1/2 \operatorname{tr}\left(\mathbf{\Sigma}^{-1} \left(\mathbf{\Gamma}-\widehat{\mathbf{\Gamma}}\right)' \mathbf{Z}'\mathbf{Z} \left(\mathbf{\Gamma}-\widehat{\mathbf{\Gamma}}\right)\right)\right)\right]\end{aligned}$$

where $\widehat{\mathbf{\Gamma}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \mathbf{Z}'\mathbf{Y}$ and $\mathbf{S} = \left(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Gamma}}\right)' \left(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Gamma}}\right)$. It is being assumed overall that matrix $\mathbf{Z}'\mathbf{Z}$ is invertible, a condition common to the VAR and OLS models (see Lütkepohl (2005) section 3.2).

One last thing is noted, the second factor of the right hand side of the last expression is proportional to the pdf of a matricvariate normal distribution for $\mathbf{\Gamma}$, and the first factor to the pdf of an inverse Wishart distribution for $\mathbf{\Sigma}$ (see the box below). This allows an exact characterization of the likelihood function as:

$$\mathcal{L} = \mathrm{iW}_{\text{pdf}}(\mathbf{S}, T-k-n-1)\, \mathrm{MN}_{\text{pdf}}\left(\widehat{\mathbf{\Gamma}}, \mathbf{\Sigma} \otimes \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right) \tag{17}$$

where $\mathrm{iW}_{\text{pdf}}(\mathbf{S}, T-k-n-1)$ stands for the pdf of an inverse Wishart distribution with parameters $\mathbf{S}$ and $T-k-n-1$.

The parameters of the VAR-X, $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$, can be estimated by maximizing equation (17). It can be shown that the result of the likelihood maximization gives:

$$\mathbf{\Gamma}_{ml} = \widehat{\mathbf{\Gamma}} \qquad\qquad \mathbf{\Sigma}_{ml} = \mathbf{S}$$

Sometimes because practical considerations or non-invertibility of $\mathbf{Z}^{'}\mathbf{Z}$, when no restrictions are imposed, equation by equation estimation can be implemented (see Lütkepohl (2005) section 5.4.4).

The Inverse Wishart Distribution
If the variable $\mathbf{X}$ (a square, positive definite matrix of size q) is distributed $\mathrm{iW}\left(\mathbf{S}, s\right)$, with parameter $\mathbf{S}$ (also a square, positive definite matrix of size q), and s degrees of freedom, then its probability density function $\left(\mathrm{iW}_{\mathrm{pdf}}\right)$ is given by:

$$\mathrm{iW}_{\mathrm{pdf}}\left(\mathbf{S}, s\right) = \frac{\left|\mathbf{S}\right|^{\frac{s}{2}}}{2^{\frac{vq}{2}}\Gamma_q\left(\frac{s}{2}\right)} \left|\mathbf{X}\right|^{\frac{-(s+q+1)}{2}} \exp\left(-1/2\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{S}\right)\right) \tag{18}$$

where $\Gamma_q\left(x\right) = \pi^{\frac{q(q-1)}{4}} \prod\limits_{j=1}^{q} \Gamma\left(x + \frac{1-j}{2}\right)$ is the multivariate Gamma function. It is useful to have an expression for the mean and mode of the inverse Wishart distribution, these are given by:

$$\mathrm{Mean}\left(\mathbf{X}\right) = \frac{S}{s-q-1} \qquad\qquad \mathrm{Mode}\left(\mathbf{X}\right) = \frac{\mathbf{S}}{s+q+1}$$

4.2. Bayesian Estimation

If the estimation is carried out by Bayesian methods the problem is to elect an adequate prior distribution and, by means of Bayes theorem, obtain the posterior density function of the parameters. The use of Bayesian methods is encouraged because they allow inference to be done conditional to the sample, and in particular the sample size, giving a better sense of the uncertainty associated with the parameters values; it also facilitate to compute moments not only for the parameters but for their functions as is the case of the impulse responses, forecast error variance decomposition and others; it is also particularly useful to obtain a measure of skewness in this functions, specially for the policy implications of the results. As mentioned in Koop (1992), the use of Bayesian methods gives an exact finite sample density for both the parameters and their functions.

The election of the prior is a sensitive issue and will not be discussed in this document, we shall restrict our attention to the case of the Jeffreys non-informative prior (Jeffreys 1961) which is widely used in Bayesian studies of vector autoregressors. There are usually two reasons for its use. The first one is that information about the reduced form parameters of the VAR-X model is scarce and difficult to translate into an adequate prior distribution. The second is that it might be the case that the econometrician does not want to include new information to the estimation but only wishes to use Bayesian methods for inference purposes. Besides the two reasons already mentioned, the use of the Jeffreys noninformative prior constitute a computational advantage because it allows a closed form representation of the posterior density function, thus allowing to make draws for the parameters by direct methods or by the Gibbs sampling algorithm (Geman & Geman 1984).[5]

For a discussion of other usual prior distributions for VAR models we refer to Kadiyala & Karlsson (1997) and, more recently, to Kociecki (2010) for the construction of feasible prior distributions over impulse response in a structural VAR context. When the model is used for forecast purposes the so called Minnesota prior is of particular interest, this prior is due to Litterman (1986), and is generalized in Kadiyala & Karlsson (1997) for allowing symmetry of the prior across equations. This generalization is recommended and is of easy implementation in the Bayesian estimation of the model. It should be mentioned that the Minnesota prior is of little interest in the structural VAR-X context, principally because the model is conditioned to the path of the exogenous variables, adding difficulties to the forecasting process.

In general the Jeffreys Prior for the linear regression parameters correspond to a constant for the parameters in $\mathbf{\Gamma}$ and for the covariance matrix a function of the form: $|\mathbf{\Sigma}|^{\frac{-(n+1)}{2}}$, where n represents the size of the covariance matrix. The prior distribution to be used is then:

$$P\left(\mathbf{\Gamma},\mathbf{\Sigma}\right) = \overline{C}\left|\mathbf{\Sigma}\right|^{\frac{-(n+1)}{2}} \tag{19}$$

where $\overline{C}$ is the integrating constant of the distribution. Its actual value will be of no interest.

The posterior is obtained from Bayes theorem as:

$$\pi\left(\mathbf{\Gamma},\mathbf{\Sigma} \mid \mathbf{Y},\mathbf{Z}\right) = \frac{\mathcal{L}\left(\mathbf{Y},\mathbf{Z} \mid \mathbf{\Gamma},\mathbf{\Sigma}\right) P\left(\mathbf{\Gamma},\mathbf{\Sigma}\right)}{m\left(\mathbf{Y}\right)} \tag{20}$$

where $\pi\left(\mathbf{\Gamma},\mathbf{\Sigma} \mid \mathbf{Y},\mathbf{Z}\right)$ is the posterior distribution of the parameters given the data, $\mathcal{L}\left(\mathbf{Y},\mathbf{Z} \mid \mathbf{\Gamma},\mathbf{\Sigma}\right)$ is the likelihood function, $P\left(\mathbf{\Gamma},\mathbf{\Sigma}\right)$ is the prior distribution of the parameters and $m\left(\mathbf{Y}\right)$ the marginal density of the model. The value and use of the marginal density is discussed in Section 4.2.1.

Combining equations (17), (19) and (20) one gets an exact representation of the posterior function as the product of the pdf of an inverse Wishart distribution and the pdf of a matricvariate normal distribution:

$$\pi\left(\mathbf{\Gamma},\mathbf{\Sigma} \mid \mathbf{Y},\mathbf{Z}\right) = \mathrm{iW}_{\mathrm{pdf}}\left(\mathbf{S}, T-k\right) \mathrm{MN}_{\mathrm{pdf}}\left(\widehat{\mathbf{\Gamma}}, \mathbf{\Sigma} \otimes \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right) \tag{21}$$

Equation (21) implies that $\mathbf{\Sigma}$ follows an inverse Wishart distribution with parameters $\mathbf{S}$ and $T-k$, and that the distribution of $\mathbf{\Gamma}$ given $\mathbf{\Sigma}$ is matricvariate

[5] For an introduction to the use of the Gibbs sampling algorithm we refer to Casella & George (1992).

Algorithm 3 Bayesian Estimation

1. Select the specification for the reduced form VAR-X, that is to chose values of p (endogenous variables lags) and q (exogenous variables lags) such that the residuals of the VAR-X ($\mathbf{e}$) have white noise properties. With this the following variables are obtained: T, p, q, k, where:
$$k = 1 + np + m\,(q+1)$$
2. Calculate the values of $\hat{\mathbf{\Gamma}}$, $\mathbf{S}$ with the data $(\mathbf{Y}, \mathbf{Z})$ as:
$$\hat{\mathbf{\Gamma}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y} \qquad\qquad \mathbf{S} = \left(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{\Gamma}}\right)'\left(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{\Gamma}}\right)$$
3. Generate a draw for matrix $\mathbf{\Sigma}$ from an inverse Wishart distribution with parameter $\mathbf{S}$ and $T-k$ degrees of freedom.
$$\mathbf{\Sigma} \sim \mathrm{iW}_{\mathrm{pdf}}\,(\mathbf{S}, T-k)$$
4. Generate a draw for matrix $\mathbf{\Gamma}$ from a matricvariate normal distribution with mean $\hat{\mathbf{\Gamma}}$ and covariance matrix $\mathbf{\Sigma} \otimes \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}$.
$$\mathbf{\Gamma}|\mathbf{\Sigma} \sim \mathrm{MN}_{\mathrm{pdf}}\left(\hat{\mathbf{\Gamma}}, \mathbf{\Sigma} \otimes \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right)$$
5. Repeat steps 2-3 as many times as desired, save the values of each draw.

The draws generated can be used to compute moments of the parameters. For every draw the corresponding structural parameters, impulse responses functions, etc. can be computed, then, their moments and statistics can also be computed. The algorithms for generating draws for the inverse Wishart and matricvariate normal distributions are presented in Bauwens et al. (2000), Appendix B.

normal with mean $\widehat{\mathbf{\Gamma}}$ and covariance matrix $\mathbf{\Sigma} \otimes \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}$. The following two equations formalize the former statement:

$$\mathbf{\Sigma} \mid \mathbf{Y}, \mathbf{Z} \sim \mathrm{iW}_{\mathrm{pdf}}\,(\mathbf{S}, T-k) \qquad\qquad \mathbf{\Gamma} \mid \mathbf{\Sigma}, \mathbf{Y}, \mathbf{Z} \sim \mathrm{MN}_{\mathrm{pdf}}\left(\widehat{\mathbf{\Gamma}}, \mathbf{\Sigma} \otimes \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right)$$

Although further work can be done to obtain the unconditional distribution of $\mathbf{\Gamma}$ it is not necessary to do so. Because equation (21) is an exact representation of the parameters distribution function, it can be used to generate draws of them, moreover it can be used to compute any moment or statistic of interest, this can be done by means of the Gibbs sampling algorithm.

4.2.1. Marginal Densities and Lag Structure

The marginal density ($m\left(\mathbf{Y}\right)$) can be easily obtained under the Jeffreys prior and can be used afterward for purposes of model comparison. The marginal density gives the probability that the data is generated by a particular model, eliminating the uncertainty due to the parameters values. Because of this $m\left(\mathbf{Y}\right)$ is often used for model comparison by means of the Bayes factor (BF): the ratio between the marginal densities of two different models that explain the same set of data ($\text{BF}_{12} = {}^{m(Y|\mathcal{M}_1)}/{}_{m(\mathbf{Y}|\mathcal{M}_2)}$). If the Bayes factor is bigger than one then the first model ($\mathcal{M}_1$) would be preferred.

From Bayes theorem (equation 20) the marginal density of the data, given the model, is:

$$m\left(\mathbf{Y}\right) = \frac{\mathcal{L}\left(\mathbf{Y},\mathbf{Z}|\mathbf{\Gamma},\mathbf{\Sigma}\right) P\left(\mathbf{\Gamma},\mathbf{\Sigma}\right)}{\pi\left(\mathbf{\Gamma},\mathbf{\Sigma}|\mathbf{Y},\mathbf{Z}\right)} \tag{22}$$

its value is obtained by replacing for the actual forms of the likelihood, prior and posterior functions (equations 17, 19 and 21 respectively):

$$m\left(\mathbf{Y}\right) = \frac{\Gamma_n\left(\frac{T-k}{2}\right)}{\Gamma_n\left(\frac{T-k-n-1}{2}\right)} \left|\mathbf{S}\right|^{\frac{-n-1}{2}} 2^{\frac{n(n+1)}{2}} \overline{C} \tag{23}$$

Although the exact value of the marginal density for a given model cannot be known without the constant $\overline{C}$, this is no crucial for model comparison if the only difference between the models is in their lag structure. In that case the constant $\overline{C}$ is the same for both models, and the difference between the marginal density of one specification or another arises only in the first two factors of the right hand side of equation (23) $\left[\frac{\Gamma_n\left(\frac{T-k}{2}\right)}{\Gamma_n\left(\frac{T-k-n-1}{2}\right)} \left|\mathbf{S}\right|^{\frac{-n-1}{2}}\right]$. When computing the Bayes factor for any pair of models the result will be given by those factors alone.

