Revista Colombiana de Estadística

Volumen 34. Número 3 - diciembre - 2011 **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://www.matematicas.unal.edu.co/revcoles
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. 34	N^o 3
ISSN 0120 - 1751	COLOMBIA	diciembre-2011	Págs. 403-588

Contenido

Editorial

Journal News and Francis Galton

Leonardo Trujillo[a]

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

Welcome to the third issue of the volume 34th of the Revista Colombiana de Estadistica (Colombian Journal of Statistics). This is the first time that this Journal is publishing three numbers in the same year. The first number was the regular one in June and the second was a Special Issue about Applications of Industrial Statistics with Professors Piedad Urdinola from the National University of Colombia and Jorge Romeu from the Syracuse University as Guest Editors. Then, this issue corresponds to the regular one for December 2011 and it has a very special connotation as being the first issue entirely published in English in our long history since 1968.

The reason is that we were the winners of an Internal Grant at the National University of Colombia (Universidad Nacional de Colombia) among many Journals in order to receive funding to publish entire issues in English and then, to strengthen its participation in international indexes according to the editorial policies of quality, visibility and impact of our published papers. Further information can be found at `http://www.dib.unal.edu.co/convocatorias/r20110803_revistas.html?ref=dibhome`. In this way, we will be only receiving papers in English language during the period of receiving this Grant until the end of 2012. We are repeating the successful experience of having three issues for the next year 2012 as we will be publishing a Special Issue about Biostatistics on July 2012 having as Guest Editors, Professors Piedad Urdinola and Liliana Lopez-Kleine. We are also happy to welcome new members in our Editorial and Scientific Committees: Professors Alex Rojas from Carnegie Mellon in Qatar and Liliana Lopez-Kleine from the National University of Colombia.

The topics in this current issue range over diverse areas of statistics: five papers in Probability by Almaraz; Bran-Cardona, Orozco-Castaneda and Nagar; Fierro and Tapia and Ozel; two more papers in Survey Sampling by Gutierrez and Zhang and by Soberanis and Miranda; one paper in Biostatistics by Tovar and Achcar; one paper in Econometrics by Gomez and Gallon and one paper in Industrial Statistics by Gonzalez and Bueno.

I would not like to finish this Editorial without paying a tribute for the 100 years of the death of Sir Francis Galton (1822-1911). He was not only a statistician, also an anthropologist, geographer, inventor, meteorologist and psychometrician (Forrest 1974). Galton founded many concepts in statistics, among them

[a] General Editor of the Colombian Journal of Statistics, Assistant Professor. E-mail: ltrujilloo@bt.unal.edu.co

correlation, percentile, quartile and widely promoted regression toward the mean (Galton 1886, Bulmer 2003). Galton compared the height of children to that of their parents and he found that adult children are closer to average height than their parents are. Galton's later statistical study of the probability of extinction of surnames led to the concept of Galton-Watson stochastic processes.

He was the first to apply statistical methods to the study of human differences and inheritance of intelligence, and introduced the use of questionnaires and surveys for collecting demographic and social data for anthropometric, biographical and genealogical studies (Senn 2003). In one of these studies, he asked to describe mental images to fellow members of the Royal Society. In another one, he collected surveys in order to study the effects of nature and nurture on the propensity toward scientific thinking from eminent scientists (Clauser 2007).

The idea that data have a central tendency or mean but also a deviation around this central value or the variance is core to any statistical analysis. Galton conceived the idea of a standardized measure, the standard deviation, on the late 1860s.

The year 2011 is coming to its end; however, for statisticians around the world is going to be remembered as the Galton year - a celebration of Francis Galton, a genius -. However, he was not a very well-known one. His cousin, Charles Darwin was more famous. Despite of this, he did many surprising things: he was the first person to use fingerprints in detective work and the first to publish a weather map in a newspaper in 1875 (Jones 2011).

Referencias

Bulmer, M. (2003), *Francis Galton: Pioneer of Heredity and Biometry*, Johns Hopkins University Press.

Clauser, B. E. (2007), 'The life and labours of Francis Galton: A review of four recent books about the father of behavioural statistics', *Journal of Educational and Behavioral Statistics* **32**(4), 440–444.

Forrest, D. W. (1974), *Francis Galton: The Life and Work of a Victorian Genius*, Paul Elek, London.

Galton, F. (1886), 'Regression towards mediocrity in hereditary stature', *Journal of the Anthropological Institute of Great Britain and Ireland* **15**, 246–263.

Jones, S. (2011), 'Francis Galton: The man who drew up the 'ugly map' of Britain', BBC News.
*http://www.bbc.co.uk/news/magazine-13775520

Senn, S. J. (2003), *Dicing with Death*, Cambridge University Press, Cambridge.

Editorial

Noticias de la Revista y Francis Galton

Leonardo Trujillo[a]

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

Me es grato presentar el tercer número del volumen 34 de la Revista Colombiana de Estadística. Esta es la primera vez que esta Revista publica tres números en un mismo año. El primer número fue el regular del mes de junio y el segundo correspondió a un Numero Especial en Aplicaciones de la Estadística en la Industria con los profesores Piedad Urdinola de la Universidad Nacional y Jorge Romeu de Syracuse University como Editores Invitados. El presente número corresponde al regular de diciembre de 2011 y tiene la connotación especial de ser el primer número enteramente publicado en idioma inglés en la historia de la Revista desde 1968.

La razón de este nuevo formato es que hemos sido los ganadores de una Convocatoria Interna en la Universidad Nacional de Colombia entre otras revistas con el fin de recibir financiamiento para publicar ediciones enteras en ingles y fortalecer la participación en índices internacionales de acuerdo con las políticas editoriales de calidad, impacto y visibilidad de nuestros artículos publicados. Más información se puede encontrar en la página web `http://www.dib.unal.edu.co/convocatorias/r20110803_revistas.html?ref=dibhome`. De esta manera, estaremos recibiendo solo artículos en ingles durante el periodo de la convocatoria hasta finales de 2012 y estaremos repitiendo la exitosa experiencia de tener tres números por volumen para el próximo año 2012 cuando se publicara un Numero Especial en Bioestadística en el mes de julio, teniendo como Editoras Invitadas a las profesoras Liliana Lopez-Kleine y Piedad Urdinola. Queremos también dar la bienvenida a algunos miembros nuevos de los Comités Científico y Editorial: los profesores Alex Rojas de la Universidad Carnegie Mellon en Qatar y Liliana Lopez-Kleine de la Universidad Nacional de Colombia.

Los tópicos del presente número abarcan diferentes áreas de la estadística: cinco artículos en Probabilidad escritos por Almaraz; Bran-Cardona, Orozco-Castañeda y Nagar; Fierro y Tapia y Ozel; dos artículos más en Muestreo por Gutiérrez y Zhang y por Soberanis y Miranda; un artículo en Bioestadística por Tovar y Achcar; un artículo en Econometría por Gómez y Gallón y finalmente uno en Estadística Industrial por González y Bueno.

No quisiera terminar esta Editorial sin rendir un tributo por la celebración de los 100 años de la muerte de Francis Galton (1822-1911). El no fue solamente un estadístico, sino también antropólogo, geógrafo, inventor, meteorólogo y psicometrista (Forrest 1974). Galton fue el fundador de muchos conceptos en estadística,

[a] Editor de la Revista Colombiana de Estadística, Profesor asistente.
E-mail: ltrujilloo@bt.unal.edu.co

entre ellos la definición de correlación, cuantil, percentil y la ampliamente difundida regresión hacia la media (Galton 1886, Bulmer 2003). Galton comparó la altura de los hijos a la de sus padres y encontró que cuando los niños se hacían adultos el promedio de sus alturas era cercano a la altura promedio de sus padres. Un estudio posterior de la extinción de algunos apellidos conllevó al concepto de los procesos estocásticos de Galton-Watson.

Galton fue el primero en aplicar métodos estadísticos para el estudio de las diferencias entre humanos y la herencia de la inteligencia e introdujo el uso de cuestionarios y encuestas para recolectar datos de tipo demográfico y social en estudios antropométricos, biográficos y genealógicos (Senn 2003). En uno de estos estudios, pidió a sus colegas miembros de la Sociedad Real el describir imágenes mentales obtenidas ante ciertos estímulos. En otro estudio, recolectó información para medir los efectos de cualidades innatas (nature) y cualidades aprendidas (nurture) en la probabilidad de desarrollar pensamiento científico en científicos eminentes (Clauser 2007).

La idea que los datos tienen una tendencia central o media pero también una desviación alrededor de esta medida central o varianza es el núcleo de cualquier análisis estadístico. Galton concibió la idea de una medida estandarizada, la desviación estándar a finales de 1860.

El año 2011 se acerca rápidamente a su final; sin embargo, para los estadísticos alrededor del mundo el 2011 será recordado como el año de Galton - una celebración por su ingenio. Sin embargo, no fue un genio famoso. Incluso, su primo Charles Darwin lo fue mucho más que él. A pesar de esto, Galton hizo muchas cosas sorprendentes: fue la primera persona en usar las huellas digitales en el trabajo de los detectives y el primero en publicar un mapa con el estado del tiempo en un periódico en 1875 (Jones 2011).

Referencias

Bulmer, M. (2003), *Francis Galton: Pioneer of Heredity and Biometry*, Johns Hopkins University Press.

Clauser, B. E. (2007), ‘The life and labours of Francis Galton: A review of four recent books about the father of behavioural statistics’, *Journal of Educational and Behavioral Statistics* **32**(4), 440–444.

Forrest, D. W. (1974), *Francis Galton: The Life and Work of a Victorian Genius*, Paul Elek, London.

Galton, F. (1886), ‘Regression towards mediocrity in hereditary stature’, *Journal of the Anthropological Institute of Great Britain and Ireland* **15**, 246–263.

Jones, S. (2011), ‘Francis Galton: The man who drew up the ‘ugly map’ of Britain’, BBC News.
*http://www.bbc.co.uk/news/magazine-13775520

Senn, S. J. (2003), *Dicing with Death*, Cambridge University Press, Cambridge.

Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 403 a 420

Hierarchical Design-Based Estimation in Stratified Multipurpose Surveys

Estimación jerárquica basada en el diseño muestral para encuestas estratificadas multi-propósito

Hugo Andrés Gutiérrez[a], Hanwen Zhang[b]

Centro de Investigaciones y Estudios Estadísticos (CIEES), Facultad de Estadística, Universidad Santo Tomás, Bogotá, Colombia

Abstract

This paper considers the joint estimation of population totals for different variables of interest in multi-purpose surveys using stratified sampling designs. When the finite population has a hierarchical structure, different methods of unbiased estimation are proposed. Based on Monte Carlo simulations, it is concluded that the proposed approach is better, in terms of relative efficiency, than other suitable methods such as the generalized weight share method.

***Key words*:** Design based inference, Finite population, Hierarchical population, Stratified sampling.

Resumen

Este artículo considera la estimación conjunta de totales poblacionales para distintas variables de interés en encuestas multi-propósito que utilizan diseños de muestreo estratificados. En particular, se proponen distintos métodos de estimación insesgada cuando el contexto del problema induce una población con una estructura jerárquica. Con base en simulaciones de Monte Carlo, se concluye que los métodos de estimación propuestos son mejores, en términos de eficiencia relativa, que otros métodos de estimación indirecta como el recientemente publicado método de ponderación generalizada.