The Bayes factor between a model, $\mathcal{M}_1$, with k_1 regressors and residual covariance matrix $\mathbf{S}_1$, and another model, $\mathcal{M}_2$, with k_2 regressors and residual covariance matrix $\mathbf{S}_2$, can be reduced to:

$$\text{BF}_{12} = \frac{\frac{\Gamma_n\left(\frac{T-k_1}{2}\right)}{\Gamma_n\left(\frac{T-k_1-n-1}{2}\right)} \left|\mathbf{S}_1\right|^{\frac{-n-1}{2}}}{\frac{\Gamma_n\left(\frac{T-k_2}{2}\right)}{\Gamma_n\left(\frac{T-k_2-n-1}{2}\right)} \left|\mathbf{S}_2\right|^{\frac{-n-1}{2}}} \tag{24}$$

5. Applications

There are several applications for the structural VAR-X, all of them useful for policy analysis. In this Section four of those applications are covered, they all use the structural VMA-X representation of the model (equation 4).

5.1. Impulse Response Functions (IRF), Multiplier Analysis (MA), and Forecast Error Variance Decomposition (FEVD)

Impulse response functions (IRF) and multiplier analysis (MA) can be constructed from the matrices in $\mathbf{C}(L)$ and $\mathbf{\Lambda}(L)$. The IRF shows the endogenous variables response to a unitary change in a structural shock, in an analogous way the MA shows the response to a change in an exogenous variable. The construction is simple and is based on the interpretations of the elements of the matrices in $\mathbf{C}(L)$ and $\mathbf{\Lambda}(L)$.

For the construction of the IRF consider matrix $\mathbf{C}_h$. The elements of this matrix measure the effect of the structural shocks over the endogenous variables h periods ahead, thus c_h^{ij} (i-throw, j-th column) measures the response of the i-th variable to a unitary change in the j-th shock h periods ahead. The IRF for the i-th variable to a change in j-th shock is constructed by collecting elements c_h^{ij} for $h = 0, 1, \ldots, H$, with H the IRF horizon.

Matrices $\mathbf{C}_h$ are obtained from the reduced form parameters according to the type of identification (Section 3). For a more detailed discussion on the construction and properties of the IRF we refer to Lütkepohl (2005), Section 2.3.2.

The MA is obtained similarly from matrices $\mathbf{\Lambda}_h$, which are also a function of the reduced form parameters.[6] The interpretation is the same as before.

A number of methods for inference over the IRF and MA are available. If the estimation is carried out by classical methods intervals for the IRF and MA can be computed by means of their asymptotic distributions or by bootstrapping methods.[7] Nevertheless, because the OLS estimators are biased, as proved in Nicholls & Pope (1988), the intervals that arise from both asymptotic theory and usual bootstrapping methods are also biased. As pointed out by Kilian (1998) this makes necessary to conduct the inference over IRF, and in this case over MA, correcting the bias and allowing for skewness in the intervals. Skewness is common in the small sample distributions of the IRF and MA and arises from the non-linearity of the function that maps the reduced form parameters to the IRF or MA. A double bootstrapping method that effectively corrects the bias and accounts for the skewness in the intervals is proposed in Kilian (1998).

In the context of Bayesian estimation, it is noted that, applying Algorithm 1 or 2 for each draw of the reduced form parameters (Algorithm 3), the distribution for each c_h^{ij} and λ_h^{ij} is obtained. With the distribution function inference can be done over the point estimate of the IRF and MA. For instance, standard deviations in each horizon can be computed, as well as asymmetry measures and *credible sets* (or intervals), the Bayesian analogue to a classical confidence interval.

In the following we shall restrict our attention to *credible sets* with minimum size (length), these are named Highest Posterior Density regions (HPD from now on). An $(1-\alpha)\,\%$ HPD for the parameter θ is defined as the set $\mathcal{I} =$

[6] See Lütkepohl (2005), Section 10.6.

[7] The asymptotic distribution of the IRF and FEVD for a VAR is presented in Lütkepohl (1990). A widely used non-parametric bootstrapping method is developed in Runkle (1987).

$\{\theta \in \mathbf{\Theta} : \pi\left({}^{\theta}\!/\!{}_{\mathbf{Y}}\right) \geq k(\alpha)\}$, where $k(\alpha)$ is the largest constant satisfying $P(\mathcal{I}|y) = \int_{\theta} \pi\left({}^{\theta}\!/\!{}_{\mathbf{Y}}\right) d\theta \geq 1-\alpha$.[8] From the definition just given is clear that HPD regions are of minimum size and that each value of $\theta \in \mathcal{I}$ has a higher density (probability) than any value of θ outside the HPD. The second property makes possible direct probability statements about the likelihood of θ falling in $\mathcal{I}$, i.e., "The probability that θ lies in $\mathcal{I}$ given the observed data $\mathbf{Y}$ is at least $(1-\alpha)\%$", this contrast with the interpretation of the classical confidence intervals. An HPD region can be disjoint if the posterior density function $\left(\pi\left({}^{\theta}\!/\!{}_{\mathbf{Y}}\right)\right)$ is multimodal. If the posterior is symmetric, all HPD regions will be symmetric about posterior mode (mean).

Koop (1992) presents a detailed revision of how to apply Bayesian inference to the IRF in a structural VAR context, his results can be easily adapted to the structural VAR-X model. Another reference on the inference over IRF is Sims & Zha (1999). Here we present, in Algorithm 4, the method of Chen & Shao (1998) for computing HPD regions from the output of the Gibbs sampler.[9]

It is important to note that Bayesian methods are by nature conditioned to the sample size and, because of that, avoid the problems of asymptotic theory in explaining the finite sample properties of the parameters functions, this includes the skewness of the IRF and MA distribution functions. Then, if the intervals are computed with the HPD, as in Chen & Shao (1998), they would be taking into account the asymmetry in the same way as Kilian's method. This is not the case for intervals computed using only standard deviations although, with them, skewness can be addressed as in Koop (1992), although bootstrap methods can be used to calculate approximate measures of this and others moments, for instance, skewness and kurtosis, Bayesian methods are preferable since exact measures can be calculated.

Another application of the structural VAR-X model is the forecast error variance decomposition (FEVD), this is no different to the one usually presented in the structural VAR model. FEVD consists in decomposing the variance of the forecast error of each endogenous variable h periods ahead, as with the IRF, the matrices of $C(L)$ are used for its construction. Note that, since the model is conditioned to the path of the exogenous variables, all of the forecast error variance is explained by the structural shocks. Is because of this that the FEVD has no changes when applied in the structural VAR-X model. We refer to Lütkepohl (2005), Section 2.3.3, for the details of the construction of the FEVD. Again, if Bayesian methods are used for the estimation of the VAR-X parameters, the density function of the FEVD can be obtained and several features of it can be explored, Koop (1992) also presents how to apply Bayesian inference in this respect.

[8] Integration can be replaced by summation if θ is discrete.

[9] The method presented is only valid if the distribution of the parameters of interest is unimodal. For a more general treatment of the highest posterior density regions, including multimodal distributions, we refer to the work of Hyndman (1996).

Algorithm 4 Highest Posterior Density Regions

As in Chen & Shao (1998), let $\left\{\theta^{(i)},\, i = 1\,,\,\ldots\,,\, N\right\}$ be an ergodic sample of $\pi\left({}^{\theta}/_{\mathbf{Y}}\right)$, the posterior density function of parameter θ. $\pi\left({}^{\theta}/_{\mathbf{Y}}\right)$ is assumed to be unimodal. The $(1-\alpha)\,\%$ HPD is computed as follows:

1. Sort the values of $\theta^{(i)}$. Define $\theta_{(j)}$ as the $j - th$ larger draw of the sample, so that:

$$\theta_{(1)} = \min_{i \in \{1,\ldots,N\}} \left\{\theta^{(i)}\right\} \qquad\qquad \theta_{(N)} = \max_{i \in \{1,\ldots,N\}} \left\{\theta^{(i)}\right\}$$

2. Define $\overline{N} = \lfloor (1-\alpha)\, N \rfloor$ the integer part of $(1-\alpha)\, N$. The HPD will contain $\overline{N}$ values of θ.

3. Define $\mathcal{I}_{(j)} = \left(\theta_{(j)}\,,\, \theta_{\left(j+\overline{N}\right)}\right)$ an interval in the domain of the parameter θ, for $j \epsilon \left\{1,\ldots, N - \overline{N}\right\}$. Note that although $\mathcal{I}_{(j)}$ contains always $\overline{N}$ draws of θ, its size may vary.

4. The HPD is obtained as the interval $\mathcal{I}_{(j)}$ with minimum size. $\text{HPD}\,(\alpha) = \mathcal{I}_{(j^\star)}$, with $j^\star$ such that:

$$\theta_{\left(j^\star+\overline{N}\right)} - \theta_{(j^\star)} = \min_{j \in \left\{1,\ldots,N-\overline{N}\right\}} \left(\theta_{\left(j+\overline{N}\right)} - \theta_{(j)}\right)$$

5.2. Historical Decomposition of the Endogenous Variables (HD)

The historical decomposition (HD) consists in explaining the observed values of the endogenous variables in terms of the structural shocks and the path of the exogenous variables. This kind of exercise is present in the DSGE literature (for example, in Smets & Wouters (2007)) but mostly absent in the structural VAR literature. There are nonotheless various exceptions, an early example is the work of Burbidge & Harrison (1985) on the role of money in the great depression, there is also the textbook by Canova (2007), and the paper of King & Morley (2007), where the historical decomposition of a structural VAR is used for computing a measure of the natural rate of unemployment for the US.

Unlike the applications already presented, the historical decomposition allows to make a statement over what has actually happened to the series in the sample period, in terms of the recovered values for the structural shocks and the observed paths of the exogenous variables. It allows to have all shocks and exogenous variables acting simultaneously, thus making possible the comparison over the relative effects of them over the endogenous variables, this means that the HD is particularly useful when addressing the relative importance of the shocks over some set of variables. The possibility of explaining the history of the endogenous variables instead of what would happen if some hypothetical shock arrives in the absence of any other disturbance is at least appealing.

Here we describe a method for computing the HD in a structural VAR and structural VAR-X context. The first case is covered in more detail and the second presented as an extension of the basic ideas.

5.2.1. Historical Decomposition for a Structural VAR Model

In a structural VAR context is clear, from the structural VMA representation of the model, that variations of the endogenous variables can only be explained by variations in the structural shocks. The HD uses the structural VMA representation in order to compute what the path of each endogenous variable would have been conditioned to the presence of only one of the structural shocks. It is important to note that the interpretation of the HD in a stable VAR model is simpler than the interpretation in a VAR-X. This is because in the former there is no need for a reference value that indicates when a shock is influencing the path of the variables. In that case, the reference value is naturally zero, and it is understood that deviations of the shocks below that value are interpreted as negative shocks and deviations above as positive shocks. As we shall see, when dealing with exogenous variables a reference value must be set, and its election is not necessarily “natural”.

Before the HD is computed it is necessary to recover the structural shocks from the estimation of the reduced form VAR. Define $\widehat{\mathbf{E}} = [\widehat{\mathbf{e}}_1 \ldots \widehat{\mathbf{e}}_t \ldots \widehat{\mathbf{e}}_T]^{'}$ as the matrix of all fitted residuals from the VAR model (equation (14) in the absence of exogenous variables). Recalling equation (9), the matrix $\mathbf{C}_0$ can be used to recover the structural shocks from matrix $\widehat{\mathbf{E}}$ as in the following expression:

$$\widehat{\mathcal{E}} = \widehat{\mathbf{E}} \left(\mathbf{C}_0^{'}\right)^{-1} \tag{25}$$

Because zero is the reference value for the structural shocks the matrix $\widehat{\mathcal{E}} = [\widehat{\epsilon}_1 \ldots \widehat{\epsilon}_t \ldots \widehat{\epsilon}_T]^{'}$ can be used directly for the HD.

The HD is an in-sample exercise, thus is conditioned to the initial values of the series. It will be useful to define the structural infinite VMA representation of the VAR model, as well as the structural VMA representation conditional on the initial values of the endogenous variables, equations (26) and (27) respectively.

$$\mathbf{y}_t = \mu + \mathbf{C}\left(L\right)\epsilon_t \tag{26}$$

$$\mathbf{y}_t = \sum_{i=0}^{t-1} \mathbf{C}_i \epsilon_{t-i} + \mathbf{K}_t \tag{27}$$

Note that in equation (26) the endogenous variables depend on an infinite number of past structural shocks. In equation (27) the effect of all shocks that are realized previous to the sample is captured by the initial values of the endogenous variables. The variable $\mathbf{K}_t$ is a function of those initial values and of the parameters of the reduced form model, $\mathbf{K}_t = f_t\left(\mathbf{y}_0\,, \ldots\,, \mathbf{y}_{-(p-1)}\right)$. It measures the effect of the initial values over the period t realization of the endogenous variables, thus the effect of all shocks that occurred before the sample. It is clear that if the VAR is stable $\mathbf{K}_t \longrightarrow \mu$ for t sufficiently large, this is because the shocks that are too far in the past have no effect in the current value of the variables. $\mathbf{K}_t$ will be refer to as the reference value of the historical decomposition.

Starting from the structural VMA representation, the objective is now to decompose the deviations of $\mathbf{y}_t$ from $\mathbf{K}_t$ into the effects of the current and past values of the structural shocks (ϵ_i for i from 1 to t). The decomposition is made over the auxiliary variable $\widetilde{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{K}_t = \sum\limits_{i=0}^{t-1} \mathbf{C}_i \epsilon_{t-i}$. The information needed to compute $\widetilde{\mathbf{y}}_t$ is contained in the first t matrices $\mathbf{C}_i$ and the first t rows of matrix $\widehat{\mathcal{E}}$.

The historical decomposition of the i-th variable of $\widetilde{\mathbf{y}}_t$ into the j-th shock is given by:

$$\widetilde{y}_t^{(i,j)} = \sum_{i=0}^{t-1} c_i^{ij} \widehat{\epsilon}_{t-i}^{j} \tag{28}$$

Note that it must hold that the sum over j is equal to the actual value of the i-th element of $\widetilde{\mathbf{y}}_t$, $\widetilde{y}_t^i = \sum\limits_{j=1}^{n} \widetilde{y}_t^{(i,j)}$. For t sufficiently large, when $\mathbf{K}_t$ is close to μ, $\widetilde{y}_t^{(i,j)}$ can be interpreted as the deviation of the i-th endogenous variable from its mean caused by the recovered sequence for the j-th structural shock.

Finally, the endogenous variables can be decomposed as well. The historical decomposition for the i-th endogenous variable into the j-th shock is given by:

$$y_t^{(i,j)} \quad = K_t^i + \widetilde{y}_t^{(i,j)} = \quad K_t^i + \sum_{i=0}^{t-1} c_i^{ij} \widehat{\epsilon}_{t-i}^{j} \tag{29}$$

the new variable $y_t^{(i,j)}$ is interpreted as what the i-th endogenous variable would have been if only realizations of the j-th shock had occurred. The value of $\mathbf{K}_t$ can be obtained as a residual of the historical decomposition, since $\mathbf{y}_t$ is known and $\widetilde{\mathbf{y}}_t$ can be computed from the sum of the HD or from the definition.

The HD of the endogenous variables ($y_t^{(i,j)}$) can be also used to compute what transformations of the variables would have been conditioned to the presence of only one shock. For instance, if the i-th variable enters the model in quarterly differences, the HD for the annual differences or the level of the series can be computed by applying to $y_t^{(i,j)}$ the same transformation used over y_t^i, in this example, a cumulative sum. Algorithm 5 summarizes the steps carried out for the historical decomposition.

5.2.2. Historical Decomposition for a Structural VAR-X Model

The structure already described applies also for a VAR-X model. The main difference is that now it is necessary to determine a reference value for the exogenous

Algorithm 5 Historical Decomposition for a Structural VAR Model

1. Estimate the parameters of the reduced form VAR.
   a) Save a matrix with all fitted residuals $\left(\hat{\mathbf{E}} = [\hat{\mathbf{e}}_1 \ldots \hat{\mathbf{e}}_t \ldots \hat{\mathbf{e}}_T]^{'}\right)$.
   b) Compute matrices $\mathbf{C}_i$ according to the identifying restrictions (Algorithm 1 or 2).
2. Compute the structural shocks $\left(\hat{\mathcal{E}} = [\hat{\epsilon}_1 \ldots \hat{\epsilon}_t \ldots \hat{\epsilon}_T]^{'}\right)$ with matrix $\mathbf{C}_0$ and the fitted residuals of the reduced form VAR:
$$\hat{\mathcal{E}} = \hat{\mathbf{E}}\left(\mathbf{C}_0^{'}\right)^{-1}$$
3. Compute the historical decomposition of the endogenous variables relative to $\mathbf{K}_t$:
$$\tilde{y}_t^{(i,j)} = \sum_{i=0}^{t-1} c_i^{ij} \hat{\epsilon}_{t-i}^{j}$$
4. Recover the values of $\mathbf{K}_t$ with the observed values of $\mathbf{y}_t$ and the auxiliary variable $\tilde{\mathbf{y}}_t$:
$$\mathbf{K}_t = \mathbf{y}_t - \tilde{\mathbf{y}}_t$$
5. Compute the historical decomposition of the endogenous variables:
$$y_t^{(i,j)} = K_t^i + \tilde{y}_t^{(i,j)}$$

Steps 3 and 5 are repeated for $t = 1, 2, \ldots, T$, $i = 1, \ldots, n$ and $j = 1, \ldots, n$. Step 4 is repeated for $t = 1, 2, \ldots, T$.

variables.[10] It shall be understood that realizations of the exogenous variables different to this value are what explain the fluctuations of the endogenous variables. We shall refer to $\overline{\mathbf{x}}_t$ as the reference value for the exogenous variables in t.

As before, it is necessary to present the structural VMA-X representation conditional to the initial values of the endogenous variables (equation 30), with $\mathbf{K}_t$ defined as above. It is also necessary to express the exogenous variables as deviations of the reference value, for this we define an auxiliary variable $\widetilde{\mathbf{x}}_t = \mathbf{x}_t - \overline{\mathbf{x}}_t$. Note that equation (30) can be written in terms of the new variable $\widetilde{\mathbf{x}}_t$ as in equation (31). In the latter, the new variable $\widetilde{\mathbf{K}}_t = \sum_{i=0}^{t-1} \mathbf{\Lambda}_i \bar{\mathbf{x}}_{t-i} + \mathbf{K}_t$ has a role

[10] The reference value for the exogenous variables need not be a constant. It can be given by a linear trend, by the sample mean of the series,or by the initial value. When the exogenous variables enter the model in their differences, it may seem natural to think in zero as a natural reference value, identifying fluctuations of the exogenous variables in an analogous way to whats done with the structural shocks.

analogous to that of $\mathbf{K}_t$ in the VAR context. $\widetilde{\mathbf{K}}_t$ properties depend on those of $\bar{\mathbf{x}}_t$ and, therefore, it can not be guaranteed that it converges to any value.

$$\mathbf{y}_t = \sum_{i=0}^{t-1} \mathbf{C}_i \epsilon_{t-i} + \sum_{i=0}^{t-1} \mathbf{\Lambda}_i \mathbf{x}_{t-i} + \mathbf{K}_t \tag{30}$$

$$\mathbf{y}_t = \sum_{i=0}^{t-1} \mathbf{C}_i \epsilon_{t-i} + \sum_{i=0}^{t-1} \mathbf{\Lambda}_i \widetilde{\mathbf{x}}_{t-i} + \widetilde{\mathbf{K}}_t \tag{31}$$

The historical decomposition is now computed using matrices $\mathbf{C}_i$, the recovered matrix of structural shocks $\widehat{\mathcal{E}}$, matrices $\mathbf{\Lambda}_i$ and the auxiliary variables $\widetilde{\mathbf{x}}_i$, for i from 1 to T. Matrix $\widehat{\mathcal{E}}$ is still computed as in equation (25). The new reference value for the historical decomposition is $\widetilde{\mathbf{K}}_t$, and the decomposition is done to explain the deviations of the endogenous variables with respect to it as a function of the structural shocks and deviations of the exogenous variables from their own reference value, $\bar{\mathbf{x}}_t$. For notational simplicity, variable $\widetilde{\mathbf{x}}_t$ is redefined: $\widetilde{\mathbf{y}}_t = \mathbf{y}_t - \widetilde{\mathbf{K}}_t = \sum_{i=0}^{t-1} \mathbf{C}_i \epsilon_{t-i} + \sum_{i=0}^{t-1} \mathbf{\Lambda}_i \widetilde{\mathbf{x}}_{t-i}$. The decomposition of the i-th variable of $\widetilde{\mathbf{y}}_t$ into the j-th shock is still given by equation (28), and the decomposition into the k-th exogenous variable is given by:

$$\widetilde{y}_t^{(i,k)} = \sum_{i=0}^{t-1} \lambda_i^{ik} \widetilde{x}_{t-i}^k \tag{32}$$

Variable $\widetilde{y}_t^{(i,k)}$, for k from 1 to m, is interpreted as what the variable $\widetilde{y}_t^i$ would have been if, in the absence of shocks, only the k-th exogenous variable is allowed to deviate from its reference value. As in the VAR model, the following equation holds: $\widetilde{y}_t^i = \sum_{j=1}^{n} \widetilde{y}_t^{(i,j)} + \sum_{k=1}^{m} \widetilde{y}_t^{(i,k)}$. The variable $\widetilde{\mathbf{K}}_t$ is recovered in the same way used before to recover $\mathbf{K}_t$.

The historical decomposition of the endogenous variables can be computed by using the recovered values for $\widetilde{\mathbf{K}}_t$. The decomposition of the i-th variable into the effects of the j-th shock is still given by equation (29), if K_t^i is replaced by $\widetilde{K}_t^i$. The decomposition of the i-th variable into the deviations of the k-th exogenous variable from its reference value is obtained from the following expression:

$$y_t^{(i,k)} = K_t^i + \widetilde{y}_t^{(i,k)} \tag{33}$$

Variable $y_t^{(i,k)}$ has the same interpretation as $\widetilde{y}_t^{(i,k)}$ but applied to the value of the endogenous variable, and not to the deviation from the reference value.

Although the interpretation and use of the HD in exogenous variables may seem strange and impractical, it is actually of great utility when the reference value for the exogenous variables is chosen correctly. The following example describes a case in which the interpretation of the HD in exogenous variables is more easily understood. Consider the case in which the exogenous variables are introduced

Algorithm 6 Historical Decomposition for a Structural VAR-X Model

1. Estimate the parameters of the reduced form VAR-X.

   a) Save a matrix with all fitted residuals $\left(\hat{\mathbf{E}} = [\hat{\mathbf{e}}_1 \dots \hat{\mathbf{e}}_t \dots \hat{\mathbf{e}}_T]^{'}\right)$.

   b) Compute matrices $\mathbf{C}_i$ and $\mathbf{\Lambda}_i$ according to the identifying restrictions (Algorithm 1 or 2).

2. Compute the structural shocks $\left(\hat{\mathcal{E}} = [\hat{\epsilon}_1 \dots \hat{\epsilon}_t \dots \hat{\epsilon}_T]^{'}\right)$ with matrix $\mathbf{C}_0$ and the fitted residuals of the reduced form VAR-X:

$$\hat{\mathcal{E}} = \hat{\mathbf{E}}\left(\mathbf{C}_0^{'}\right)^{-1}$$

3. Compute the historical decomposition of the endogenous variables relative to $\tilde{\mathbf{K}}_t$:

$$\tilde{y}_t^{(i,j)} = \sum_{i=0}^{t-1} c_i^{ij} \hat{\epsilon}_{t-i}^{j} \qquad \tilde{y}_t^{(i,k)} = \sum_{i=0}^{t-1} \lambda_i^{ik} \tilde{x}_{t-i}^{k}$$

4. Recover the values of $\tilde{\mathbf{K}}_t$ with the observed values of $\mathbf{y}_t$ and the auxiliary variable $\tilde{\mathbf{y}}_t$:

$$\tilde{\mathbf{K}}_t = \mathbf{y}_t - \tilde{\mathbf{y}}_t$$

5. Compute the historical decomposition of the endogenous variables:

$$y_t^{(i,j)} = \tilde{K}_t^i + \tilde{y}_t^{(i,j)} \qquad y_t^{(i,k)} = \tilde{K}_t^i + \tilde{y}_t^{(i,k)}$$

Steps 3 and 5 are repeated for $t = 1, 2, \dots, T$, $i = 1, \dots, n$, $j = 1, \dots, n$ and $k = 1, \dots, m$. Step 4 is repeated for $t = 1, 2, \dots, T$.

in the model in their first differences. The person performing the study may be asking himself the effects of the shocks and the changes in the exogenous variables over the endogenous variables. In this context, the criteria or reference value for the exogenous variables arises naturally as a base scenario of no change in the exogenous variables and no shocks. Under the described situation one has, for all t, $\overline{\mathbf{x}}_t = 0$ and $\widetilde{\mathbf{K}}_t = \mathbf{K}_t$. This also allows to interpret both $y_t^{(i,k)}$ and $\widetilde{y}_t^{(i,k)}$ as what would have happened to the i-th endogenous variable if it were only for the changes of the k-th exogenous variable.

Algorithm 6 summarizes the steps carried out for the historical decomposition in a structural VAR-X setup.

6. An Example

In this Section some of the concepts presented in the document are exemplified by an application of Galí's (1999) structural VAR, augmented with oil prices as an exogenous variable. The exercise has illustrative purposes only and does not mean to make any assessment on the economics involved.

The Section is organized as follows: first a description of the model to be used is made, then the lag structure of the reduced form VAR-X is chosen and the estimation described. Finally, impulse response functions, multiplier analysis and the historical decomposition are presented for one of the model's endogenous variables.