***Palabras clave*:** inferencia basada en el diseño, población finita, población jerárquica, muestreo estratificado.

[a] Lecturer. E-mail: hugogutierrez@usantotomas.edu.co
[b] Lecturer. E-mail: hanwenzhang@usantotomas.edu.co

1. Background

The reality of surveys is complex; as Holmberg (2002) states, most of the real applications in survey sampling involve not one, but several characteristics of study; and as Goldstein (1991) claims, real populations have hierarchical structures. Moreover, in certain occasions, the survey methodologist is faced with the estimation of several parameters of interest in different levels of the population and he/she is commanded with the seeking of proper approaches to estimate those parameters as required in the study. The problem of proposing sampling strategies (optimal sampling design and efficient estimators) that contemplate joint estimation of several parameters in multipurpose survey has been widely discussed in recent statistical literature. Although there is a vast number of papers about estimation of hierarchical populations (Gelman & Hill 2006) and model-based (or model-assisted) multilevel survey data (Skinner, Holt & Smith 1989, Lehtonen & Veijanen 1999, Goldstein 2002, Rabe-Hesketh & Skrondal 2006), the design-based estimation for finite populations with hierarchical structures seems to be omitted by survey statisticians. The aim of this paper is to provide a multipurpose approach to the joint estimation of several parameters for different variables in a stratified finite population with two levels.

Next are detailed some clarifying ideas concerning the concept of hierarchical structures in finite populations. Many kinds of data have a hierarchical or clustered structure. Note that in biological studies it is natural to think in a hierarchy where the offspring of the races is clustered into families; in educational surveys, students belong to schools and schools belong to districts, and so on; in social studies, a person belongs to a household and households are grouped geographically. In this paper, the concept of hierarchy is related with the multipurpose approach in the sense that the survey statistician often needs to make inferences on different levels of the finite population. For example, consider an establishment survey. It would be of interest to estimate the total sales of the market sections of the stores in detail (sales by toys, grocery, electronics or pharmacy sections) and at the same time it would be of interest to estimate the number of employees working in the stores. It is clear that the multipurpose approach is given by the joint inference of two different study variables (sales by market section and number of employees in the stores) but these variables of interest are in different levels of the population: sales are related with the market section level and the number of employees with the store level. Note that as the market sections belong to the stores, then the set of all market sections defines the second level and the set of all stores defines the first level.

In some occasions, it is impossible to obtain a sampling frame for the first level, however this is available for the second level. For example, Särndal, Swensson & Wretman (1992, example 1.5.1) reports on the Swedish household survey where there is not a good complete list of households and the sampling frame used was the Swedish Register of the Total Population, which is a list of individuals. In this case, the first level is composed of households, the second level is composed of individuals and the inferences about households are induced directly from the population of individuals. If the requirements of that survey were to obtain inferences about both households and individuals, then it would be a clear example of a study involving multipurpose estimation within a hierarchical structure in the finite population, with the restriction that the sampling frame is only available in the second level. In other cases, it is possible that both sampling frames are available in the design stage. However, if the requirements of the survey are focused in the estimation of the population totals in both levels, the most trivial, but in some cases useless, solution would be planning two sampling designs. In this paper we propose another solution requiring just the use of a sampling frame in order to simultaneously estimate several parameters for different study variables in two different levels of a stratified population, when the sampling frame to be used is related with the units of the second level. Note that, since the sampling frame is not available (or available but useless) in the first level, sampling designs such as cluster, or multi-stage sampling designs are no longer valid to solve this kind of problems.

The outline of this paper is as follows: after a brief introduction explaining the hierarchical concept, different levels of estimation in such populations, and its implications in the survey sampling context; Section 2, explains in detail, by means of a simple example, the foundations of the hierarchical finite population and the issue of this paper. Section 3, refers to the proposal of an indirect estimation in the first level involving different variables of interest than those considered in the second level. This approach is based on the computation of the first and second order inclusion probabilities, given by the induced sampling design in the first level, using the principles of the well-known Horvitz-Thompson and Hájek estimators for a population total. Besides, in this section, the authors show how this problem is related with the indirect sampling approach (Lavallée 2007). This section also presents a simple case study to illustrate the procedures of the proposed approach in the case of simple random stratified sampling (STSI) in the second level. In Section 4, we present an empirical study based on several Monte Carlo simulations that show how our proposal outperforms, in the sense of relative efficiency, other methods of indirect estimation such as the generalized weight share method (indirect sampling). Finally, some recommendations and conclusions are given in Section 5.

2. Multipurpose Estimation

Let $U = \{1, \dots, k, \dots, N\}$ denote the second level finite population of N elements in which a sampling frame is available. Suppose that the sampling frame is stratified and for each element $k \in U$ the stratum to which k belongs is completely identified by means of some discrete auxiliary variable. That is, the population U is partitioned into H subsets $U_1, U_2, ..., U_H$ called strata, where

$$\bigcup_{h=1}^{H} U_h = U, \qquad U_h \bigcap U_{h'} = \emptyset \quad \text{for all } h \neq h'$$

On the other hand, assume that each element $k \in U$ in the second level belongs to a unique cluster in the first level. It is assumed that there exist N_I clusters denoted by $U_1, \dots, U_i, \dots, U_{N_I}$. This set of clusters is symbolically represented as $U_I = \{1, \dots, i, \dots, N_I\}$. This way, the first level population is U_I, the second level population is U and, clearly, the data show a notorious hierarchical structure.

Although there is an available sampling frame for U, suppose that it is impossible to obtain a frame for the population of the first level U_I and that the requirements of the survey imply the inference of parameters, say population totals or means, for both levels. Hence, it is assumed that there are two variables of interest, say, y in the second level, and z in the first level, and it is requested the estimation of both population totals, defined by

$$t_y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k$$

and

$$t_z = \sum_{i \in U_I} z_i$$

In this paper, the notation of any pair of elements in the second level will be denoted by the letters k and l; meanwhile for the units in the first level, the letters i and j will be used.

By taking advantage of the sampling frame in the second level, a stratified sample s is drawn. For each $k \in s$, the value of the variable of interest y_k is observed. Besides, it is supposed that unit k can also provide the information of its corresponding cluster, say U_i. This way, the value of the other variable of interest z_i is recorded. Note that for a particular second level sample there exists a corresponding set of units in the first level. In other words, the second level sample s induces a set, contained in the first level population, which will be called the first level sample, denoted by m and given by

$$m = \{i \in U_I \mid \text{at least one unit of the cluster } U_i \text{ belong to } s\}$$

In summary, the values of both variables of interest could be recorded ar the same time: y_k for the elements in the selected sample; s and z_i for the clusters in the induced sample m. As an example, consider the finite population showed in Table 1. The second level population, denoted by $U = \{A1, B1, D1, \dots, D4, E4\}$ of size $N = 15$ is a set of market sections in different stores. This population is stratified in four sections ($H = 4$). The population of the first level is hence $U_I = \{A, B, C, D, E\}$ with $N_I = 5$. Each stratum is present in different clusters. For example, Section 1 is present in four stores, whereas Section 3 is present in three stores. Notice that it is not required that each stratum be present in all of the clusters.

Following with the example, when a sample s is drawn, an interviewer visits the selected market section, say k, records the value of y_k and also obtains the information about z_i, the value of the variable of interest in the cluster that contains that section. Table 2, reports the first and second level population values for the variables of interest. If the sampling design is such that only one element

Table 1: Description of a possible hierarchical configuration.

	Section 1	Section 2	Section 3	Section 4
Store A	A1	A2	-	A4
Store B	B1	-	B3	-
Store C	-	C2	-	C4
Store D	D1	D2	D3	D4
Store E	E1	E2	E3	E4

of each section is selected, then a possible sample in the second level would be $s = \{A1, E2, B3, E4\}$. This way, the recorded values for this specific sample correspond to 32, 33, 26, 55 and the induced first level sample would be $m = \{A, B, E\}$ and the values of the variable of interest in this level correspond to 14.12, 10.25 and 24.81, respectively. Note that a store may be selected more than once; however, following Särndal et al. (1992, section 3.8), we omit the repeated information in the first level and carry out the inference by using the reduced sample. The parameter of interest in the first level is $t_z = 14.12 + 10.25 + 17.52 + 22.58 + 24.81 = 89.28$ and the parameter of interest in the second level is $t_y = 106 + 105 + 68 + 162 = 441$.

Table 2: Variables of interest in a possible hierarchical configuration.

Y_1	Y_2	Y_3	Y_4	Z
$y_{A1} = 32$	$y_{A2} = 12$	-	$y_{A2} = 51$	$Z_A = 14.12$
$y_{B2} = 18$	-	$y_{B3} = 26$	-	$Z_B = 10.25$
-	$y_{C2} = 36$	-	$y_{C4} = 10$	$Z_C = 17.52$
$y_{D1} = 42$	$y_{D2} = 24$	$y_{D3} = 14$	$y_{D4} = 46$	$Z_D = 22.58$
$y_{E1} = 14$	$y_{E2} = 33$	$y_{E3} = 28$	$y_{E4} = 55$	$Z_E = 24.81$

As stated at the beginning of this section, the second level population U is stratified into H strata. In each stratum h $(h = 1, \ldots, H)$ a sampling design $p_h(\cdot)$ is applied and a sample s_h is drawn. An important feature of stratified sampling design is the independence between selections. For this reason, the sampling design takes the following form

$$p(s) = \prod_{h=1}^{H} p_h(s_h) \qquad \text{where} \qquad s = \bigcup_{h=1}^{H} s_h$$

We have that an unbiased estimator of t_y and its variance are given by

$$\hat{t}_{y\pi} = \sum_{h=1}^{H} \sum_{s_h} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \hat{t}_{h\pi} \tag{1}$$

$$V(\hat{t}_{y_\pi}) = \sum_{h=1}^{H} V_h(\hat{t}_{h\pi}) = \sum_{h=1}^{H} \sum_{k \in U_h} \sum_{l \in U_h} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, and $\hat{t}_{h\pi}$ corresponds to the Horvitz-Thompson estimator in the h-th stratum, defined by

$$\hat{t}_{h\pi} = \sum_{s_h} \frac{y_k}{\pi_k}$$

In the case that the sample design is simple random sampling carried out along the strata, the first and second order inclusion probabilities are given by

$$\pi_k = P(k \in s) = P(k \in s_h) = \frac{n_h}{N_h}$$

And

$$\pi_{kl} = \begin{cases} \frac{n_h}{N_h} & \text{if } k = l \\ \frac{n_h}{N_h} \frac{n_h - 1}{N_h - 1} & \text{if } k \neq l, \text{ with } k, l \in h \\ \frac{n_h}{N_h} \frac{n_{h'}}{N_{h'}} & \text{if } k \neq l, \text{ with } k \in h \text{ y } l \in h' \end{cases}$$

where N_h and n_h denote the population size and the sample size in the stratum h, respectively.

3. Estimation in the First Level

In this section, we develop the proposed approach in order to estimate the parameter of interest in the first level and we point out that another suitable approach could be used to solve this kind of estimation problems, namely the Generalized Weight Share Method (GWSM) (Deville & Lavallée 2006). However, as it will be confirmed later, in the simulation report of Section 4, our proposal is more efficient than the GWSM.

3.1. Proposed Approach

Recalling that the second level sample s induces a first level sample m, we can obtain the induced sampling design as stated in the following result.

Result 1. *The sampling design in the first level induced by the stratified sample s is given by*

$$p(m) = \sum_{\{s:\ s \rightarrow m\}} \prod_{h=1}^{H} p_h(s_h) \tag{2}$$

where the notation $s \rightarrow m$ indicates that the second level sample s induces the first level sample m.