6.1. The Model and the Data

The model used in this application is original from Galí (1999) and is a bivariate system of labor productivity and a labor measure.[11] The labor productivity is defined as the ratio between gross domestic product (GDP) and labor. The identification of the shocks is obtained by imposing long run restrictions a la Blanchard & Quah (1989). Two shocks are identified, a technology (productivity) shock and a non-technology shock, the former is assumed to be the only shock that can have long run effects on the labor productivity. As pointed out in Galí (1999) this assumption is standard in neoclassical growth, RBC and New-Keynesian models among others.

The model is augmented with oil prices as an exogenous variable with the only purpose of turning it into a structural VAR-X model, so that it can be used to illustrate some of the concepts of the document. As mentioned in Section 3 the presence of an exogenous variable does not change the identification of the structural shocks.

All variables are included in the model in their first differences, this is done partially as a condition for the long run identification (labor productivity) and partially because of the unit root behavior of the observed series. It should be clear that, in the notation of the document, $n = 2$ (the number of endogenous variables) and $m = 1$ (the number of exogenous variables).

Noting by z_t the labor productivity, l_t the labor measure and p_t^o the oil price, the reduced form representation of the model is given by equation (1) with $\mathbf{y}_t = \left[\begin{array}{cc} \Delta z_t & \Delta l_t \end{array}\right]'$ and $x_t = \Delta p_t^o$:

$$\mathbf{y}_t = \mathbf{v} + \mathbf{B}_1 \mathbf{y}_{t-1} + \ldots + \mathbf{B}_p \mathbf{y}_{t-p} + \boldsymbol{\Theta}_0 x_t + \ldots + \boldsymbol{\Theta}_q x_{t-q} + \mathbf{e}_t$$

In the last equation vector $\mathbf{v}$ is of size 2×1, matrices $\mathbf{B}_i$ are of size 2×2 for $i = 1 : p$ and all $\boldsymbol{\Theta}_j$ are 2×1 vectors. The structural VMA-X form of the model is given (as in equation (4)) by:

[11] Galí uses total hours worked in the non-farm sector as labor measure in the main exercise but also points at the number of employees as another possible labor measure, here we take the second option and use non-farm employees.

$$\mathbf{y}_t = \mu + \mathbf{C}(L)\,\epsilon_t + \mathbf{\Lambda}(L)\,x_t$$

with μ a 2×1 vector, each matrix of $\mathbf{C}(L)$ is of size 2×2, and the "coefficients" of $\mathbf{\Lambda}(L)$ are 2×1 vectors. $\epsilon_t = \left[\begin{array}{cc} \epsilon_t^T & \epsilon_t^{NT} \end{array}\right]$ is the vector of structural shocks.

The identification assumption implies that $\mathbf{C}(1)$ is a lower triangular matrix, this allows us to use algorithm 2 for the identification of the shocks and the matrices in $\mathbf{C}(L)$. Equations (5), (6) and (7) still hold.

The data set used to estimate the model consists in quarterly GDP, non-farm employees and oil price series for the US economy that range from 1948Q4 to 1999Q1. The quarterly GDP is obtained from the Bureau of Economic Analysis, and the non-farm employees and oil price from the FRED database of the Federal Reserve Bank of St. Louis. GDP and non-farm employees are seasonally adjusted. GDP is measured in billions of chained 2005 dollars, non-farm employees in thousands of people and oil prices as the quarterly average of the WTI price in dollars per barrel.

6.2. Lag Structure and Estimation

Choosing the lag structure of the model consists in finding values for p and q so that the estimated reduced form model satisfies some conditions. In this case we shall choose values for p and q so that the residuals (e_t) are not auto-correlated.[12] The tests indicate that four lags of the endogenous variables are necessary for obtaining non-auto-correlated residuals $(p = 4)$, this result is independent of the lags of the exogenous variable. The change of the oil prices can be included only contemporary $(q = 0)$ or with up to six lags $(q = 6)$.

Since any number of lags of the exogenous variables makes the residuals satisfy the desired condition, the marginal density of the different models (under the Jeffreys prior) is used to determined the value of q. Each possible model only differs in the lags of exogenous variable, there are seven models indexed as $\mathcal{M}_i(\mathbf{Y})$ with $i = 0 \ldots 6$. The marginal density for each model is computed as in equation (23):

$$\mathcal{M}_i(\mathbf{Y}) = \frac{\Gamma_n\left(\frac{T-k_i}{2}\right)}{\Gamma_n\left(\frac{T-k_i-n-1}{2}\right)} \left|\mathbf{S}_i\right|^{\frac{-n-1}{2}} 2^{\frac{n(n+1)}{2}} \overline{C}$$

A presample is taken so that all models have the same effective T, since all have the same number of endogenous variables $(n = 2)$, the only difference between the marginal density of two models is in k_i (the total number of regressors) and $\mathbf{S}_i$ (the estimated covariance of the residuals). Recalling from Section 4: $k_i = (1 + np + m(q_i + 1))$ and $\mathbf{S}_i = \left(\mathbf{Y} - \mathbf{Z}_i\widehat{\mathbf{\Gamma}}_i\right)'\left(\mathbf{Y} - \mathbf{Z}_i\widehat{\mathbf{\Gamma}}_i\right)$.

Table 1 presents the results of the marginal densities, it is clear that the marginal density does not increase monotonically in the exogenous lag and that

[12] The auto-correlation of the residual is tested whit Portmanteau tests at a 5% significance level. See Lütkepohl (2005), Section 4.4.3.

$\mathcal{M}_4(\mathbf{Y})$ $(q = 4)$ is preferred to the other models. Then, the VAR-X model is estimated with four lags in both the endogenous and the exogenous variables, and the contemporary value of the change in the oil price.

Table 1: Marginal Densities.

$\mathcal{M}_0(\mathbf{Y})$	$\mathcal{M}_1(\mathbf{Y})$	$\mathcal{M}_2(\mathbf{Y})$	$\mathcal{M}_3(\mathbf{Y})$	$\mathcal{M}_4(\mathbf{Y})$	$\mathcal{M}_5(\mathbf{Y})$	$\mathcal{M}_6(\mathbf{Y})$
6.1379	6.1268	6.1664	6.1817	6.2414	6.1733	6.1115

The values presented are proportional to the marginal densities of the models by a factor of $10^{13}\overline{C}$.

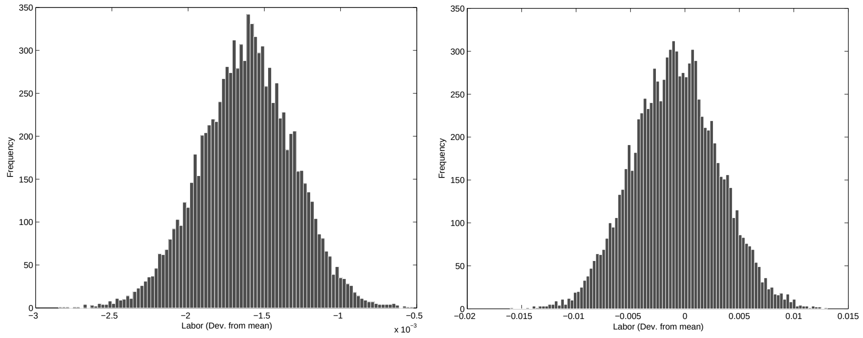
The estimation is carried out by Bayesian methods under the Jeffreys prior as in Section 4.2. Algorithm 3 is applied to obtain 10,000 draws of the reduced form parameters, for every draw Algorithm 2 is applied, along with the identification restriction over the technology shock, to obtain the parameters of the structural VMA-X representation of the model.

6.3. Impulse Response Functions and Multiplier Analysis

From the output of the Bayesian estimation of the model the impulse response function and multipliers are computed. Note that the distributions of the IRF and the multipliers are available since the estimation allows to obtain both for each draw of the reduced form parameters. This makes possible to compute highest posterior density regions (HPD) as mentioned in Section 5.1. For doing so we presented, in Algorithm 4, the steps to be carried out in the case in which the distribution of the IRF and the multipliers in every period is unimodal. Here we present only the response of labor to a technology shock and a change in oil price as the posterior mean of the responses generated for each of the 10,000 draws of the parameters, the responses are presented along with HPD regions at 68% and 90% probability.

Before presenting the HPD for the IRF and the multipliers, it is necessary to check if the distribution of the responses in every period are unimodal. Although no sufficient, a preliminary test of the mentioned condition is to check the histograms of the IRF and the multipliers before computing the HPD. Figure 1 presents the histograms for the response of labor to a technology shock (Figure 1(a)) and to a change in oil price (Figure 1(b)) at impact, the histograms for up to 20 periods ahead are also checked, but not presented. In all cases Algorithm 4 can be used.

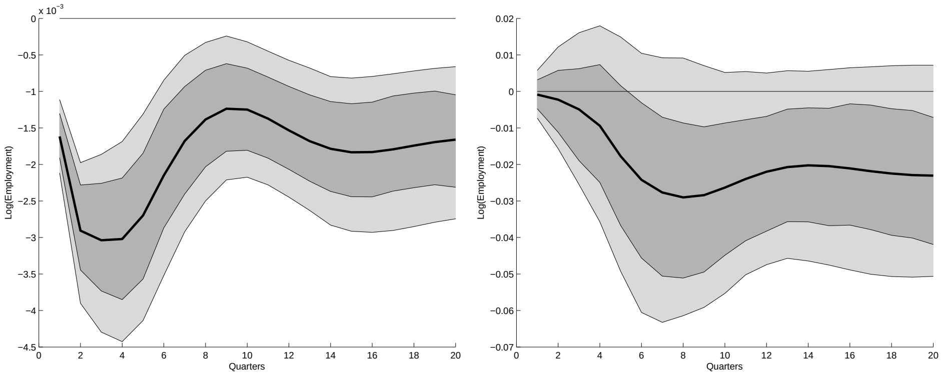
The results are presented in Figure 2 and point to a decrease of labor in response to both a positive technology shock and an increase in oil prices, although the decrease is only significant for the response to a technology shock. The response of labor to an increase in the oil price is never significant at 90% probability and only significant at 68% probability after period 5.



(a) IRF: Labor to tech shock at impact (b) MA: Labor to oil price at impact

Histograms of the response of labor to a technology shock and a change in the oil price at impact. The histograms are obtained from 10000 draws of the parameters of the structural VAR-X model, and are computed with 100 bins.

Figure 1: Histograms.



(a) IRF: Labor to tech shock (b) MA: Labor to oil price at impact

Response of labor to a unitary technology shock and a unit change in the oil price. The point estimate (dark line) corresponds to the posterior mean of the distribution of the IRF and the multipliers of labor, the distributions are obtained from 10000 draws of the parameters of the structural VAR-X model. HPD regions at 68% and 90% probability are presented as dark and light areas correspondingly.

Figure 2: Impuse Response Functions and Multiples Analysis.

6.4. Historical Decomposition

Finally, the historical decomposition of labor into the two structural shocks and the changes in the oil price is computed. As mentioned in Section 5.2 it is necessary to fix a reference value for the exogenous variable. Since the oil price enters the model in its first difference, the reference value will be set to zero ($\forall_t\, \overline{x}_t = 0$). This means that all changes in the oil price are understood by the model as innovations to that variable.[13] In this exercise all computations are carried out with the posterior mean of the parameters. Since the Jeffreys prior was used in the estimation, the posterior mean of the parameters equals their maximum likelihood values.

Applying Algorithm 6, steps 1 to 3, the historical decomposition for the first difference of labor (relative to $\widetilde{\mathbf{K}}_t$) is obtained, this is presented in Figure 3. Yet, the results are unsatisfactory, principally because the quarterly difference of labor lacks a clear interpretation, its scale is not the one commonly used and might be too volatile for allowing an easy understanding of the effects of the shocks.[14]

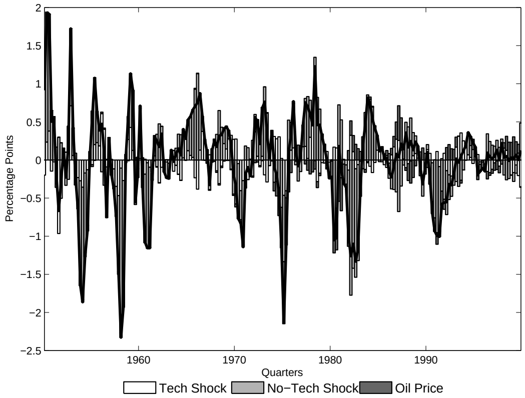


FIGURE 3: Historical Decomposition - Labor in first difference

An alternative to the direct historical decomposition is to use the conditioned series (step 5 of Algorithm 6) to compute the historical decomposition of the annual differences of the series, this is done by summing up the quarterly differences conditioned to each shock and the exogenous variable. The advantage of this transformation is that it allows for an easier interpretation of the historical decomposition, since the series is now less volatile and its level is more familiar for the researcher (this is the case of the annual inflation rate or the annual GDP growth rate). The result is presented in Figure 4, it is clear that labor dynamics have been governed mostly by non-technology shocks in the period under consideration, with technology shocks and changes in the oil price having a minor effect.

It is worth to note that decomposing the first difference of the series (as in Figures 3 and 4) has another advantage. The decomposition is made relative to $\widetilde{\mathbf{K}}_t$ with $\overline{x}_t = 0$, hence $\widetilde{\mathbf{K}}_t = \mathbf{K}_t$ and $\widetilde{\mathbf{K}}_t \longrightarrow \mu$, this means, for Figure 3, that the decomposition is made relative to the sample average of the quarterly growth rate of the series, in that case if the black solid line is, for example, 0.1 at some point it can be read directly as the growth rate of labor being 10% above its sample

[13] Another possibility is to use the sample mean of the change in the oil price as a reference value, in this case the innovations are changes of the oil price different to that mean.

[14] In fact the series used is not too volatile, but there are other economically relevant series whose first difference is just too volatile for allowing any assessment on the results, the monthly inflation rate is usually an example of this.

average. Since Figure 4 is also presenting differences it can be shown that the new $\widetilde{\mathbf{K}}_t$ converges to the sample mean of the annual growth rate of the series, making interpretation of the decomposition easier to read.

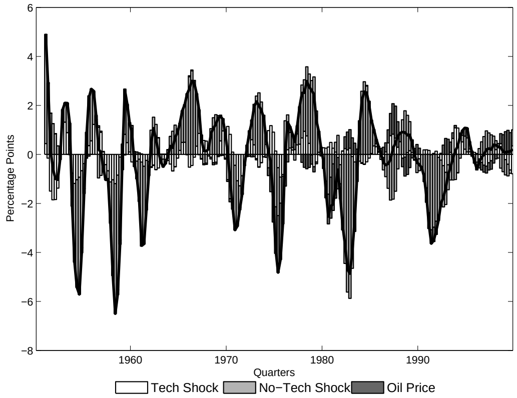


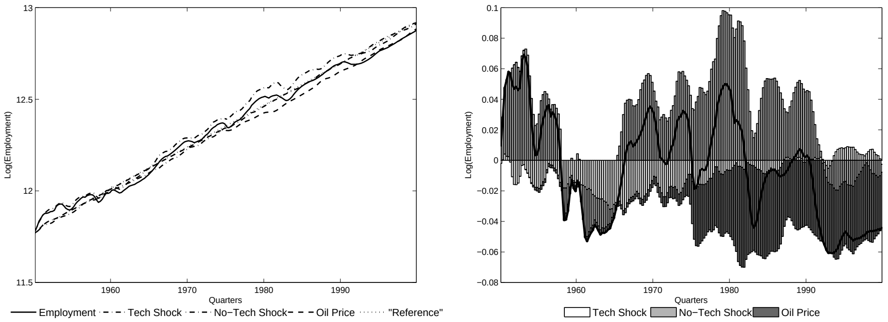
Figure 4: Historical Decomposition - Labor in annual differences

Another alternative is to accumulate the growth rates (conditioned to each shock and the exogenous variable) starting from the observed value of the series in the first period, this generates the historical decomposition of the level of the variable. The results of this exercise are presented in Figure 5.

There are several points to be made about the historical decomposition in levels, the first one is that, since $\widetilde{\mathbf{K}}_t$ is also being accumulated from some initial value, the decomposition is not made relative to a constant but relative to a line, this line corresponds to the linear tendency of the series. Figure 5(a) plots the actual path of labor along with paths conditioned to each shock and the exogenous variable and the "Reference" line, which is the accumulation of $\widetilde{\mathbf{K}}_t$. Interpretation of Figure 5(a) is difficult because the effect of each shock and the exogenous variable is obtained as the difference between its conditioned path and the "Reference" line, because all are moving in each period identifying that effect becomes a challenging task.

The second point arises from the interpretation of Figure 5(b), which presents the decomposition of the level of labor relative to the "Reference" line, this is similar to what was presented in Figures 3 and 4. The interpretation is nevertheless more complicated. In the former Figures the decomposition was made relative to a constant, but the decomposition in levels is made relative to a line, whose value is changing in each period, this makes the reading of the level of the bars and the line more difficult. If the line is in 3 it means that the observed series is 3 units above its linear tendency.

Another characteristic of decomposition in level must be mentioned, although it is not clear from Figure 5(b), the accumulated effects of the shocks over any series in the first and last period are, by construction, equal to zero. This means that the bars associated with the structural shocks are not present in both the first and last period of the sample, and that the value of the observed variable has to be explained entirely by the exogenous variables, moreover, it means that the accumulated effect of the shocks has to be dis-accumulated when the sample is getting to its end. This occurs because the accumulated effect of the shocks has to be zero at the beginning of the sample, since the effect of the shocks before that point is summarized in the initial value of the series, and because the mean of the shocks over the sample is zero (one of the properties of the estimation), this implies that $\sum_{t=1}^{T}\epsilon_t^i = 0$. When the conditioned difference series is accumulated, the effect of the shock is accumulated so that it also sums to zero. This last problem is not present in the historical decomposition in differences (or annual differences) and makes the results of the decomposition in levels to be unreliable.



(a) Decomposition in level (b) Decomposition around reference value

FIGURE 5: Historical Decomposition - Labor in level

7. Final Remarks

This paper presents a review of VAR-X topics with emphasis in Bayesian estimation, and different applications of the model, covering impulse response functions, multiplier analysis, forecast error variance decomposition and historical decomposition calculations. The treatment and discussion of the latter constitutes a novelty in the literature, since it has been largely ignored (with few exceptions) despite its usefulness in the context of multivariate time series analysis. A short exercise in presented using much of the technique reviewed.

Bayesian methods are presented with detail and shown as an easy to implement option for overcoming the usual small sample restrictions faced by frequentist methods. These methods are off course recommended to scholars when using the VAR or the VAR-X model.

Finally, this document is intended as an introductory review to the VAR-X model, and does not exhausts all the literature available about it. A couple of examples of further topics in the literature are the VAR model with mixed frequencies, like Rodriguez & Puggioni (2010) or Chiu, Eraker, Foerster, Kim &

Seoane (2011) (and the references therein), and the recently proposed Copula-VAR-X as in Bianchi, Carta, Fantazzini, Giuli & Maggi (2010), who use flexible multivariate distributions, different from the normal distribution, allowing a rich dependence structure and more flexible marginal distributions for better fit of empirical data, specially leptokurtosis.

Acknowledgements

We are in debt with the editor of the Colombian Journal of Statistics and three anonymous referees, their suggestions where crucial for this paper. We wish to thank also Eliana González, Martha Misas, Andrés González, Luis Fernando Melo, and Christian Bustamante for useful comments on earlier drafts of this document, of course, all remaining errors are our own. Needless to say, the results and opinions expressed in this document do not compromise in any way the Banco de la República, its board of governors or the Inter-American Development Bank.

[Recibido: enero de 2012 — Aceptado: noviembre de 2012]

References

Amisano, G. & Giannini, C. (1997), *Topics in Structural VAR Econometrics*, Springer.

Bauwens, L., Lubrano, M. & Richard, J.-F. (2000), *Bayesian Inference in Dynamic Econometric Models*, Oxford University Press.

Bianchi, C., Carta, A., Fantazzini, D., Giuli, M. E. D. & Maggi, M. (2010), 'A Copula VAR-X Approach for Industrial Production Modelling and Forecasting', *Applied Economics* **42**(25), 3267–3277.

Blanchard, O. J. & Quah, D. (1989), 'The Dynamic Effects of Aggregate Demand and Supply Disturbances', *American Economic Review* **79**(4), 655–73.

Burbidge, J. & Harrison, A. (1985), 'A Historical Decomposition of the Great Depression to Determine the Role of Money', *Journal of Monetary Economics* **16**(1), 45–54.

Canova, F. (2007), *Methods for Applied Macroeconomic Research*, Princeton University Press, Nueva Jersey.

Canova, F. & De Nicolo, G. (2002), 'Monetary Disturbances Matter for Business Fluctuations in the G-7', *Journal of Monetary Economics* **49**(6), 1131–1159.

Canova, F. & Pappa, E. (2007), 'Price Differentials in Monetary Unions: The Role of Fiscal Shocks', *Economic Journal* **117**(520), 713–737.

Casella, G. & George, E. I. (1992), 'Explaining the Gibbs Sampler', *The American Statistician* **46**(3), 167–174.

Chen, M. & Shao, Q. (1998), ‘Monte Carlo Estimation of Bayesian Credible and HPD Intervals’, *Journal of Computational and Graphical Statistics* **8**, 69–92.

Chiu, C. W. J., Eraker, B., Foerster, A. T., Kim, T. B. & Seoane, H. D. (2011), Estimating VAR’s Aampled at Mixed or Irregular Spaced Frequencies : a Bayesian Approach, Research Working Paper RWP 11-11, Federal Reserve Bank of Kansas City.

Christiano, L. J., Eichenbaum, M. & Evans, C. L. (1999), Monetary Policy Shocks: What Have We Learned and to What End?, *in* J. B. Taylor & M. Woodford, eds, ‘Handbook of Macroeconomics’, Vol. 1 of *Handbook of Macroeconomics*, Elsevier, chapter 2, pp. 65–148.

Galí, J. (1999), ‘Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?’, *American Economic Review* **89**(1), 249–271.

Geman, S. & Geman, D. (1984), ‘Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Hyndman, R. J. (1996), ‘Computing and Graphing Highest Density Regions’, *The American Statistician* **50**, 120–126.

Jeffreys, H. (1961), *Theory of Probability*, International Series of Monographs on Physics, Clarendon Press.

Kadiyala, K. R. & Karlsson, S. (1997), ‘Numerical Methods for Estimation and Inference in Bayesian VAR-Models’, *Journal of Applied Econometrics* **12**(2), 99–132.

Kilian, L. (1998), ‘Small-Sample Confidence Intervals For Impulse Response Functions’, *The Review of Economics and Statistics* **80**(2), 218–230.

King, T. B. & Morley, J. (2007), ‘In Search of the Natural Rate of Unemployment’, *Journal of Monetary Economics* **54**(2), 550–564.

Kociecki, A. (2010), ‘A Prior for Impulse Responses in Bayesian Structural VAR Models’, *Journal of Business & Economic Statistics* **28**(1), 115–127.

Koop, G. (1992), ‘Aggregate Shocks and Macroeconomic Fluctuations: A Bayesian Approach’, *Journal of Applied Econometrics* **7**(4), 395–411.

Litterman, R. B. (1986), ‘Forecasting with Bayesian Vector Autoregressions - Five Years of Experience’, *Journal of Business & Economic Statistics* **4**(1), 25–38.

Lütkepohl, H. (1990), ‘Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models’, *The Review of Economics and Statistics* **72**(1), 116–25.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer.

Moon, H. R., Schorfheide, F., Granziera, E. & Lee, M. (2011), Inference for VARs Identified with Sign Restrictions, NBER Working Papers 17140, National Bureau of Economic Research, Inc.

Mountford, A. & Uhlig, H. (2009), 'What are he Effects of Fiscal Policy Shocks?', *Journal of Applied Econometrics* **24**(6), 960–992.

Nicholls, D. F. & Pope, A. L. (1988), 'Bias in the Estimation of Multivariate Autoregressions', *Australian Journal of Statistics* **30A**(1), 296–309.

Rodriguez, A. & Puggioni, G. (2010), 'Mixed Frequency Models: Bayesian Approaches to Estimation and Prediction', *International Journal of Forecasting* **26**(2), 293–311.

Runkle, D. E. (1987), 'Vector Autoregressions and Reality', *Journal of Business & Economic Statistics* **5**(4), 437–42.

Sims, C. A. (1980), 'Macroeconomics and Reality', *Econometrica* **48**(1), 1–48.

Sims, C. A. & Zha, T. (1999), 'Error Bands for Impulse Responses', *Econometrica* **67**(5), 1113–1156.

Smets, F. & Wouters, R. (2007), 'Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach', *American Economic Review* **97**(3), 586–606.

Uhlig, H. (2005), 'What are the Effects of Monetary Policy on Output? Results From an Agnostic Identification Procedure', *Journal of Monetary Economics* **52**(2), 381–419.

Zellner, A. (1996), *An Introduction to Bayesian Inference in Econometrics*, Wiley Classics Library, John Wiley.

Revista Colombiana de Estadística
Diciembre 2012, volumen 35, no. 3, pp. 509 a 521

Some Alternative Predictive Estimators of Population Variance

Algunos estimadores predictivos alternativos de la varianza poblacional

Radhakanta Nayak[1,a], Lokanath Sahoo[2,b]

[1]Department of Statistics, Khallikote College, Berhampur, India
[2]Department of Statistics, Utkal University, Bhubaneswar, India

Abstract

Using a predictive estimation procedure, an attempt has been made to develop some estimators for the finite population variance in the presence of an auxiliary variable. Analytical and simulation studies have been undertaken for understanding the performance of the suggested estimators compared to some existing ones.

***Key words*:** Auxiliary variable, Bias, Efficiency, Prediction approach.

Resumen

Mediante el uso de un procedimiento de estimación predictivo, se desarrollan algunos estimadores de la varianza poblacional en la presencia de una variable auxiliar. Estudios analíticos y de simulación son implementados para entender el desempeño de los estimadores sugeridos en comparación con otros ya existentes.