Proof. Considering that even though a particular first level sample m may be induced by different samples in the second level, it is clear that a second level sample s may only induce a unique first level sample m, then we have that

$$\begin{aligned} p(m) &= \sum_{\{s:\ s\rightarrow m\}} p(s) \\ &= \sum_{\{s:\ s\rightarrow m\}} \prod_{h=1}^{H} p_h(s_h) \end{aligned}$$

The last equation follows because of the independence in the selection of s_h for $h = 1, \ldots, H$. □

For example, continuing with the population described in Table 1, if the sampling design in the second level is simple random sampling in each stratum such that $N_3 = 3$, $N_1 = N_2 = N_4 = 4$ and $n_h = 1$ for $h = 1, 2, 3, 4$, then in order to compute the selection probability of the particular first level sample $m = \{A, B\}$, it is necessary to find all of the second level samples inducing that specific sample m. Given the data structure, the set $\{s : \ s \rightarrow m\}$ has only two second level samples; these samples are: $\{A1, A2, B3, A4\}$ and $\{B1, A2, B3, A4\}$. For that m, we have that its selection probability corresponds to

$$\begin{aligned} p(m) &= p(\{A1, A2, B3, A4\}) + p(\{B1, A2, B3, A4\}) \\ &= \prod_{h=1}^{4} \frac{1}{N_h} + \prod_{h=1}^{4} \frac{1}{N_h} = \frac{1}{96} = 0.0104 \end{aligned}$$

Given that one parameter of interest is the population total of the variable z in the first level, we can obtain the first and second order inclusion probability of clusters in U_I in order to propose some estimators for t_z. These inclusion probabilities are given in the following results.

Result 2. *The first order inclusion probability of the cluster U_i, denoted by π_i, is given by*

$$\pi_i = Pr(i \in m) = 1 - \prod_{h=1}^{H} q_h^{(i)} \tag{3}$$

where $q_h^{(i)} = Pr($None of the units of U_i belongs to $s_h)$ and s_h denotes the selected sample in the stratum U_h, for $h = 1, \ldots, H$.

***Proof*.**

$$\begin{aligned} \pi_i = Pr(i \in m) &= Pr(\text{At least one unit of } U_i \text{ belongs to } s) \\ &= 1 - Pr(\text{None of the units of } U_i \text{ belongs to } s) \\ &= 1 - \prod_{h=1}^{H} q_h^{(i)} \end{aligned}$$

□

***Note* 1.** Note that the computation of the quantities $q_h^{(i)}$ depends on the sampling design used in each stratum. Moreover, if $a_h^{(i)}$ denotes the number of units of cluster U_i belonging to stratum U_h, then $a_h^{(i)} \geq 0$. Which implies that each cluster is not necessarily present in each stratum.

***Note* 2.** The stratified sampling design on the second level population implies independence across strata. However, depending on the sampling design used within each stratum, the independence of units selection may not be guaranteed. For example, in the case of simple random sampling designs, there is no independence. On the other hand, other sampling designs such as Bernoulli and Poisson do provide that independence feature.

Result 3. *The second order inclusion probability for any pair of clusters U_i, U_j is given by*

$$\pi_{ij} = 1 - \prod_{h=1}^{H} q_h^{(i)} - \prod_{h=1}^{H} q_h^{(j)} + \prod_{h=1}^{H} q_h^{(ij)} \tag{4}$$

With $q_h^{(ij)} = Pr$(None of the units of U_i belongs to s_h and none of the units of U_j belongs to s_h) and $q_h^{(i)}$, $q_h^{(j)}$ are defined analogously in Result 3.2.

***Proof*.** After some algebra, we have that

$$\begin{aligned}
\pi_{ij} &= Pr(i \in m, j \in m)\\
&= 1 - Pr(i \notin m \text{ or } j \notin m)\\
&= 1 - [Pr(i \notin m) + Pr(j \notin m) - Pr(i \notin m, j \notin m)]\\
&= 1 - [(1-\pi_i) + (1-\pi_j) - Pr(i \notin m, j \notin m)]\\
&= 1 - \prod_{h=1}^{H} q_h^{(i)} - \prod_{h=1}^{H} q_h^{(j)} + Pr(i \notin m, j \notin m)\\
&= 1 - \prod_{h=1}^{H} q_h^{(i)} - \prod_{h=1}^{H} q_h^{(j)} + \prod_{h=1}^{H} q_h^{(ij)}
\end{aligned}$$

□

Once these inclusion probabilities are computed, it is possible to estimate t_z by means of the well known Horvitz-Thompson estimator given by

$$\hat{t}_{z\pi} = \sum_{i \in m} \frac{z_i}{\pi_i} \tag{5}$$

Note that $\hat{t}_{z\pi}$ is unbiased for t_z and, if the stratified sampling design in the second level is such that $n_h \geq 2$ for $h = 1, \ldots, H$, its variance is given by

$$V(\hat{t}_{z_\pi}) = \sum_{i \in U_I} \sum_{j \in U_I} \Delta_{ij} \frac{z_i}{\pi_i} \frac{z_j}{\pi_j}$$

Where $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$. However, since the first level sample is induced by the second level sample, the size of m is random, even when the stratified sample design of the second level is of fixed size. For a more detailed discussion about the randomness of the sample size and its effects when a Horvitz-Thompson estimator is used, an interested reader can see Särndal et al. (1992, Example 5.7.3 and Example 7.4.1). In order to avoid extreme estimates, sometimes obtained with the previous estimator, and taking into account that N_I is known, we propose to use the expanded sample mean estimator (denoted in this paper as Hájek estimator) given by

$$\widetilde{t}_z = N_I \frac{\widehat{t}_{z\pi}}{\widehat{N}_{I,\pi}} \tag{6}$$

Where $\widehat{N}_{I,\pi} = \sum_{i\in m} \frac{1}{\pi_i}$. It is well known that its approximate variance is given by

$$AV(\widetilde{t}_z) = \sum_{i\in U_I}\sum_{j\in U_I} \Delta_{ij} \frac{z_i - \overline{z}_{U_I}}{\pi_i}\frac{z_j - \overline{z}_{U_I}}{\pi_j} \tag{7}$$

With $\overline{z}_{i\in\ U_I} = \sum_{U_I} z_i/N_I$. For more comprehensive details, see Gutiérrez (2009, expressions 9.3.7. and 9.3.9.) and Särndal et al. (1992, expression 7.2.10.).

3.1.1. Some Particular Cases

In the case that in each stratum of the second level population a Bernoulli sampling design is used, with the same inclusion probability θ across the strata, then the first order inclusion probability for a cluster U_i is given by

$$\begin{aligned}\pi_i = 1 - \prod_{h=1}^{H} q_h^{(i)} &= 1 - \prod_{h=1}^{H}(1-\theta)^{a_h^{(i)}}\\ &= 1-(1-\theta)^{\sum_{h=1}^{H} a_h^{(i)}} = 1-(1-\theta)^{N_i}\end{aligned}$$

Where $N_i = \#(U_i)$. The second order inclusion probability for clusters U_i and U_j is given by

$$\begin{aligned}\pi_{ij} &= 1 - \prod_{h=1}^{H} q_h^{(i)} - \prod_{h=1}^{H} q_h^{(j)} + \prod_{h=1}^{H} q_h^{(ij)}\\ &= 1-(1-\theta)^{N_i} - (1-\theta)^{N_j} + \prod_{h=1}^{H}(1-\theta)^{a_h^{(i)}+a_h^{(j)}}\\ &= 1-(1-\theta)^{N_i} - (1-\theta)^{N_j} + (1-\theta)^{N_i+N_j}\end{aligned}$$

Other interesting case is carrying out simple random sampling in each stratum. This way, the resulting formulaes for the proposed approach are quite simple. Denoting the population size and the sample size in the h-th stratum by N_h and

n_h, respectively, and by following the assumptions of the Result 3.2, the first inclusion probability for a cluster U_i is given in terms of $q_h^{(i)}$, where

$$q_h^{(i)} = \begin{cases} \frac{\binom{N_h - a_h^{(i)}}{n_h}}{\binom{N_h}{n_h}}, & \text{if } n_h \leq N_h - a_h^{(i)} \\ 0, & \text{otherwise} \end{cases}$$

On the other hand, for the computation of the second order inclusion probability for clusters U_i and U_j, we have that

$$q_h^{(ij)} = \begin{cases} \frac{\binom{N_h - a_h^{(i)} - a_h^{(j)}}{n_h}}{\binom{N_h}{n_h}}, & \text{if } n_h \leq N_h - a_h^{(i)} - a_h^{(j)} \\ 0, & \text{otherwise} \end{cases}$$

For example, following the finite population in Table 1, the first inclusion probabilities of the store A and store B are given by

$$\begin{aligned} \pi_{store_{(A)}} &= 1 - \left(1 - \frac{n_1}{N_1}\right)\left(1 - \frac{n_2}{N_2}\right)\left(1 - \frac{n_4}{N_4}\right) \\ \pi_{store_{(B)}} &= 1 - \left(1 - \frac{n_1}{N_1}\right)\left(1 - \frac{n_3}{N_3}\right) \end{aligned}$$

And the second order inclusion probability for these two stores is given by

$$\begin{aligned} \pi_{store_{(A)},store_{(B)}} = 1 &- \left(1 - \frac{n_1}{N_1}\right)\left(1 - \frac{n_2}{N_2}\right)\left(1 - \frac{n_4}{N_4}\right) - \left(1 - \frac{n_1}{N_1}\right)\left(1 - \frac{n_3}{N_3}\right) \\ &+ \frac{(N_1 - n_1)}{N_1}\frac{(N_1 - n_1 - 1)}{(N_1 - 1)}\left(1 - \frac{n_2}{N_2}\right)\left(1 - \frac{n_3}{N_3}\right)\left(1 - \frac{n_4}{N_4}\right) \end{aligned}$$

Once the inclusion probabilities are computed, it is possible to obtain estimations of t_z, by using (5) and (6), along with its respective estimated coefficients of variation by means of the expression for the estimated variances.

3.2. Indirect Sampling

This kind of situations can also be handled by using the indirect sampling approach (Lavallée 2007). We introduce it briefly: it is assumed that the first level population U_I is related to the second level population U through a link matrix representing the correspondence between the elements of U_I and U. Since there is no available sampling frame for U_I, an estimate for t_z can be obtained indirectly using a sample from U and the existing links between the two populations. The link matrix is denoted by $\mathbf{\Theta}$ with size $N \times N_I$, and the ki-th element of the matrix $\mathbf{\Theta}$ is defined as

$$[\mathbf{\Theta}]_{ki} = \begin{cases} 1 & \text{if the element } k \text{ is related with the cluster } U_i \\ 0 & \text{otherwise} \end{cases}$$

for $k = 1, \dots, N$, $i = 1, \dots, N_I$.

The formulation of the standardized link matrix is needed to carry out the estimation of t_z. This matrix is defined as

$$\widetilde{\mathbf{\Theta}} = \mathbf{\Theta}[diag(\mathbf{1}'_N\mathbf{\Theta})]^{-1}$$

where $\mathbf{1}_N$ is the vector of ones of dimension N. It can be shown that $\widetilde{\mathbf{\Theta}}\mathbf{1}_N = \mathbf{1}_{N_I}$. This way, the population total t_z can be expressed as

$$t_z = \mathbf{1}'_{N_I}\mathbf{z} = \mathbf{1}'_N\widetilde{\mathbf{\Theta}}\mathbf{z}$$

Where $\mathbf{z} = (z_1, \dots, z_{N_I})$. By using the previous expression and taking into account the principles of GWSM, as pointed in Deville & Lavallée (2006), we have the following estimator:

$$\widehat{t}_z = \mathbf{1}'_N\mathbf{I}_N\mathbf{\Pi}_N^{-1}\widetilde{\mathbf{\Theta}}\mathbf{z} \tag{8}$$

where $\mathbf{\Pi}_N = diag(\pi_1, \dots, \pi_N)$, is a matrix of dimension $N \times N$ that contains the inclusion probabilities for all the elements in the second level population and $\mathbf{I}_N$ is the diagonal matrix containing the indicator variables I_k for the membership of elements in the second level sample s. Note that (8) may be expressed as

$$\widehat{t}_z = \mathbf{w}\mathbf{z}$$

where $\mathbf{w} = \mathbf{1}'_N\mathbf{I}_N\mathbf{\Pi}_N^{-1}\widetilde{\mathbf{\Theta}}$. We can see that the elements of $\mathbf{w}$ are given by

$$w_i = \begin{cases} \sum_{k\in U} I_k \dfrac{\widetilde{\mathbf{\Theta}}_{ki}}{\pi_k}, & \text{if } i \in m \\ 0, & \text{if } i \notin m \end{cases}$$

for $i = 1, \dots, N_I$. Note that $\widehat{t}_z$ is a weighted sum upon all units in the induced sample m of U_I.