***Palabras clave*:** variable auxiliar, sesgo, eficiencia, enfoque de predicción.

1. Introduction

Let $U = \{1, 2, \dots, i \dots, N\}$ be a finite population, and y and x denote the study variable and the auxiliary variable taking values y_i and x_i respectively on the ith unit ($i = 1, 2, \dots, N$). Let $\overline{Y} = \sum_{i=1}^{N} y_i/N$ and $\overline{X} = \sum_{i=1}^{N} x_i/N$ be the population means, $S_y^2 = \sum_{i=1}^{N}(y_i - \overline{Y})^2/(N-1)$ and $S_x^2 = \sum_{i=1}^{N}(x_i - \overline{X})^2/(N-1)$ be the population variances of y and x respectively. Assume that a sample s of n units is drawn from U according to simple random sampling without replacement

[a]Lecturer. E-mail: rkn2010@gmail.com
[b]Professor. E-mail: lnsahoostatuu@rediffmail.com

(SRSWOR) in order to estimate the unknown parameter S_y^2. Let $\overline{y} = \sum_{i\in s} y_i/n$ and $\overline{x} = \sum_{i\in s} x_i/n$ be the sample means, $s_y^2 = \sum_{i\in s}(y_i - \overline{y})^2/(n-1)$ and $s_x^2 = \sum_{i\in s}(x_i - \overline{x})^2/(n-1)$ the sample variances.

In certain situations, estimation of population variance S_y^2 has received considerable attention from survey statisticians. For example, in manufacturing industries and pharmaceutical laboratories, sometimes the researchers are interested in the variation of their products. Although, the literature describes a great variety of techniques for using auxiliary information by means of ratio, and product and regression methods for estimating population mean, variance estimation using auxiliary information has received scarce attention. This is perhaps due to the belief that the gain in efficiency we could obtain by involving an auxiliary variable may not be too much relevant to motivate the use of more complex estimators. However, some efforts in this direction are due to Das & Tripathi (1978), Isaki (1983), Prasad & Singh (1990)(1992), Singh & Kataria (1990), Srivastava & Jhajj (1980)(1995), Singh & Singh (2001), Ahmed, Walid & Ahmed (2003), Giiancarlo & Chiara (2004), Jhajj, Sharma & Grover (2005), Kadilar & Cingi (2006)(2007) and Grover (2007). Two notable estimators that are very much popular in the literature are due to Isaki (1983) defined by

$$\nu_1 = s_y^2 S_x^2 / s_x^2$$

and

$$\nu_2 = s_y^2 + b^*(S_x^2 - s_x^2)$$

where b^* is an estimate of the regression coefficient of s_y^2 on s_x^2 defined by $b^* = \frac{s_y^2(\widehat{\lambda}-1)}{s_x^2(\widehat{\beta}_2(x)-1)}$, such that $\widehat{\lambda} = m_{22}/m_{20}m_{02}$ and $\widehat{\beta}_2(x) = m_{40}/m_{20}^2$ with $m_{rs} = \sum_{i\in s}(x_i - \overline{x})^r(y_i - \overline{y})^s/n$ [cf., Garcia & Cebrain (1996), and Kadilar & Cingi (2006)].

During the years that followed, much emphasis has been given on the prediction of population mean or total [cf., Srivastava (1983)]. But, little interest has been shown towards the prediction of the population variance. Under this approach, the survey data at hand i.e., the sample observations are treated as fixed and unassailable. Uncertainty is then attached only to the unobserved values which need to be predicted. Bolfarine & Zacks (1992) indicated various techniques for predicting population variance. Biradar & Singh (1998), using classical estimation theory, provided some predictive estimators for S_y^2. In this paper, using auxiliary variable x, we develop some more estimators under the prediction approach of Basu (1971) with regards to a finite population setup.

2. Prediction Criterion

Let us decompose U into two mutually exclusive domains s and r of n and $N-n$ units respectively, where $r = U - s$ denotes the collection of units in U which are not included in s. Then, under the usual prediction criterion given in Bolfarine & Zacks (1992), it is possible to express

$$(N-1)S_y^2 = (n-1)s_y^2 + (N-n-1)S_{y(r)}^2 + (1-f)n(\overline{y} - \overline{Y}_r)^2, \qquad (1)$$

where $f = n/N$, and $\overline{Y}_r = \sum_{i\in s} y_i/(N-n)$ and $S^2_{y(r)} = \sum_{i\in r}(y_i - \overline{Y}_r)^2/(N-n-1)$ are respectively the mean and variance of y-values belonging to r.

Notice that the first component on the right hand side of (1) is known while the second and third components are unknown. Hence, the prediction of $(N-1)S^2_y$ is possible when $S^2_{y(r)}$ and $\overline{Y}_r$ are simultaneously predicted by some means from the sample data. Using V_r and M_r as their respective predictors, a predictor of S^2_y can be provided by the equation:

$$(N-1)\widehat{S}^2_y = (n-1)s^2_y + (N-n-1)V_r + (1-f)n(\overline{y} - M_r)^2 \tag{2}$$

Most of the predictions are based either on distributional forms or an assumed model [cf., Royall (1988), Bolfarine & Zacks (1992)]. However, Sampford (1978) argued that the consideration of a model free prediction can generate a new, estimator possessing some desirable properties. Basu (1971) also encouraged the use of tools of the classical estimation theory to find out suitable predictors for $\overline{Y}$. Biradar & Singh (1998) formulated some estimators of S^2_y from (2) by considering suitable choices of the predictors V_r and M_r in terms of the auxiliary variable x under the tools of classical estimation theory. Defining $\overline{X}_r = \sum_{i\in r} x_i/(N-n)$ and $S^2_{x(r)} = \sum_{i\in r}(x_i - \overline{X}_r)^2/(N-n-1)$, we report below their estimators along with the corresponding selections of V_r and M_r:

$$\nu_3 = \left(\frac{N-2}{N-1}\right) s^2_y$$

when $V_r = s^2_y$ and $M_r = \overline{y}$,

$$\nu_4 = \frac{s^2_y}{s^2_x} S^2_x + \frac{nN(\overline{x} - \overline{X})^2}{(N-n)(N-1)} \left(\frac{\overline{y}^2}{\overline{x}^2} - \frac{s^2_y}{s^2_x}\right)$$

when $V_r = s^2_y S^2_{x(r)}/s^2_x$ and $M_r = \overline{y}\overline{X}_r/\overline{x}$, and

$$\nu_5 = \frac{s^2_y}{s^2_x} S^2_x + \frac{nN(\overline{x} - \overline{X})^2}{(N-n)(N-1)} \left(b^2_{yx} - \frac{s^2_y}{s^2_x}\right)$$

when $V_r = s^2_y S^2_{x(r)}/s^2_x$ and $M_r = \overline{y} + b_{yx}(\overline{X}_r - \overline{x})$, where $b_{yx} = s_{yx}/s^2_x$.

Biradar & Singh (1998) also identified Isaki's (1983) estimator ν_1 as a special case of (2) for $V_r = s^2_y S^2_{x(r)}/s^2_x$ and $M_r = \overline{y} + s_y(\overline{X}_r - \overline{x})/s_x$. This shows that the estimator possesses a predictive character.

3. Some New Predictive Estimators of S^2_y

In the following discussions, we introduce some alternative approaches in order to develop a few more predictive estimators of S^2_y.

1. Consider the following alternative but equivalent representation of S_y^2:

$$(N-1)S_y^2 = (n-1)s_y^2 + (N-n)[\sigma^2_{y(r)} + f(\overline{y} - \overline{Y}_r)^2] \tag{3}$$

where $\sigma^2_{y(r)} = \sum_{i\in r}(y_i - \overline{Y}_r)^2/(N-n)$. Denoting V_r^* as a predictor of $\sigma^2_{y(r)}$ and M_r, as the predictor of $\overline{Y}_r$, the following alternative predictive equation can be considered:

$$(N-1)S_y^2 = (n-1)s_y^2 + (N-n)[V_r^* + f(\overline{y} - M_r)^2] \tag{4}$$

Then, for $V_r^* = \left(\frac{n-1}{n}\right)s_y^2$ and $M_r = \overline{y}$ in (4) we get an estimator of S_y^2 defined by

$$\nu_6 = \left(\frac{n-1}{n}\right)\left(\frac{N}{N-1}\right)s_y^2$$

2. Biradar & Singh (1998) developed the estimator ν_5 from (2) with $V_r = s_y^2 S^2_{x(r)}/s_x^2$ and $M_r = \overline{y} + b_{yx}(\overline{X}_r - \overline{x})$. See that in such an attempt V_r has been assumed a ratio version of the variance estimator while the connected mean estimator is a regression estimator. Hence as a matter of curiosity, we may also think in the light of Isaki (1983) to use a regression version of the variance estimator i.e., $V_r = s_y^2 + b^*(S^2_{x(r)} - s_x^2)$ along with the mean estimator $M_r = \overline{y} + b_{yx}(\overline{X}_r - \overline{x})$ in the predictive equation (2) to predict S_y^2. This operation, after a considerable simplification, leads to produce the following estimator:

$$\nu_7 = \frac{N-2}{N-1}\left[s_y^2 + b^*\left(\frac{N-1}{N-2}S_x^2 - s_x^2\right)\right]$$

3. Srivastava (1983) considered the predictive equation:

$$\hat{\hat{\overline{Y}}} = f\overline{y} + (1-f)M_r \tag{5}$$

where M_r is the implied predictor of $\overline{Y}_r$, for predicting $\overline{Y}$ and shown that when $M_r = \overline{y}\overline{X}_r/\overline{x}, \hat{\hat{\overline{Y}}} = \overline{y}_R = \overline{y}\overline{X}/\overline{x}$, the classical ratio estimator of $\overline{Y}$, and when $M_r = \overline{y} + b_{yx}(\overline{X}_r - \overline{x}), \hat{\hat{\overline{Y}}} = \overline{y}_L = \overline{y} + b_{yx}(\overline{X} - \overline{x})$, the classical regression estimator of $\overline{Y}$. Thus, both the ratio and regression estimators ($\overline{y}_R$ and $\overline{y}_L$) of the mean possess a predictive character, the origin of which actually lies in predicting y_i's, $i \in r$, by $y_i = \overline{y}x_i/\overline{x}$ and $y_i = \overline{y} + b_{yx}(x_i - \overline{x})$ in that order. In view of this, we designate these two estimators as basic estimators of the population mean. Notice that the predictive estimators ν_1, ν_4, ν_5 and ν_7 suggested so far have been obtained by using either $V_r = s_y^2 S^2_{x(r)}/s_x^2$ or $V_r = s_y^2 + b^*(S^2_{x(r)} - s_x^2)$ as the case may be. This means that the unknown quantity $S^2_{y(r)}$ is estimated as a whole with the same principle as that applied to estimate $\overline{Y}_r$. But, such a choice of V_r seems to be arbitrary by nature. Rather, we feel that it is more appropriate if the variance is established by predicting individual y_i's, $i \in r$, for which we need to express S_y^2 in the following form:

$$(N-1)S_y^2 = (n-1)s_y^2 + \sum_{i\in r} y_i^2 - (N-n)\overline{Y}_r^2 + (1-f)n(\overline{y} - \overline{Y}_r)^2 \tag{6}$$

A number of new estimators can be easily generated from this equation on the basis how $\sum_{i\in r} y_i^2$ is predicted. But, for simplicity, here we consider the prediction of $y_i, i \in r$, either by $y_i = \overline{y}x_i/\overline{x} = \overline{y} + \overline{y}(x_i - \overline{x})/\overline{x}$ or by $y_i = \overline{y} + b_{yx}(x_i - \overline{x})$ and prediction of $\overline{Y}_r$ by $\overline{y}\overline{X}_r/\overline{x}$.

Then, accordingly after a considerable simplification, we obtain the following two new estimators:

$$\nu_8 = \left(\frac{n-1}{N-1}\right)\left[s_y^2 + \left(\frac{\overline{y}}{\overline{x}}\right)^2\left(\frac{N-1}{n-1}S_x^2 - s_x^2\right)\right]$$

$$\nu_9 = \left(\frac{n-1}{N-1}\right)\left[s_y^2 + b_{yx}^2\left(\frac{N-1}{n-1}S_x^2 - s_x^2\right)\right]$$

4. Performance of the Proposed Estimators

Out of the nine estimators considered or proposed in the preceding sections, the estimators ν_3 and ν_6 were achieved without using any auxiliary information whereas others were achieved through the use of information on the auxiliary variable x. A desirable goal here is to study the performance of the proposed estimators ν_6 to ν_9 compared to ν_1 to ν_5 at least in respect of bias and mean square error (MSE) i.e., efficiency, where bias and MSE of an estimator ν_i of S_y^2 are defined respectively by $B(\nu_i) = E(\nu_i) - S_y^2$ and $M(\nu_i) = E(\nu_i - S_y^2)^2 (i = 1, 2, \ldots, 9)$. But, we see that some of the estimators are so complex that it is not possible to derive exact expressions for their bias and MSE. Biradar & Singh (1998) presented asymptotic expressions for these performance measures for the estimators ν_1 to ν_5. On the other hand, Nayak (2009) derived these expressions in favor of ν_1 to ν_9. But, the sufficient conditions for superiority of one estimator over other derived by the authors using asymptotic expressions are so complicated that it is not conducive to compare different estimators meaningfully. However, to facilitate our comparison, these expressions are considered under the following widely used linear regression model:

$$y_i = \boldsymbol{\beta} x_i + e_i,\ i = 1, 2, \ldots, N \tag{7}$$

where $\boldsymbol{\beta}(> 0)$ is the model parameter and e_i is the error component such that $E(e_i/x_i) = 0, E(e_i^2/x_i) = \delta x^g(\delta > 0, 0 \le g \le 1)$, and $E(e_ie_j/x_i, x_j) = 0$ for $i \neq j$. Further, we also assume that $E(e_i^4/x_i) = \xi x^g$ and $E(e_i^3/x_i) = E(e_i^3e_j/x_i, x_j) = E(e_ie_j^3/x_i, x_j) = 0, (i \neq j)$. It may be pointed out here that the asymptotic expressions for bias and MSE of different estimators under this assumed model are derived through the Taylor linearization method.