Deville & Lavallée (2006) have shown that $\widehat{t}_z$ is an unbiased estimator for t_z and its variance is given by

$$V(\widehat{t}_z) = \mathbf{z}'\mathbf{\Delta}_{N_I}\mathbf{z}$$

with $\mathbf{\Delta}_{N_I} = \widetilde{\mathbf{\Theta}}'\mathbf{\Delta}_N\widetilde{\mathbf{\Theta}}$, where the kl-th element of $\mathbf{\Delta}_N$ is given by

$$[\mathbf{\Delta}_N]_{kl} = \frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l}$$

for $k, l = 1, \dots, N$.

It is important to comment that despite the resulting inferences of indirect sampling from the GSWM are defined for the first level population, they are directly induced by the probability measure of the sampling design in the second level $p(s)$. However, the inferences from our proposed approach are given directly by the induced sampling design of the first level $p(m)$.

4. Simulation Study

In this section, by means of Monte Carlo simulations, we compare the performance of the two proposed estimators given by (5) and (6) and the indirect sampling estimator. We simulate several stratified populations with hierarchical structure where all clusters are presented in each stratum, that is, $N_h = N_I$ in all strata. The values of the variables of interest y and z are generated from different gamma distributions. Wu (2003) claims that heavy tail distributions such as the log-normal and the gamma distribution with large scale parameters should not be used to generate sampling observations. For this reason, we use the gamma distribution with small shape and scale parameters.

In each stratum, a simple random sample of equal size n is selected, then the two proposed estimators and the indirect sampling estimator are computed in order to estimate t_z. The process was repeated $G = 1000$ times with $N_I = 20, 50, 100, 400$ clusters, and $H = 5, 5, 10, 50$ for each of these values of N_I. The simulation was programmed in the statistical software R (R Development Core Team 2009) and the source codes are available from the author upon request. In the simulation, the performance of an estimator $\widehat{t}$ of the parameter t was tracked by the Percent Relative Bias (RB), defined by

$$RB(\widehat{t}) = 100\% G^{-1} \sum_{g=1}^{G} \frac{\widehat{t}_g - t}{t}$$

and the Relative Efficiency (RE), that corresponds to the ratio of the Mean Square Error (MSE) of the estimator of the GWSM approach to the Horvitz-Thompson and the Hájek estimators defined as

$$RE(\widehat{t}_{z\pi}) = \frac{MSE(\widehat{t}_z)}{MSE(\widehat{t}_{z\pi})} \qquad \text{and} \qquad RE(\widetilde{t}_z) = \frac{MSE(\widehat{t}_z)}{MSE(\widetilde{t}_z)}$$

respectively. Note that $\widehat{t}_g$ is computed in the g-th simulated sample and the Mean Square Error is given by

$$MSE(\widehat{t}) = G^{-1} \sum_{g=1}^{G} (\widehat{t}_g - t)^2$$

The estimators are considered under a wide range of specifications. The simulation results correspond to the ratio of MSE, since the ratio of bias is in all cases negligible indicating that no estimator takes advantage over others in terms of the RB.

Table 3, reports the simulated ratio of MSE for the proposed estimators with the indirect sampling estimator for $N_I = 20$, $H = 5$ and $n = 1, 5, 10, 15$. It can be seen that the Hájek estimator is always more efficient, even when the sample size is $n = 1$. The gain in efficiency increases with increasing sample size. The Horvitz-Thompson estimator has a quite poor performance.

Table 3: MSE ratio of the indirect sampling estimator to HT and Hájek estimators for $H = 5$ strata and $N_I = 20$ clusters.

Sample size per stratum	HT	Hájek
n=1	0,08	1,06
n=5	0,03	1,84
n=10	0,05	5,50
n=15	0,52	73,75

Table 4: MSE ratio of the indirect sampling estimator to HT and Hájek estimators for $H = 5$ strata and $N_I = 50$ clusters.

Sample size per stratum	HT	Hájek
n=1	0,12	1,02
n=5	0,03	1,29
n=10	0,02	1,57
n=20	0,02	3,24
n=40	1,06	175,83

Table 5: MSE ratio of the indirect sampling estimator to HT and Hájek estimators for $H = 10$ strata and $N_I = 100$ clusters.

Sample size per stratum	HT	Hájek
n=1	0,09	1,03
n=10	0,02	1,83
n=20	0,02	3,64
n=50	0,44	101,47

Table 6: MSE ratio of the indirect sampling estimator to HT and Hájek estimators for $H = 50$ strata and $N_I = 40$ clusters.

Sample size per stratum	HT	Hájek
n=1	0,02	1,98
n=5	0,77	110,25
n=10	Inf	Inf
n=20	Inf	Inf

Table 7: MSE ratio of the stratified estimator to indirect sampling (IND), HT and Hájek estimators for $H = 5$ strata and $N_I = 20$ clusters.

Sample size per stratum	IND	HT	Hájek
n=1	4,84	3.45	5.39
n=5	4,92	2.53	9.42
n=10	4,34	4.94	27.08
n=15	5,37	40.88	342.90

In the simulation reported in Table 4, we increased the number of clusters to $N_I = 50$, and the sample size to $n = 40$. We see that the Hájek estimator maintains its advantage over the indirect sampling estimator, and it is particularly large when $n = 40$. On the other hand, the Horvitz-Thompson still performs poorly, although when n is close to N_I it is slightly better. The results reported in the Table 5 with $N_I = 100$ and $H = 10$, are similar to those reported in Table 3.

In Table 6, we set $N_I = 40$ and $H = 50$, that is, there are more strata than first level population clusters. We see that the advantage of the Hájek estimator increases substantially even when $n = 5$. The symbol Inf indicates that the MSE of the Horvitz-Thompson and the Hájek estimator are both close to zero in comparison with the MSE of the indirect sampling estimator; that is, the ratio of MSE is huge.)

In order to visualize the average performance of these three approaches, Figure 1, presents the histogram of the Horvitz-Thompson, Hájek and indirect sampling estimators with $N_I = 20$, $H = 5$, $n = 5$. The vertical dotted line indicates the value of the parameter of interest. We observe that the three estimators are unbiased and the estimations obtained with the Hájek estimator are highly concentrated around the population total, while the Horvitz-Thompson estimator has a larger variance.

An interesting, but less practical, situation arises when the parameter of interest in the second level coincides with the parameter of interest in the first level. That is, if $z_i = \sum_{k \in U_i} y_k$, the variable of interest in the cluster U_i corresponds to the total of the variable y in the cluster U_i. In this case, both population totals are the same ($t_y = t_z$) and they can be estimated by using the four mentioned estimators, namely: the stratified estimator given in (1), the Horvitz-Thompson estimator given in (5), the Hájek estimator given in (6) and the indirect sampling estimator given in (8). Notice that in this case, the Horvitz-Thompson, Hájek and indirect estimators use first level information, whereas the stratified estimator uses second level information. Then, it is interesting to evaluate these estimators and compare them. Figure 2 shows the average performance of the four estimators with $N_I = 20$, $H = 5$, $n = 5$. We conclude, once more, that the Hájek estimator is the most efficient and that the estimator of indirect sampling has an acceptable performance, while the stratified and the Horvitz-Thompson estimators have large variances.

Table 7, reports simulation results when comparing the stratified estimator with respect to the remaining three estimators which use the first level information, in terms of relative efficiency. We can see that estimators using first level information are always more efficient than the classical stratified estimator; on the other hand, for each n, the Hájek estimator is the most efficient when increasing the sample size.

The above simulations involve the case that any cluster contains at most one member per stratum, this way the sample includes at most one member in each cluster. However, since our approach may be extended to the general case where a cluster might contain more than one member in some strata, then a more realistic situation arises when we set $a_h > 1$ in some strata. Table 8, reports the simulated

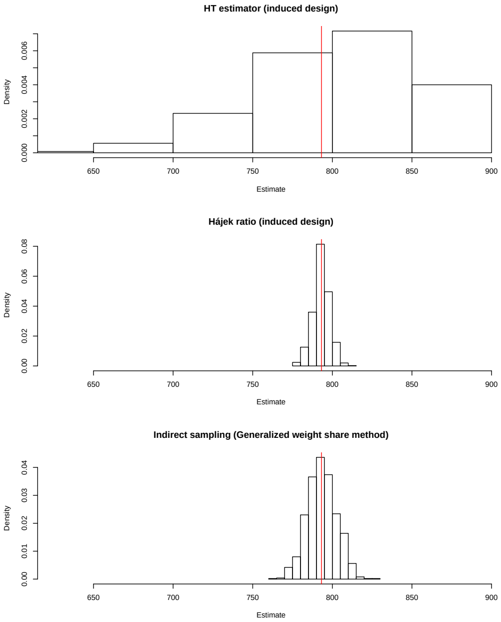

Figure 1: Histogram of estimates in 1000 iterations with $N_I = 20$, $H = 5$, $n = 5$.

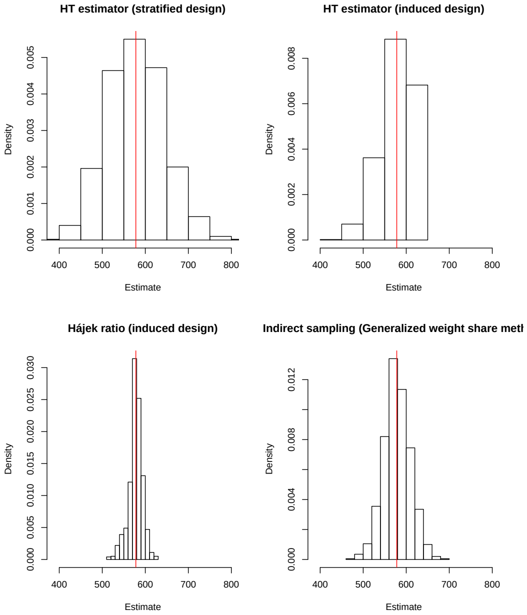

Figure 2: Histogram of estimates in 1000 iterations with $N_I = 20$, $H = 5$, $n = 5$.

MSE ratio for the proposed estimators with the indirect sampling estimator for $N_I = 20$, $H = 5$, $a_h = 3$ for each $h = 1, \ldots, H$ and each cluster. Finally, the sample size considered per stratum was $n = 1, 5, 10, 15$. It can be seen that the Hájek estimator is always more efficient, even when sample size is $n = 1$; its gain in efficiency increases with the sample size augmenting. Figure 3, shows the average performance of the three estimators with $N_I = 20$, $H = 5$, $n = 5$.

Table 8: MSE ratio of the indirect sampling estimator to HT and Hájek estimators for $H = 5$ strata, $N_I = 20$ clusters and $a_h = 3$.