4.1. Comparison of Bias

After some algebraic manipulations (suppressed to save space), we get the following model-based results in respect of the bias of different estimators up to $O(n^{-1})$

$$B(\nu_1) = \mathcal{C}\delta E(x^g) \tag{8}$$

$$B(\nu_2) = 0 \tag{9}$$

$$B(\nu_3) = -\frac{1}{N-1}[\beta^2 S_x^2 + \delta E(x^g)] \tag{10}$$

$$B(\nu_4) = -(\mathcal{B} - \mathcal{C})\delta E(x^g) \tag{11}$$

$$B(\nu_5) = -(\mathcal{K} - \mathcal{C})\delta E(x^g) \tag{12}$$

$$B(\nu_6) = -\frac{N-n}{N-1}[\beta^2 S_x^2 + \delta E(x^g)] \tag{13}$$

$$B(\nu_7) = -\frac{1}{N-1}\left(\frac{n-2}{n-1}\right)\delta E(x^g) \tag{14}$$

$$B(\nu_8) = -(N-n)\mathcal{B}\delta E(x^g) \tag{15}$$

$$B(\nu_9) = -\left(\frac{N-n}{N-1}\right)\left(\frac{n-2}{n-1}\right)\delta E(x^g) \tag{16}$$

where $\mathcal{B} = \frac{1}{N-1}\left(1 - \frac{C_x^2}{n}\right)$, $\mathcal{C} = \frac{1}{n}(\beta_2(x) - 2)$ and $\mathcal{K} = \left(\frac{n}{n-1}\right)\left(\frac{1}{N-1}\right)$, such that C_x and $\beta_{2(x)}$ are respectively the coefficient of variation and β_2- coefficient of the auxiliary variable x.

In the light of the expressions (8) to (16), we state the following comments on the bias of the estimators:

(i) The regression estimator ν_2 is model-unbiased, ν_1 is positively biased and the rest seven estimators are negatively biased.

(ii) $|B(\nu_3)| < |B(\nu_6)|$. This indicates that the bias of ν_6 is always greater than that of ν_3.

(iii) $|B(\nu_8)| < |B(\nu_7)|$i.e., ν_8 is less biased than ν_7.

(iv) $|B(\nu_7)| < |B(\nu_9)|$i.e., ν_7 is less biased than ν_9.

(v) $|B(\nu_9)| \lesseqgtr |B(\nu_8)|$ according as $C_x^2 \lesseqgtr \frac{n}{n-1}$.

(vi) $|B(\nu_4)| < |B(\nu_7)|$, when $|\mathcal{B} - \mathcal{C}| < \frac{1}{N-1}\left(\frac{n-2}{n-1}\right)$.

(vii) $|B(\nu_5)| < |B(\nu_7)|$, when $|\mathcal{K} - \mathcal{C}| < \frac{1}{N-1}\left(\frac{n-2}{n-1}\right)$.

(viii) $|B(\nu_7)| < |B(\nu_1)|$, when $\mathcal{C} > \mathcal{K}$ and $n > 2$.

In view of (iii) and (iv), although we can conclude that ν_8 is less biased than ν_7 and ν_9, we fail to obtain a clear-cut idea on the magnitude of bias of ν_8 compared to ν_1,ν_4 and ν_5. Because, comparison of (15) with (8) or (11) or (12) does not lead to any meaningful conditions.

4.2. Comparison of Efficiency

We present below model-based asymptotic expressions of the MSEs of different estimators up to $O(n^{-1})$ together with the exact expression for the variance of the traditional unbiased estimator s_y^2.

$$V(s_y^2) = V(\nu_2) + \mathcal{C}\beta^4 S_x^4 \tag{17}$$

$$M(\nu_1) = M(\nu_2) + \mathcal{C}\delta^2 E^2(x^g) \tag{18}$$

$$M(\nu_2) = \xi(x^g) + 4\boldsymbol{\beta}^2 S_x^2 \frac{\delta E(x^g)}{n-1} - \frac{n-3}{n(n-1)}\delta^2 E^2(x^g) \tag{19}$$

$$M(\nu_3) = \left(\frac{N-2}{N-1}\right)^2 V(s_y^2) \cong V(s_y^2) \tag{20}$$

$$M(\nu_4) = M(\nu_2) + \mathcal{C}\delta^2 E^2(x^g) + \frac{2}{N-1}\delta^2 E^2(x^g) \tag{21}$$

$$M(\nu_5) = M(\nu_2) + \mathcal{C}\delta^2 E^2(x^g) + \frac{2}{N-1}\left(\frac{n}{n-1}\right)\delta^2 E^2(x^g) \tag{22}$$

$$M(\nu_6) = \left(\frac{n-1}{n}\right)^2 \left(\frac{N}{N-1}\right)^2 V(s_y^2) \cong \left\{1 - 2\left(\frac{1}{n} + \frac{1}{N-1}\right)\right\} V(s_y^2) \tag{23}$$

$$M(\nu_7) = M(\nu_2) \tag{24}$$

$$\begin{aligned} M(\nu_8) = M(\nu_2) - 4\left(\frac{N-n}{N-1}\right)^2 \boldsymbol{\beta}^2 S_x^2 \frac{\delta E(x^g)}{n-1} + \\ 2\left(\frac{N-n}{N-1}\right)^2 \frac{C_x^2}{n}(2\boldsymbol{\beta}^2 S_x^2 - 1)\delta E(x^g) \end{aligned} \tag{25}$$

$$M(\nu_9) = M(\nu_2) + 2\left(1 - 2\frac{N-n}{N-1}\right)\delta^2 \frac{E^2(x^g)}{n-1} + \left(\frac{N-n}{N-1}\right)^2 \delta^2 E^2(x^g). \tag{26}$$

From these expressions, as ν_2 appears to be more efficient than $s_y^2, \nu_1, \nu_3, \nu_4$ and ν_5, we present the following results concerning efficiencies of the suggested estimators:

(ix) $M(\nu_6) < M(\nu_3) < V(s_y^2)$. This indicates that ν_6 is more efficient than both s_y^2 and ν_3.

(x) $M(\nu_7) = M(\nu_2)$ i.e., ν_7 and ν_2 are equally efficient even though they are configurationally different.

(xi) ν_8 is more efficient than ν_2 when $\boldsymbol{\beta}^2 S_x^2 < \frac{1}{2}$ which is very often satisfied in practice. This means that there is a scope to improve upon the Isaki's regression estimator ν_2 through ν_8.

(xii) The estimator ν_9 is less efficient than ν_2 when $n < \dfrac{N+1}{2}$.

(xiii) $M(\nu_8) < M(\nu_2) = M(\nu_7) < M(\nu_9)$, when $n < \dfrac{N+1}{2}$ and $\boldsymbol{\beta}^2 S_x^2 < \frac{1}{2}$. This shows that ν_8 is preferred to ν_2, ν_7 and ν_9 when the stated conditions are satisfied. The first condition is not a serious one. The second condition is easily satisfied for characters being measured in smaller magnitudes. We can also reduce the mean square error by considering transformations on the auxiliary variable and making the second condition more feasible.

4.3. Some Remarks

From the previous model-based comparisons, we see that the proposed estimator ν_8 turns out to be more efficient than others. But no meaningful conclusion could be drawn in favor of the four proposed estimators $\nu_i, i = 6, 7, 8, 9$ in respect of bias. This negative finding may be discouraging but not very decisive as our comparisons are based on the asymptotic expressions derived through Taylor linearization. However, as a counterpart to these analytical comparisons, we do carry out a simulation study in the next section with an objective to examine the overall performance of the different variance estimators. The performance measures of an estimator ν_i taken into consideration in this study are (i) *Absolute Relative Bias* $(ARB) = |B(\nu_i)|/S_y^2$, and (ii) *Percentage Relative Efficiency* $(PRE) = 100 \times V(s_y^2)/M(\nu_i), (i = 1, 2, \dots, 9)$

5. Description of the Simulation Study

Our simulation study involves repeated draws of simple random (without replacement) samples from 20 natural populations described in Table 1. 2,000 independent samples, for $n = 6, 8$ and 10, were selected from a population and for each sample several estimators were calculated. Then, considering 2,000 such combinations, simulated values of the performance measures were calculated and displayed in Tables 2 and 3. To save space, the numerical values of the performance measures for $n = 8$ and 10 are not shown, but the results based on these values are only reported. Major findings of the study are discussed in subsections 5.1 and 5.2.

5.1. Results Based on the ARB

The numerical values on the ARB reveal that there is no definite pattern in the performances of different estimators. The estimator ν_1 possesses the least ARB in 7 populations for $n = 6$ and in 6 populations for $n = 8$ and 10. ν_8 is found to have least ARB in 8, 10 and 11 populations for $n = 6, 8$ and 10 respectively. This clearly indicates that the overall performance of ν_8 improves with the increase in sample size. Searching for an estimator as the third choice is difficult owing to very erratic results in favor of the estimators (except ν_1 and ν_8).

Table 1: Description of the populations.

Pop	Source	N	y	x
1	Cochran (1977) p. 152	49	no of inhabitants in 1930	no. of inhabitants in 1920
2	Sukhatme & Sukhatme (1977) p. 185	34	area under wheat in 1937	area under wheat in1936
3	Sukhatme & Sukhatme (1977) p. 185	34	area under wheat in 1937	area under wheat in1931
4	Sampford (1962) p. 61	35	acreage under oats in 1957	acreage of crops and grass in 1947
5	Wetherill (1981) p. 104	32	yield of petroleum sprit	petroleum fraction end point
6	Murthy (1967) p. 398	43	no of absentees	no of workers
7	Murthy (1967) p. 399	34	area under wheat in 1964	cultivated area in 1961
8	Murthy (1967) p. 399	34	area under wheat in 1964	area under wheat in 1963
9	Steel & Torrie (1960) p. 282	30	leaf burn in secs.	percentage of potassium
10	Shukla (1966)	50	fiber yield	height of plant
11	Shukla (1966)	50	fiber yield	base diameter
12	Murthy (1967) p. 178	108	area under winter paddy	geographical area
13	Dobson (1990) p. 83	30	cholesterol	age in years
14	Dobson (1990) p. 83	30	cholesterol	body mass
15	Yates (1960) p. 159	25	measured volume of timber	eye estimated volume of timber
16	Yates (1960) p. 159	43	no. of absentees	total no. of persons
17	Panse & Sukhatme (1985) p. 118	25	progeny mean	parental plant value
18	Panse & Sukhatme (1985) p. 118	25	progeny mean	parental plot mean
19	Dobson (1990) p. 69	20	total calories from carbohydrate	calories as protein
20	Horvitz & Thompson (1952)	20	actual no. of households	eye estimated number of households

5.2. Results Based on the PRE

Results on the PRE of the competing estimators show that the estimator ν_8 is decidedly more efficient than the rest of the estimators in all populations for $n = 6$ and in 18 populations (except populations 1 and 17) for $n = 8$ and 10. Also the efficiency gain due to this estimator is noticeably high. The estimator ν_9 is found to be the second best estimator being more efficient than others (except ν_8) in 12 populations for $n = 6$ and in 10 populations for $n = 8$ and 10.

Further, it is observed that both ν_3 and ν_6 i.e., the estimators exploiting no auxiliary information, perform satisfactorily with ν_6 being better than ν_3 in all populations. It may also be noted here that for $n = 6, \nu_8$ is the only estimator using auxiliary variable x that is better than s_y^2 in all populations. However, this situation slightly changes with the increase in the sample size as it is worse than s_y^2 in one population for $n = 8$ and in two populations for $n = 10$. The estimators ν_1 ,ν_2, ν_4 and ν_5 do not fare well in most of the cases.

Table 2: ARB of the estimators for $n = 6$.