Sample size per stratum	HT	Hájek
n=1	0,07	1,06
n=5	0,03	1,89
n=10	0,04	4,85
n=15	0,11	17,65

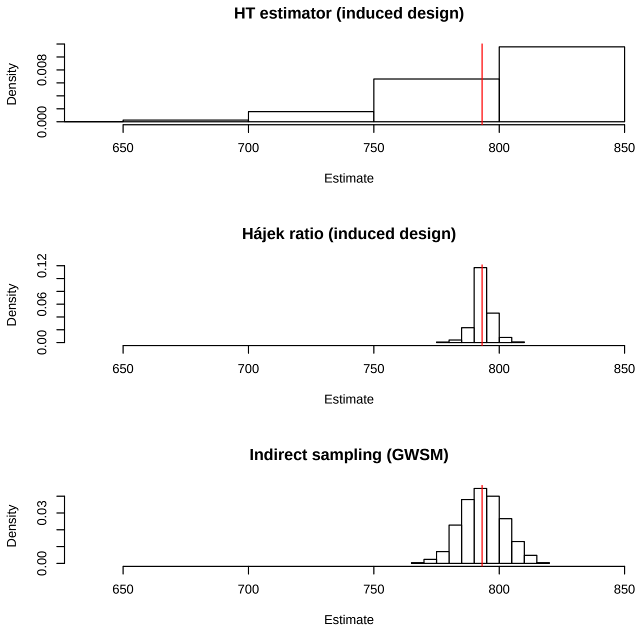

Figure 3: Histogram of estimates in 1000 iterations with $N_I = 20$, $H = 5$, $n = 10$ and $a_h = 3$.

It is worth commenting that the Hajek estimator is asymptotically unbiased. However, for samples of size 20 or more, the bias may be important not to be ignored (Särndal et al. 1992, p. 251). There are some proposals available in the literature to modify either the estimator or the sampling design to reduce the bias of this estimator. For a review of some variations of the Hajek estimator, see Rao (1988). Note that even though the sample size in the stratified second level is small, the induced sample size in the first level is not. This way, it is understandable that the bias for the Hajek estimator is negligible.

5. Discussion and Conclusion

In this paper, we have proposed a design-based approach that yields the unbiased estimation of the population total in the first level based on a stratified sampling design in the second level. With this in mind, the proposed approach is multipurpose in the sense that, for the same survey, different parameters can be estimated in different levels of the population. An important feature of this method is its suitability in the estimation of parameters in the first level where there is no sampling frame available. The empirical study shows that by using the same information, our proposal outperforms the indirect sampling approach because our proposal always has a smaller mean squared error.

The reduction of variability in our proposal may be explained because different second level samples may induce the same first level sample m. In this case, the estimates obtained by applying the GWSM principles will be generally different because the vector of weights $\mathbf{w}$, that depends on the inclusion probabilities of the selected elements in s, differs from sample to sample in the second level. Then we will have different estimates for the same induced sample m. This feature is not present if we follow the approach proposed in this paper, since $\widehat{t}_{z,\pi}$ and $\widetilde{t}_z$ remain constant for different second level samples that induce the same first level sample m. However, $\widehat{t}_{z,\pi}$ does not perform as well as $\widetilde{t}_z$ because, in general, the Horvitz-Thompson approach does not work well under random size sample designs, which is the nature of the sampling design $p(m)$.

This research is still open, further work could be focused in the development of a general methodology conducive to joint estimation in more than two levels when the sampling frame is only available in the last level of the hierarchical population. Besides, the proposed approach could be easily extended in some situations where there is a suitable auxiliary variable (continuous or discrete) that helps to improve the efficiency of the resulting estimators, just as in the functional form of the GWSM with the calibration approach (Lavallée 2007, ch. 7).

Acknowledgements

We thank God for guiding our research. We are grateful to the two anonymous referees for their valuable suggestions and to the Editor in Chief for his advice during the publication process and his comments on the asymptotic unbiased property of the Hajek estimator. Our posthumous gratitude to Leonardo Bautista who motivated this research some years ago. This research was supported by a grant of the Unidad de Investigación from Universidad Santo Tomás.

[Recibido: noviembre de 2009 — Aceptado: mayo de 2011]

References

Deville, J. C. & Lavallée, P. (2006), 'Indirect sampling: the foundation of the generalized weight shared method', *Survey Methodology* **32**(2), 165–176.

Gelman, A. & Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.

Goldstein, H. (1991), 'Multilevel modelling of survey data', *Journal of the Royal Statistical Society: Series D (The Statistician)* **40**(2), 235–244.

Goldstein, H. (2002), *Multilevel Statistical Models*, third edn, Wiley.

Gutiérrez, H. A. (2009), *Estrategias de Muestreo. Diseño de Encuestas y Estimación de Parámetros*, Universidad Santo Tomás.

Holmberg, A. (2002), 'A multiparameter perspective on the choice of sampling design in surveys', *Statistics in Transition* **5**, 969–994.

Lavallée, P. (2007), *Indirect Sampling.*, Springer.

Lehtonen, R. & Veijanen, A. (1999), Multilevel-model assisted generalized regression estimators for domain estimation, *in* 'Proceedings of the 52nd ISI Session'.

R Development Core Team (2009), *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

Rabe-Hesketh, S. & Skrondal, A. (2006), 'Multilevel modelling of complex survey data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**(4), 805–827.

Rao, P. S. R. S. (1988), Ratio and regression estimators, *in* P. R. Krishnaiah & C. Rao, eds, 'Handbook of Statistics', Vol. 6, North-Holland, pp. 449–468.

Särndal, C. E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.

Skinner, C. J., Holt, D. & Smith, T. M. F. (1989), *Analysis of Complex Surveys*, Chichester: Wiley.

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

Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 421 a 432

Testing Homogeneity for Poisson Processes

Prueba de homogeneidad para procesos de Poisson

Raúl Fierro[1,2,a], Alejandra Tapia[3,b]

[1]Instituto de Matemática, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

[2]Centro de Investigación y Modelamiento de Fenómenos Aleatorios-Valparaíso, Universidad de Valparaíso, Valparaíso, Chile

[3]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brasil

Abstract

We developed an asymptotically optimal hypothesis test concerning the homogeneity of a Poisson process over various subintervals. Under the null hypothesis, maximum likelihood estimators for the values of the intensity function on the subintervals are determined, and are used in the test for homogeneity.

***Key words*:** Poisson process, hypothesis testing, local alternatives, asymptotic distribution, asymptotically optimal, likelihood ratio test.

Resumen

Una prueba de hipótesis asintótica para verificar homogeneidad de un proceso de Poisson sobre ciertos subintervalos es desarrollada. Bajo la hipótesis nula, estimadores máximo verosímiles para los valores de la función intensidad sobre los subintervalos mencionados son determinados y usados en el test de homogeneidad.

***Palabras clave*:** proceso de Poisson, prueba de hipótesis, alternativas locales, distribución asintótica, asintóticamente óptimo, prueba de razón de verosimilitud.

1. Introduction

Poisson processes have been used to model random phenomena in areas such as communications, hydrology, meteorology, insurance, reliability, and seismology,

[a]Professor. E-mail: rfierro@ucv.cl
[b]Doctoral student. E-mail: alejandreandrea@gmail.com

among others. These processes are often appropriate for modeling a series of events over time.

Poisson processes are governed by an intensity function $\lambda(t)$, which determines the instantaneous rate of event occurrence at time t. Equivalently, a Poisson process is also governed by the cumulative intensity function $\Lambda(t) = \int_0^t \lambda(\tau)\, d\tau$. When the intensity function is a constant, the Poisson process is known as a homogeneous Poisson process. When the intensity function varies with time, the Poisson process is known as a nonhomogeneous Poisson process (NHPP). A special case of an NHPP which arises in this paper has an intensity function that is piecewise constant over various time subintervals. The main aim of this paper is to develop a hypothesis test to determine whether an observed point process is drawn from a homogeneous Poisson process or a nonhomogeneous Poisson process with a piecewise constant intensity function.

A number of authors have carried out statistical analysis on the intensity of an NHPP. For instance, Leemis (1991), Leemis (2004), Kuhl, Wilson & Johnson (1997), Arkin & Leemis (2000), Kuhl & Wilson (2000), Henderson (2003), and others have considered the nonparametric estimation of the cumulative intensity function for a NHPP, and some of these authors have devoted their attention to modeling the periodic behavior of the process.

By following ideas from Fierro (2008), and considering, as in Leemis (2004), a finite time horizon that has been partitioned into subintervals, we state a result for testing whether a Poisson process is homogeneous or not over certain time intervals. Although a nonhomogeneous Poisson process is oftentimes a more accurate model of a phenomenon occurring in a non-stationary fashion, from a statistical point of view, the modeling based on a homogeneous Poisson process is simpler due to the fact that its intensity function depends only upon a single real parameter. Even though the process could be nonhomogeneous, it is important to investigate whether the process is homogeneous at certain time intervals.

For this reason, the main aim of this paper is to develop an asymptotic likelihood test for homogeneity. It is proved that this test is asymptotically optimal. As in Neyman (1949), we study the asymptotic behavior of the log likelihood of the test, but additionally, we consider the noncentral scenario to obtain an approximation to the power of the test. A sequence of local alternative hypotheses are stated, similar to those considered in some tests, which can be found in Serfling (1980), Karr (1991) and Lehmann (1999). Under the null hypothesis, the intensity function of the process is piecewise constant and, in order to obtain sufficient information to estimate these constants, observations from each of these intervals should be considered. A maximum likelihood estimator should take this information into account, for example, when estimating the cumulative intensity function. The methods introduced here are parametric because the inference on the intensity function of a NHPP involves a finite number of parameters. This technique has been argued against by some authors because it requires the introduction of parameters by the modeler (Leemis 1991, Arkin & Leemis 2000); this technique, however, has been used by Henderson (2003) and Leemis (2004) and we believe it is appropriate in many settings. Even though, in this work, the intensity of the nonhomogeneous Poisson process is not sequentially estimated, it is worth mentioning there are other estimation methods. One of them is the Shiryayev-Roberts test which was introduced by Shiryaev (1963) and Roberts (1966). This procedure is concerned with the sequential detection of changes in distributions occurring at unknown points in time.

The article is organized as follows. In Section 2 we introduce the null hypothesis and derive maximum likelihood estimators under the null hypothesis and without restrictions on the parameters. In Section 3, the asymptotic normality of the estimators is stated with the main results introduced in Section 4. In Section 5, an example is presented. Finally a method for simulations of NHPP variables under the null hypothesis, is proposed in Section 5.

2. Preliminaries

Let T be a fixed strictly positive real number and let us partition the interval $[0,T]$ into m subintervals $[t_0,t_1],(t_1,t_2],\dots,(t_{m-1},t_m]$, where $t_0=0$ and $t_m=T$. The subintervals do not necessarily have equal widths. Let us denote by $\mathcal{C}$ the class of all functions λ which are piecewise constant on each subinterval defined above. From now on, the constant value of λ on $(t_{i-1},t_i]$ will be denoted by λ_i. Consequently,

$$\lambda(t)=\lambda_1\mathrm{I}_{[t_0,t_1]}(t)+\sum_{i=2}^{m}\lambda_i\mathrm{I}_{(t_{i-1},t_i]}(t)$$

where I_C stands for the indicator function on a set C.

This work refers to the hypothesis test that the intensity λ is constant in certain groups of the above subintervals. To do this, we need to partition the set $J=\{1,\dots,m\}$ into r subsets $J(1),\dots,J(r)$, (r groups), that is, $J=J(1)\cup\cdots\cup J(r)$, and for $u\neq v$, $J(u)\cap J(v)=\emptyset$. Let us denote by $m(u)$ the cardinality of $J(u)$. Hence $m(1)+\cdots+m(r)=m$. With these notations, we are interested in finding out whether or not $\lambda(t)$ is constant on the sets $\bigcup_{i\in J(u)}(t_{i-1},t_i]$, ($u\in\{1,\dots,r\}$). Consequently, the null hypothesis should be stated in mathematical terms as follows:

$$\mathrm{H}_0:\forall u\in\{1,\dots,r\},\forall i,j\in J(u),\lambda_i=\lambda_j \tag{1}$$

This hypothesis can be stated in the following simpler equivalent form:

$$\mathrm{H}_0:\forall u\in\{1,\dots,r\},\forall i\in J(u),\lambda_i=\lambda^u$$

where $\lambda^u=\sum_{i\in J(u)}\lambda_i/m(u)$.

Considering $r=1$, H_0 is the hypothesis corresponding to λ is the intensity of an homogeneous Poisson process.

Assume there are $N_1,\dots,N_k$ independent realizations of a nonhomogeneous Poisson process with intensity $\lambda\in\mathcal{C}$ and as before, λ_i denotes the constant value of λ on $(t_{i-1},t_i]$. Put $N^k=N_1+\cdots+N_k$. An estimation of λ_i can be obtained by counting the jumps of N_k into the interval $(t_{i-1},t_i]$.

From Theorem 2.31, in Karr (1991), a likelihood function for $\lambda_1, \dots, \lambda_m$ is given on $[0, T]$ by

$$\mathcal{L}(\lambda_i; 1 \leq i \leq m) = \exp\left[\int_0^T \log(\lambda(t))\ \mathrm{d}N^k(t) - k\int_0^T \lambda(t)\ \mathrm{d}t\right]$$

Hence,

$$\mathcal{L}(\lambda_i; 1 \leq i \leq m) = \exp\left[\sum_{i=1}^{m}\{\log(\lambda_i)\triangle N_i^k - k\lambda_i\triangle t_i\}\right]$$

where $\triangle N_i^k = N^k(t_i) - N^k(t_{i-1})$ and $\triangle t_i = t_i - t_{i-1}$.

Under H_0, this likelihood function on $[0, T]$ becomes

$$\mathcal{L}_0(\lambda^u; 1 \leq u \leq r) = \exp\left[\sum_{u=1}^{r}\log(\lambda^u)\sum_{i\in J(u)}\triangle N_i^k - k\lambda^u\sum_{i\in J(u)}\triangle t_i\right]$$

It is easy to see that the maxima of $\mathcal{L}_0$ and $\mathcal{L}$ are attained at $\lambda^u = \widehat{\lambda^u}, 1 \leq u \leq r$, and $\lambda_i = \widehat{\lambda}_i, 1 \leq i \leq m$, respectively, where

$$\widehat{\lambda^u} = \frac{\sum_{i\in J(u)}\triangle N_i^k}{k\sum_{i\in J(u)}\triangle t_i} \qquad \text{and} \qquad \widehat{\lambda}_i = \frac{\triangle N_i^k}{k\triangle t_i}$$

For the sake of simplicity, the reference to k in these maximum likelihood estimators has been omitted.

Notice that, under H_0, $\widehat{\lambda}_i$ is not sufficient for λ_i and thus there exists information from the data which is not contained in the statistic $\widehat{\lambda}_i$. This lack of information is contained in $T = \sum_{i\in J(u)}\triangle N_i^k$, for instance. Consequently, it is prominent the convenience of using, under H_0, $\widehat{\lambda^u}$ instead of $\widehat{\lambda}_i$, in any estimation of a function of λ_i, $(i \in J(u))$. This fact is relevant in Section 5, where an estimation of the cumulative intensity function of the process is considered in variate generation by inversion and by thinning for a NHPP from event count data.

3. Asymptotic Normality of Estimators

For making inference about the parameters λ_i, $(i = 1, \dots, r)$, for instance, in order to obtain asymptotic confidence intervals for these parameters, we need the corresponding estimators to be consistent and asymptotically normal. This fact is stated in Proposition 1 below. Moreover, Corollary 1, provides us the asymptotical distribution for the parameters under the null hypothesis.

Proposition 1. *For each $i = 1, \dots, m$, $\widehat{\lambda}_i$ is consistent and asymptotically normal $\mathcal{N}(0, \lambda_i)$, which means that the following two conditions hold:*

(C1) *For each $i = 1, \dots, m$, $\widehat{\lambda_i}$ converges to λ_i, with probability 1, as $k \to \infty$.*

(C2) $\sqrt{k}(\widehat{\lambda_1} - \lambda_1, \dots, \widehat{\lambda_m} - \lambda_m)$ *converges in distribution to an m-variate normal random vector having mean zero and covariance matrix $\mathbf{\Sigma}$ given by*

$$\mathbf{\Sigma} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m \end{pmatrix}$$

Proof . Conditions (C1) and (C2) directly follow from Kolmogorov's Law of Large Numbers, the independent increments property of Poisson processes and the classical Central Limit Theorem. □

Corollary 1. *Under* H$_0$*, for each $u = 1, \dots, r$, $\widehat{\lambda^u}$ converges to λ^u as $k \uparrow \infty$ and $\sqrt{k}(\widehat{\lambda^1} - \lambda^1, \dots, \widehat{\lambda^r} - \lambda^r)$ converges in distribution to an r-variate normal random vector having mean zero and covariance matrix $\mathbf{\Sigma}$ given by*

$$\mathbf{\Sigma} = \begin{pmatrix} \lambda^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda^r \end{pmatrix}$$

The above corollary enables us to obtain the usual confidence-interval estimate for λ^u, and hence for λ_i where $i \in J(u)$. Indeed, an asymptotically $100(1-\alpha)\%$ confidence interval for λ^u is

$$\widehat{\lambda^u} - z_{\alpha/2}\sqrt{\widehat{\lambda^u}/k} < \lambda^u < \widehat{\lambda^u} + z_{\alpha/2}\sqrt{\widehat{\lambda^u}/k}$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ percentile of the standard normal distribution.

4. Main Result

The main result of this paper is stated in Theorem 1 below. For testing H$_0$ against H$_0$ fails to be true, we denote by $\mathcal{R}_k$ the likelihood ratio, that is

$$\mathcal{R}_k = \frac{\mathcal{L}_0(\widehat{\lambda^u}; 1 \le u \le r)}{\mathcal{L}(\widehat{\lambda}_i; 1 \le i \le m)}$$

and hence

$$\mathcal{R}_k = \exp\left[\sum_{u=1}^{r} \sum_{i \in J(u)} [\log(\widehat{\lambda^u}/\widehat{\lambda}_i)\triangle N_i^k - k(\widehat{\lambda^u} - \widehat{\lambda}_i)\triangle t_i]\right] \tag{2}$$

Even though $\mathcal{R}_k$ depends on N^k, it is worth noting $\mathcal{R}_k$ does not depend on k, i.e., $\mathcal{R}_k$ depends on k only through N^k.

In order to state the main result, for each $u \in \{1, \dots, r\}$, we consider the following sequence of local alternatives to the null hypothesis:

$$\mathrm{H}^{(k)} : \forall u \in \{1, \dots, r\}, \forall i \in J(u), \lambda_i = \lambda^u + \delta_i/\sqrt{k} \tag{3}$$

where $\delta = (\delta_1, \dots, \delta_m)$ is a fixed vector in $\mathbb{R}^m$ satisfying $\sum_{i \in J(u)} \delta_i = 0$, for each $u \in \{1, \dots, r\}$.

Theorem 1. *Under* $\mathrm{H}^{(k)}$ *as* $k \to \infty$, $-2\log(\mathcal{R}_k)$ *has noncentral asymptotically* χ^2 *distribution with* $m - r$ *degrees of freedom and noncentrality parameter*

$$\begin{aligned}\Phi^2 &= \sum_{u=1}^{r} \frac{1}{\lambda^u} \sum_{i \in J(u)} \triangle t_i \left[\delta_i - \frac{\sum_{j \in J(u)} \delta_j \triangle t_j}{\sum_{j \in J(u)} \triangle t_j} \right]^2 \\ &= \sum_{u=1}^{r} \frac{1}{\lambda^u} \sum_{i \in J(u)} \triangle t_i \left[\delta_i^2 - \left(\frac{\sum_{j \in J(u)} \delta_j \triangle t_j}{\sum_{j \in J(u)} \triangle t_j} \right)^2 \right]\end{aligned}$$

***Proof*.** By taking into account that $\log(x) = (x-1) - (x-1)^2/2 + \mathrm{O}((x-1)^3)$, from (2) it is obtained

$$\begin{aligned}-2\log(\mathcal{R}_k) &= \sum_{u=1}^{r} \sum_{i \in I(u)} \left([(\widehat{\lambda}_i - \widehat{\lambda^u})/\widehat{\lambda}_i]^2 \triangle N_i^k + \mathrm{O}([(\widehat{\lambda}_i - \widehat{\lambda^u})/\widehat{\lambda}_i]^3) \right) \\ &= \sum_{i=1}^{m} \left[(U_i^k)^2 + \mathrm{O}_{\mathrm{P}}((U_i^k)^3/\sqrt{\triangle N_i^k}) \right]\end{aligned}$$

where $U_i^k = (\widehat{\lambda}_i - \widehat{\lambda^u})\sqrt{k\triangle t_i/\widehat{\lambda}_i}$ whenever $i \in J(u)$, and in general, $A_n = \mathrm{O}_{\mathrm{P}}(B_n)$ means that given any $\eta > 0$, there is a constant $M = M(\eta)$ and a positive integer $n_0 = n_0(\eta)$ such that $\Pr\{|A_n| \le M|B_n|\} \ge 1 - \eta$ for every $n > n_0$.

For each $i = 1, \dots, m$, let $\triangle M_i^k = (\triangle N_i^k - k\lambda_i \triangle t_i)/\sqrt{k}$. Since $\triangle M_1^k, \dots, \triangle M_m^k$ are independent, by the classical Central Limit Theorem, $\{(\triangle M_1^k, \dots, \triangle M_m^k)\}_{k \in \mathbb{N}}$ converges in distribution to a normal random vector having mean zero and covariance matrix $\mathbf{\Sigma}$ given by the diagonal matrix $\mathbf{\Sigma} = \mathrm{diag}(\lambda_1 \triangle t_1, \dots, \lambda_m \triangle t_m)$. Under $\mathrm{H}^{(k)}$, for each $u = 1, \dots, r$ and each $i \in J(u)$, $\triangle N_i^k = \lambda^u k \triangle t_i + \delta_i \sqrt{k} \triangle t_i + \sqrt{k} \triangle M_i^k$. Hence,

$$\widehat{\lambda}_i = \lambda^u + \frac{\delta_i}{\sqrt{k}} + \frac{\triangle M_i^k}{\sqrt{k}\triangle t_i} \tag{4}$$

and

$$\widehat{\lambda^u} = \lambda^u + \frac{\sum_{j \in J(u)} \delta_j \triangle t_j}{\sqrt{k}\sum_{j \in J(u)} \triangle t_j} + \frac{\sum_{j \in J(u)} \triangle M_j^k}{\sqrt{k}\sum_{j \in J(u)} \triangle t_j} \tag{5}$$

From (4) and (5), for each $i \in J(u)$ one obtains

$$\begin{aligned} U_i^k &= \sqrt{\frac{\triangle t_i}{\widehat{\lambda}_i}}\left(\delta_i + \frac{\triangle M_i^k}{\triangle t_i} - \frac{\sum_{j\in J(u)} \triangle M_j^k}{\sum_{j\in J(u)} \triangle t_j} - \frac{\sum_{j\in J(u)} \delta_j \triangle t_j}{\sum_{j\in J(u)} \triangle t_j}\right) \\ &= V_i^k - \frac{\sqrt{\triangle t_i}\sum_{j\in J(u)} \sqrt{\triangle t_j} V_j^k \sqrt{\widehat{\lambda}_j/\widehat{\lambda}_i}}{\sum_{j\in J(u)} \triangle t_j} \end{aligned}$$

where $V_i^k = \triangle M_i^k/\sqrt{\widehat{\lambda}_i \triangle t_i} + \delta_i\sqrt{\triangle t_i/\widehat{\lambda}_i}$.