Pop No	ν_1	ν_2	ν_3	ν_4	ν_5	ν_6	ν_7	ν_8	ν_9
1	10.24	12.23	4.10	10.24	10.29	9.55	12.27	1.85	13.75
2	18.81	18.45	8.20	18.63	18.69	18.72	18.33	5.68	15.13
3	1.19	4.53	8.20	2.03	1.39	18.72	4.84	36.58	13.16
4	57.23	18.51	39.30	49.81	50.48	46.35	18.35	13.50	13.99
5	24.85	26.94	26.39	32.36	27.66	34.57	27.62	79.72	44.42
6	31.04	45.37	44.37	41.58	41.73	51.83	45.88	33.46	64.22
7	0.57	4.13	7.01	1.57	0.36	17.76	4.46	39.28	13.35
8	1.19	0.56	7.01	0.92	1.14	17.76	1.49	0.35	1.47
9	32.96	13.67	22.47	24.87	30.23	30.78	16.04	81.23	70.68
10	19.36	24.10	35.40	20.17	19.76	43.93	24.67	78.83	49.51
11	62.42	3.47	35.40	57.74	58.13	43.93	4.08	73.58	30.10
12	25.10	11.15	51.06	23.71	22.81	58.44	11.19	8.64	15.69
13	61.77	14.72	35.31	62.93	60.25	42.24	13.68	7.95	10.23
14	27.91	34.55	35.31	28.71	28.75	42.24	36.14	72.76	72.55
15	7.04	3.13	3.08	3.73	10.98	12.21	4.61	2.02	31.25
16	43.05	46.28	44.62	44.10	54.08	51.59	46.77	67.22	63.75
17	33.62	29.05	19.07	25.71	36.39	26.70	30.55	46.47	57.58
18	40.92	18.98	19.07	21.61	21.92	26.70	11.23	8.13	51.70
19	33.30	5.06	25.42	24.22	27.79	30.95	2.80	4.32	26.57
20	0.74	2.34	16.31	1.27	1.34	22.51	2.97	15.91	11.19

6. Conclusions

Our model-assisted analytical and simulated studies lead to an overall conclusion that the estimator ν_8 is preferable to others on the ground of efficiency. Although the analytical comparison fails to conclude which estimator is decidedly better than others on the ground of bias, the simulation study gives an indication that on this ground ν_8 is the better performer than other estimators. In view of these findings, if computational difficulty is not a matter of great concern, the variance estimator ν_8 may be considered as the most suitable estimator. Of course, these findings are only indicative and are no able to reveal essential features of the comparable estimators in a straightforward manner. Further investigations in this direction may be made for arriving at the conclusions.

Table 3: PRE of the estimators for $n = 6$.

Pop No	ν_1	ν_2	ν_3	ν_4	ν_5	ν_6	ν_7	ν_8	ν_9
1	285	223	104	226	283	138	228	395	246
2	419	444	106	427	416	135	444	531	403
3	270	261	106	281	271	135	262	425	301
4	21	68	106	29	28	135	68	313	68
5	122	127	106	154	129	135	129	646	176
6	191	197	104	196	195	137	201	816	368
7	410	370	106	437	414	135	372	804	440
8	958	908	106	985	960	136	911	1037	989
9	16	78	107	17	17	135	82	206	202
10	61	83	104	63	67	138	84	400	194
11	11	45	104	12	12	138	45	663	45
12	15	65	108	17	16	141	65	398	66
13	13	19	107	14	13	135	19	475	206
14	72	104	107	74	73	135	109	665	435
15	146	146	109	153	152	133	149	196	139
16	211	211	105	218	215	138	215	478	393
17	208	169	109	226	239	133	179	615	406
18	6	52	108	11	10	133	54	190	70
19	9	27	111	12	11	130	27	841	23
20	121	124	111	130	122	131	126	817	155

Acknowledgement

The authors are grateful to the referees and the Editor-in-Chief for providing some useful comments on an earlier draft of the paper.

[Recibido: diciembre de 2011 — Aceptado: noviembre de 2012]

References

Ahmed, M. S., Walid, A. D. & Ahmed, A. O. H. (2003), ‘Some estimators for finite population variance under two-phase sampling’, *Statistics in Transition* **6**, 143–150.

Basu, D. (1971), An essay on the logical foundations of survey sampling, *in* V. P. Godambe & D. A. Sprott, eds, ‘Foundations of Statistical Inference’, Vol. I, Holt, Rinehart and Wintson, Toronto, Canada, pp. 203–242.

Biradar, R. S. & Singh, H. P. (1998), ‘Predictive estimation of finite population variance’, *Calcutta Statistical Association Bulletin* **48**, 229–235.

Bolfarine, H. & Zacks, S. (1992), *Prediction Theory for Finite Populations*, Springer-Verlag.

Cochran, W. (1977), *Sampling Techniques*, Wiley Eastern Limited.

Das, A. & Tripathi, T. P. (1978), 'Use of auxiliary information in estimating the finite population variance', *Sankhyā* **C40**, 139–148.

Dobson, A. J. (1990), *An Introduction to Generalized Linear Models*, Chapman and Hall, New York.

Garcia, M. R. & Cebrain, A. A. (1996), 'Repeated substitution method: The ratio estimator for the population variance', *Metrika* **43**, 101–105.

Giiancarlo, D. & Chiara, T. (2004), 'Estimation for finite population variance in double sampling', *Metron* **62**, 223–232.

Grover, L. K. (2007), 'A wide and efficient class of estimators of population variance under sub-sampling scheme', *Model Assisted Statistics and Applications* **2**, 41–51.

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

Isaki, C. T. (1983), 'Variance estimation using auxiliary information', *Journal of the American Statistical Association* **78**(117-123).

Jhajj, H. S., Sharma, M. K. & Grover, L. K. (2005), 'An efficient class of chain estimators of population variance under sub-sampling scheme', *Journal of the Japan Statistical Society* **35**, 273–286.

Kadilar, C. & Cingi, H. (2006), 'Improvement in variance estimation using auxiliary information', *Hacettepe Journal of Mathematics and Statistics* **35**, 111–115.

Kadilar, C. & Cingi, H. (2007), 'Improvement in variance estimation in simple random sampling', *Communications in Statistics-Theory and Methods* **36**, 2075–2081.

Murthy, M. N. (1967), *Sampling Theory and Methods*, Statistical Publishing Society, Kolkata.

Nayak, R. K. (2009), Some Estimation Strategies in Finite Population Survey Sampling Using Auxiliary Information, PhD., Utkal University, India.

Panse, V. G. & Sukhatme, P. V. (1985), *Statistical Methods for Agricultural Workers*, Indian Council for Agricultural Research, New Delhi.

Prasad, B. & Singh, H. P. (1990), 'Some improved ratio-type estimators of finite population variance in sample surveys', *Communications in Statistics-Theory and Methods* **19**, 1127–1139.

Prasad, B. & Singh, H. P. (1992), 'Unbiased estimators of finite population variance using auxiliary information in sample surveys', *Communications in Statistics-Theory and Methods* **21**, 1367–1376.

Royall, R. M. (1988), The prediction approach to sampling theory, *in* P. R. Krishnaish & C. R. Rao, eds, 'Handbook of Statistics', Vol. 6, North Holland, pp. 351–358.

Sampford, M. R. (1962), *An Introduction to Sampling Theory*, Oliver and Boyd, Edinburg.

Sampford, M. R. (1978), Predictive estimation and internal congruency, *in* H. A. David, ed., 'Contribution to Survey Sampling and Applied Statistics', Academic Press, New York, pp. 29–39.

Shukla, G. K. (1966), 'An alternative multivariate ratio estimate for finite population', *Calcutta Statistical Association Bulletin* **15**, 127–134.

Singh, H. P. & Singh, R. (2001), 'Improved ratio-type estimators for variance using auxiliary information', *Journal of the Indian Society of Agricultural Statistics* **54**, 276–287.

Singh, S. & Kataria, P. (1990), 'An estimator of the finite population variance', *Journal of the Indian Society of Agricultural Statistics* **42**, 186–188.

Srivastava, S. K. (1983), 'Predictive estimation of finite population using product estimators', *Metrika* **30**, 93–99.

Srivastava, S. K. & Jhajj, H. S. (1980), 'A class of estimators using auxiliary information for estimating finite population variance', *Sankhyā* **C42**, 87–96.

Srivastava, S. K. & Jhajj, H. S. (1995), 'Classes of estimators of finite population mean and variance using auxiliary information', *Journal of the Indian Society of Agricultural Statistics* **47**, 119–128.

Steel, R. G. D. & Torrie, J. H. (1960), *Principles and Procedures of Statistics*, Mc. Graw Hill Book Company.

Sukhatme, P. V. & Sukhatme, B. V. (1977), *Sampling Theory of Surveys with Applications*, Asia Publishing House, New Delhi.

Wetherill, G. B. (1981), *Intermediate Statistical Methods*, Chapman and Hall, London.

Yates, F. (1960), *Sampling Methods for Censuses and Surveys*, Charls and Griffin, London.

Información para los autores

La **Revista Colombiana de Estadística** publica artículos originales de carácter teórico o aplicado en cualquiera de las ramas de la estadística. Los artículos puramente teóricos deberán incluir la ilustración de las técnicas presentadas con datos reales o por lo menos con experimentos de simulación, que permitan verificar la utilidad de los contenidos presentados. Se consideran también artículos divulgativos de gran calidad de exposición sobre metodologías o técnicas estadísticas aplicadas en diferentes campos del saber. Únicamente se publican artículos en español e inglés, si el autor escribe en una lengua diferente a la nativa debe enviar un certificado de un traductor oficial o de un corrector de estilo que haya revisado el texto.

El Comité Editor únicamente acepta trabajos para evaluación que no han sido publicados previamente y que no están siendo propuestos simultáneamente para publicación en otros medios, ni lo serán sin previo consentimiento del Comité, a menos que, como resultado de la evaluación, se decida no publicarlos en la Revista. Se supone además que cuando los autores hacen entrega de un documento con fines de publicación en la **Revista Colombiana de Estadística**, conocen las condiciones anteriores y que están de acuerdo con ellas.

Material

Los artículos remitidos a la **Revista Colombiana de Estadística** deben ser presentados en archivo PDF o PS, con textos, gráficas y tablas en color negro y, además, los autores deben agregar una versión del artículo sin nombres ni información de los autores, que se utilizará para el arbitraje. Se debe enviar una carta firmada por cada uno de los autores, donde manifiesten estar de acuerdo con someter el artículo y con las condiciones de la Revista. Si un artículo es aceptado, los autores deben poner a disposición del Comité Editorial los archivos: fuente en LaTeX y de gráficas en formato EPS en blanco y negro.

Para facilitar la preparación del material publicado se recomienda utilizar MiKTeX[1], usando los archivos de la plantilla y del estilo *revcoles* disponibles en la página Web de la Revista[2] y siguiendo las instrucciones allí incorporadas.

Todo artículo debe incluir:

- Título en español y su traducción al inglés.
- Los nombres completos y el primer apellido, la dirección postal o electrónica y la afiliación institucional de cada autor.
- Un resumen con su versión en inglés (*abstract*). El resumen en español no debe pasar de 200 palabras y su contenido debe destacar el aporte del trabajo en el tema tratado.

[1] http://www.ctan.org/tex-archive/systems/win32/miktex/

[2] http://www.estadistica.unal.edu.co/revista

- Palabras clave (*Key words*) en número entre 3 y 6, con su respectiva traducción al inglés, siguiendo las recomendaciones del *Current Index to Statistics* (CIS)[3].
- Cuando el artículo se deriva de una tesis o trabajo de grado debe indicarse e incluirse como una referencia.
- Si se deriva de un proyecto de investigación, se debe indicar el título del proyecto y la entidad que lo patrocina.
- Referencias bibliográficas, incluyendo solamente las que se hayan citado en el texto.

Referencias y notas al pie de página

Para las referencias bibliográficas dentro del texto se debe utilizar el formato autor-año, dando el nombre del autor seguido por el año de la publicación dentro de un paréntesis. La plantilla LaTeX suministrada utiliza, para las referencias, los paquetes BibTeX y Harvard[4]. Se recomienda reducir el número de notas de pie de página, especialmente las que hacen referencia a otras notas dentro del mismo documento y no utilizarlas para hacer referencias bibliográficas.

Tablas y gráficas

Las tablas y las gráficas, con numeración arábiga, deben aparecer referenciadas dentro del texto mediante el número correspondiente. Las tablas deben ser diseñadas en forma que se facilite su presentación dentro del área de impresión de la Revista. En este sentido, los autores deben considerar en particular la extensión de las tablas, los dígitos representativos, los títulos y los encabezados. Las gráficas deben ser visualmente claras y debe ser posible modificar su tamaño. Cuando el artículo sea aceptado para su publicación, los autores deben poner la versión definitiva a disposición del Comité Editorial. Todos los elementos como barras, segmentos, palabras, símbolos y números deben estar impresos en color negro.

Responsabilidad legal

Los autores se hacen responsables por el uso de material con propiedad intelectual registrada como figuras, tablas, fotografías, etc.

Arbitraje

Los artículos recibidos serán revisados por el Comité Editorial y sometidos a arbitraje por pares especializados en el tema respectivo. El arbitraje es “doble ciego” (árbitros anónimos para los autores y viceversa). El Comité Editorial decide aceptar, rechazar o solicitar modificaciones a los artículos con base en las recomendaciones de los árbitros.

[3] http://www.statindex.org/CIS/homepage/keywords.html

[4] http://tug.ctan.org/tex-archive/macros/latex/contrib/harvard

Revista Colombiana de Estadística
Índice de autores del volumen 35, 2012