Under $\mathrm{H}^{(k)}$, for each $i \in J(u)$, $\lambda_i = \lambda^u + \delta_i/\sqrt{k}$ and by Proposition 1 in Section 3, for each $i, j \in J(u)$, $\sqrt{\widehat{\lambda}_j/\widehat{\lambda}_i} \to 1$, with probability 1, as $k \to \infty$. Consequently, from a slight modification of Proposition 1 in Section 3 and Slutzky's theorem, $\{(U_1^k, \dots, U_m^k)\}_{k\in\mathbb{N}}$ converges in distribution to $\mathbf{U} = (U_1, \dots, U_m)$, where for each $i \in J(u)$, $(u = 1, \dots, r)$,

$$U_i = V_i - \frac{\sqrt{\triangle t_i}\sum_{j\in J(u)} \sqrt{\triangle t_j} V_j}{\sum_{j\in J(u)} \triangle t_j}$$

and $\mathbf{V} = (V_1, \dots, V_m)^{\mathrm{t}}$ is a vector of m independent normal random variables with variance one, and such that for each $u \in \{1, \dots, r\}$ and each $j \in J(u)$, V_j has mean $\delta_j\sqrt{\triangle t_j/\lambda^u}$. Hence, $\{-2\log(\mathcal{R}_k))\}_{k\in\mathbb{N}}$ converges in distribution to $\|\mathbf{U}\|^2 = \sum_{i=1}^m U_i^2$, where $\|\cdot\|$ stands for the Euclidean norm in $\mathbb{R}^m$.

Let

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}(1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{P}(r) \end{pmatrix}$$

where for each $u = 1, \dots, r$, $\mathbf{P}(u) = (p_{ij}(u); i, j \in J(u))$ is the matrix defined by $p_{ij}(u) = \sqrt{\triangle t_i \triangle t_j}/\sum_{j\in J(u)} \triangle t_j$. Hence, $\mathbf{U} = (\mathbf{I} - \mathbf{P})\mathbf{V}$, and since for each $u = 1, \dots, r$, $\mathbf{P}(u)$ is an idempotent matrix having rank 1, the matrix $\mathbf{P}$ is idempotent as well and has rank r. Consequently, $\mathbf{I} - \mathbf{P}$ is idempotent and has rank $m - r$. It follows from Theorem 3.5.1 in Serfling (1980) that $\|\mathbf{U}\|^2$ has χ^2 distribution with $m - r$ degrees of freedom and non-centrality parameter $\mu(\mathbf{I} - \mathbf{P})\mu^{\mathrm{t}}$, where

$$\mu = (\delta_1\sqrt{\triangle t_1/\lambda^{u(1)}}, \dots, \delta_m\sqrt{\triangle t_m/\lambda^{u(m)}})$$

and for each $i \in \{1, \dots, m\}$, $u(i)$ is the unique integer in $\{1, \dots, r\}$ such that $i \in J(u(i))$. Since $\mu^{\mathrm{t}}(\mathbf{I} - \mathbf{P})\mu = \|(\mathbf{I} - \mathbf{P})\mu\|^2 = \|\mu\|^2 - \|\mathbf{P}\mu\|^2$, the proof is complete. □

The corollary below is useful to test the hypothesis whether a Poisson process is homogeneous or not.

Corollary 2. *Let $\widetilde{\lambda} = \sum_{j=1}^{m} \lambda_j/m$, $(\delta_1, \dots, \delta_m) \in \mathbb{R}^m$ such that $\sum_{j=1}^{m} \delta_i = 0$ and $\mathrm{H}^{(k)}$ be the statistical hypothesis defined as*

$$\mathrm{H}^{(k)} : \forall i \in \{1, \dots, m\}, \lambda_i = \widetilde{\lambda} + \delta_i/\sqrt{k}$$

Then, under $\mathrm{H}^{(k)}$, $-2\log(\mathcal{R}_k)$ has noncentral asymptotically χ^2 distribution with $m-1$ degrees of freedom and noncentrality parameter

$$\Phi^2 = \frac{1}{\widetilde{\lambda}} \sum_{i=1}^{m} \triangle t_i \left[\delta_i - \frac{\sum_{j=1}^{m} \delta_j \triangle t_j}{\sum_{j=1}^{m} \triangle t_j} \right]^2$$

***Note* 1.** A natural application of the foregoing theorem is to calculate the approximate power of the test relative to

$$\mathrm{H}_0 : \forall u \in \{1, \dots, r\}, \forall i \in J(u), \lambda_i = \lambda^u$$

against the simple alternative

$$\mathrm{H}_1 : \forall u \in \{1, \dots, r\}, \forall i \in J(u), \lambda_i = \lambda_i^*$$

Suppose that the critical region is $\{-2\log(\mathcal{R}_k) > t_0\}$, where t_0 has been calculated for a level of significance α based upon the null hypothesis asymptotic χ^2-distribution of $-2\log(\mathcal{R}_k)$.

We interpret δ_i in $\mathrm{H}^{(k)}$ as $\sqrt{k}(\lambda_i^* - \lambda^u)$ and approximate the power of the test by means of the probability of $\{\chi^2 > t_0\}$, where χ^2 is a random variable having χ^2-distribution with $m-r$ degrees of freedom and noncentrality parameter

$$\Phi^2 = k \sum_{u=1}^{r} \frac{1}{\lambda^u} \sum_{i \in J(u)} \triangle t_i \left[\lambda_i^* - \frac{\sum_{j \in J(u)} \triangle t_j \lambda_j^*}{\sum_{j \in J(u)} \triangle t_j} \right]^2$$

***Note* 2.** By the standard Central Limit Theorem, for $m-r$ and k large enough, $-2\log(\mathcal{R}_k)$ has approximate normal distribution with mean $m-r$ and variance $2(m-r)$.

Based on Theorem 1, an asymptotically maximum likelihood test, for testing H_0, according to (1), against local alternatives, can be stated. An important property of a test is its power optimality. The following proposition allows to conclude the above-mentioned test is asymptotically uniformly most powerful.

Proposition 2. *Let $\mathcal{B}(\mathbb{R})$ be the Borel σ-algebra of subsets of $\mathbb{R}$ and for each $\nu \geq 0$, P_ν be the probability distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ corresponding to the χ^2-distribution with g degrees of freedom and noncentrality parameter ν. For testing $H : \nu = 0$ against $K : \nu > 0$, the test defined by the critical region $[t_\alpha, \infty)$, where $P_0([t_\alpha, \infty)) = \alpha$, is uniformly most powerful.*

***Proof*.** The probability density function corresponding to the χ^2-distribution with g degrees of freedom and noncentrality parameter ν is given by

$$f(x, g, \nu) = \frac{x^{g/2-1}\, \mathrm{e}^{-(x+\nu)/2}}{2^{g/2}\Gamma(g/2)} \left(1 + \sum_{j=1}^{\infty} \frac{(\nu x/4)^j \Gamma(g/2)}{j!\Gamma(j+g/2)} \right) \mathrm{I}_{]0,\infty[}(x)$$

Hence, for each $x \geq 0$, we have $f(x, g, \nu) = f(x, g, 0)G(x, g, \nu)$, where

$$G(x, g, \nu) = \mathrm{e}^{-\nu/2}\left(1 + \sum_{j=1}^{\infty} \frac{(\nu x/4)^j \Gamma(g/2)}{j!\Gamma(j + g/2)}\right)$$

Let $A = [t_\alpha, \infty[$ and B be a Borelian subset of $\mathbb{R}$ such that $P_0(B) = \alpha$. We need to prove, for each $\nu \geq 0$, $P_\nu(A) \geq P_\nu(B)$.

We have

$$P_\nu(A) - P_\nu(B) = \int_{A \cap B^{\,c}} f(x, g, 0)G(x, g, \nu)\ \mathrm{d}x - \int_{A^{\,c} \cap B} f(x, g, 0)G(x, g, \nu)\ \mathrm{d}x$$

and since G is increasing at the first variable, we derive

$$\begin{aligned} P_\nu(A) - P_\nu(B) &\geq G(t_\alpha, g, \nu)\left(\int_{A \cap B^{\,c}} f(x, g, 0)\ \mathrm{d}x - \int_{A^{\,c} \cap B} f(x, g, 0)\ \mathrm{d}x\right) \\ &= G(t_\alpha, g, \nu)(P_0(A) - P_0(B)) \\ &= 0 \end{aligned}$$

Therefore, the proof is complete. □

Corollary 3. *Let* H_0 *and* $\mathrm{H}^{(k)}$ *be the statistical hypotheses defined by (1) and (3), respectively. For testing* H_0 *against* $\mathrm{H}^{(k)}$ *with a significance level* α, $(0 < \alpha < 1)$, *the test defined by the critical region* $\{-2\log(\mathcal{R}_k) > t_\alpha\}$ *is asymptotically uniformly most powerful, where* $t_\alpha > 0$ *is determined by*

$$\lim_{k\to\infty} \Pr(-2\log(\mathcal{R}_k) > t_\alpha) = \alpha.$$

5. An Example

Let us suppose in a call center there are k employees in charge of state connections. The call number is recorded from 10.00 am to 1.00 pm every day during a seven-day period. It is possible to assume the number of phone connections made for each server follows a Poisson process and that these k Poisson processes are independent. For this purpose, it is assume the end of a working day agrees with the beginning of the next day. Even though people usually go for a walk on Saturday and Sunday, it is suspected that on weekends this number decreases due to people are not working. To find out this circumstance, the null hypothesis is defined as "H_{01}: the call rates are every day the same". From Corollary 4.1, in this case, $-2\log(\mathcal{R}_k)$ has asymptotically χ^2-distribution with six degrees of freedom. Another test could be stated by defining the null hypothesis as "H_{02}: the call rates from Monday to Friday are the same, and the Saturday call rate equals the Sunday one". In this last case, there are two groups $J(1)$ and $J(2)$ and consequently, $-2\log(\mathcal{R}_k)$ has asymptotically χ^2-distribution with five degrees of freedom.

In order to compare both tests, we simulate $k = 10$ copies of Poisson process and test H_{01} and H_{02} with the same data set. The call rate corresponding to the working days was assumed to be 50 calls/hour, and the call rate on Saturday and Sunday was assumed to be 45 calls/hour. Both hypothesis tests were performed 10^5 times with a level of significance $\alpha = 0.05$, and as result of this simulation, H_{01} and H_{02} were rejected 96.6% and 4.95%, respectively, which illustrates the test and gives an insight of its power.

6. Variate Generation

The cumulative intensity function for a NHPP is often estimated for simulating the NHPP. A number of methods for carrying out this simulation are described in Lewis & Shedler (1979), where the simulation method for a NHPP by thinning is stated. In this section, our purpose is not to give detailed variate generation algorithms, but to give an estimation of the cumulative intensity function based on $\widehat{\lambda^u}$ $(u = 1, \dots, r)$, which are, under H_0, estimators containing sufficient information for the parameters $\lambda_1, \dots, \lambda_m$. To this end, an estimator for the cumulative intensity function is defined and the basis of variate generation by inversion is recalled.

For each $t \geq 0$, let $i(t)$ denote the unique $i \in \{1, \dots, m\}$ satisfying $t \in Q_{i(t)}$, where $Q_1 = [t_0, t_1]$ and $Q_i = (t_{i-1}, t_i]$ for $j \in \{2, \dots, m\}$. By writing $U(t) = \{u : \exists i \leq i(t), i \in J(u)\}$ and $V(u,t) = \{j \in J(u) : j < i(t)\}$, the cumulative intensity function $\Lambda : [0, T] \to \mathbb{R}$ defined as $\Lambda(t) = \int_0^t \lambda(u)\, \mathrm{d}u$ satisfies

$$\Lambda(t) = \sum_{u \in U(t)} \sum_{j \in V(u,t)} \lambda_j \triangle t_j + \lambda_{i(t)}(t - t_{i(t)-1})$$

Let u_i denote the unique $u \in \{1, \dots, r\}$ such that $i \in J(u)$ and define $u(t) = u_{i(t)}$. Under H_0, we have

$$\Lambda(t) = \sum_{u \in U(t)} \lambda^u \sum_{i \in V(u,t)} \triangle t_i + \lambda^{u(t)}(t - t_{i(t)})$$

Consequently, Λ can be estimated by $\widehat{\Lambda}$, where for $t \geq 0$,

$$\widehat{\Lambda}(t) = \sum_{u \in U(t)} \widehat{\lambda^u} \sum_{i \in V(u,t)} \triangle t_i + \widehat{\lambda^{u(t)}}(t - t_{i(t)})$$

Following Leemis (2004), a realization of a Poisson process for modeling in a discrete-event simulation can be generated, under H_0, by inversion. Let

$$\widehat{\Psi}(u) = \begin{cases} \inf\{t > 0 : \widehat{\Lambda}(t) \geq u\} & \text{if} \quad u \leq \widehat{\Lambda}(T) \\ +\infty & \text{if} \quad u > \widehat{\Lambda}(T) \end{cases}$$

Note that for each $u \geq 0$, $\widehat{\Lambda}(\widehat{\Psi}(u)) = u$, almost everywhere, and consequently, if $S_1, S_2, \dots$ are the points in a homogeneous Poisson process of rate one (which have been chosen independently of $\widehat{\Lambda}$), then $\widehat{\Psi}(S_1), \widehat{\Psi}(S_2), \dots$ are the points in a nonhomogeneous Poisson process with cumulative intensity function $\widehat{\Lambda}$. This fact enables us to generate NHPP event times starting from standard Poisson random variate generation.

According to Henderson (2003), at the beginning of Section 3, pages 379-380, for a general rate function, a faster generation procedure of NHPP event times is obtained by thinning. This method for simulating the NHPP was introduced by Lewis & Shedler (1979) and it is based on an estimator of the rate function λ. Under H_0, a maximum likelihood estimator for λ is given by $\widehat{\lambda}$, which is defined for $t \geq 0$ as

$$\widehat{\lambda}(t) = \sum_{u=1}^{r} \widehat{\lambda^u} \sum_{i \in J(u)} \mathrm{I}_{(t_{i-1}, t_i]}(t).$$

Recall that thinning first generates a candidate event time T^*, and then accepts the event time with probability $\widehat{\lambda}(T^*)/\lambda^*$, where λ^* is an upper bound of λ. The novelty here is that in this case, thinning is based on the estimators $\widehat{\lambda^1}, \dots, \widehat{\lambda^r}$, which, as pointed out before, are sufficient statistics for $\lambda^1, \dots, \lambda^r$.

7. Conclusions and Recommendations

In this paper we carry out a hypothesis test that allows us to find out whether or not a NHPP could be considered homogeneous in certain time intervals. Such an inquiry becomes very important when it is assumed that the rate function is a piecewise constant on subintervals of the time. Indeed, when there exists a great non-homogeneity and an approximated piecewise constant rate function has to be defined, it is necessary to partition the time interval in many subintervals. However, if homogeneity is observed in a large subset (which need not be connected) of the time horizon, a lesser number of subintervals will be necessary and an economy of computational time and/or memory to store the information could be obtained. On the other hand, under the null hypothesis, the estimators of the constant values of the intensity function are expressed in terms of sufficient statistics, which enables us to make use of the whole information provided by the data. This fact is particularly important for generating Poisson variates by inversion or thinning procedure.

8. Acknowledgements

This work was partially supported by Dirección de Investigación e Innovación de la Pontificia Universidad Católica de Valparaíso under projects 124.722/2010 and 037.335/2011. The authors also thank the editor and referees for their helpful suggestions and corrections.

[Recibido: octubre de 2010 — Aceptado: abril de 2011]

References

Arkin, B. L. & Leemis, L. M. (2000), 'Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process from overlapping realizations', *Management Science* **46**(7), 989–998.

Fierro, R. (2008), 'Test of homogeneity for some population models based on counting processes', *Communications in Statistics-Theory and Methods* **37**(1), 46–54.

Henderson, S. G. (2003), 'Estimation for nonhomogeneous Poisson processes from aggregated data', *Operation Research Letters* **31**(5), 375–382.

Karr, A. F. (1991), *Point Processes and their Statistical Inference*, Marcel Dekker, New York.

Kuhl, M. E. & Wilson, J. R. (2000), 'Least square estimation of nonhomogeneous Poisson processes', *Journal of Statistical Computation and Simulation* **67**, 75–108.

Kuhl, M. E., Wilson, J. R. & Johnson, M. A. (1997), 'Estimating and simulating Poisson processes having trends or multiple periodicities', *IIE Transactions* **29**(3), 201–211.

Leemis, L. M. (1991), 'Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process', *Management Science* **37**(7), 886–900.

Leemis, L. M. (2004), 'Nonparametric estimation and variate generation for a nonhomogeneous Poisson process from event count data', *IIE Transactions* **36**, 1155–1160.

Lehmann, E. L. (1999), *Elements of Large-Sample Theory*, Springer-Verlag, New York.

Lewis, P. A. W. & Shedler, G. S. (1979), 'Simulation of nonhomogeneous Poisson process by thinning', *Naval Research Logistics* **26**(3), 403–413.

Neyman, J. (1949), Contribution to the theory of the χ^2 test, *in* 'First Berkeley Symposium on Mathematical Statistics and Probability', pp. 239–273.

Roberts, S. W. (1966), 'A comparison of some control chart procedures', *Technometrics* **8**, 411–430.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley and Sons, New York.

Shiryaev, A. N. (1963), 'On optimum methods in quickest detection problems', *Theory Probability and Its Applications* **8**, 22–46.

Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 433 a 450

Indexes to Measure Dependence between Clinical Diagnostic Tests: A Comparative Study

Indices para medir dependencia entre pruebas para diagnóstico clínico: un estudio comparativo

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

[1]Departamento de Estatística, Instituto de Matemática Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas, Brasil

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

Abstract

In many practical situations, clinical diagnostic procedures include two or more biological traits whose outcomes are expressed on a continuous scale and are then dichotomized using a cut point. As measurements are performed on the same individual there is a likely correlation between the continuous underlying traits that can go unnoticed when the parameter estimation is done with the resulting binary variables. In this paper, we compare the performance of two different indexes developed to evaluate the dependence between diagnostic clinical tests that assume binary structure in the results with the performance of the binary covariance and two copula dependence parameters.

***Key words*:** Copula, Farlie Gumbel Morgenstern distribution, Gumbel distribution.

Resumen

Muchos procedimientos de diagnóstico clínico médico exigen la evaluación de dos o mas rasgos biológicos que se ven alterados ante la presencia de fenómenos de enfermedad o infección, los cuales se expresan en una escala contínua de medición con posterior dicotomización usando de un valor límite o punto de corte. Dado que las mediciones son realizadas en el mismo indivíduo, los resultados probablemente presenten dependencia de algún tipo, lo cual puede ser ignorado en la etapa de análisis de datos dada la presentación binaria de los datos. En este estudio comparamos el comportamiento de dos parámetros de dependencia presentes en funciones de cópula con el de la covarianza binaria y dos índices creados para medir dependencia entre pruebas diagnósticas de respuesta dicótoma.

***Palabras clave*:** cópula, distribución Farlie, Gumbel.

[a]Professor. E-mail: rtovar34@hotmail.com
[b]Professor. E-mail: achacar@fmrp.usp.br

1. Introduction

The study of dependence between two clinical diagnostic tests has been a matter of interest in medical and statistical research. Many studies developed within clinical diagnostic tests framework have studied the conditional dependence between diagnostic tests using the binary covariance as dependence parameter (see for instance: (Thibodeau 1981), (Vacek 1985), (Torrance-Rynard & Walter 1997), (Enoe, Georgiadis & Johnson 2000) and (Dendukuri & Joseph 2001) among many others). Some authors as Georgiadis, Johnson & Gardner (2003), have used reparametrizations of the conditional correlation between binary tests to facilitate the prior specification in the implementation of a Bayesian estimation procedure. Bohning & Patilea (2008), consider one of the indexes used by Georgiadis et al. (2003) and developed another one to study the association between two diagnostic tests in designs where the individuals with negative outcome in both screening tests are not verified by "gold standard" (verification bias conditions).

Many diagnostic procedures include the measures of two or more biological traits directly observable or not, whose outcomes are initially expressed on a continuous scale and operationalized within a dichotomous representation using a cut point. It is possible that, there exists dependence between the two evaluated traits conditional on the true disease state and the same should be studied using indexes developed to study association between continuous variables, but as the data analyses is made with the binary data, the data analyst evaluates the conditional dependence hypothesis using indexes developed to binary variables.

In this paper, we use the indexes developed by Bohning & Patilea (2008) and we compare their performance with the performance of the binary covariance and the performance of the Farlie Gumbel Morgerstern (FGM) and Gumbel copula dependence parameters. The main goal is to evaluate the existing relationship among the five dependence parameters, where three of them (covariance and Böhning's indexes) are built to study dependence between binary variables and the other two (copula parameters) are developed to model dependence between continuous variables.

The paper is organized as follows: in Section 2, we introduce the estimation model formulation for two associated diagnostic tests, in Section 3, we present the comparative study among Böhning and Patilea's indexes, the binary covariance and the copula dependence parameters, in Section 4, we introduce two examples, one of them with simulated data and the other with published data. Finally, in Section 5, we present some conclusions on the results obtained.

2. Statistical Model with Two Dependent Screening Tests

Let us assume that we have a clinical diagnostic procedure that uses two screening tests whose performance we are interested to study and a reference procedure that classifies individuals as diseased and non-diseased without error called "gold standard". Sometimes, the design of the study considers that those individuals with negative outcomes in both screening tests are not verified by the "gold standard" which is known as "verification bias". We assume that, the screening test outcomes are expressed on a continuous scale and they are exposed to a process of dichotomization using a cut point. We also assume that the test outcomes have a continuous dependent structure but the same can not be considered in the data analysis since they are presented in a binary form.

2.1. Modelling Dependence with Binary Structure

Let us denote by p the prevalence of a disease and by D a random variable related to the true disease status, where $D = 1$ denotes a diseased individual and $D = 0$ denotes a non-diseased individual. That is, $p = P(D = 1)$. Also, denote by T_1 and T_2, the two random variables associated to the test results, where $T_j = 1$, denotes a positive result and $T_j = 0$, denotes a negative result, $P(T_j = 1 \mid D = 1) = S_j$ is the sensitivity of the test j and $P(T_j = 0 \mid D = 0) = E_j$ is the specificity of the test j, for $j = 1, 2$. If we assume that, the dependence between tests can be modeled by the covariance (ψ parameter) using a Bernoulli distribution on the test outcome and we also assume the covariance is not necessarily the same in both populations ($\psi_D \neq \psi_{ND}$), we can use the Dendukuri's procedure to obtain the likelihood function contributions, as shown in Table 1.

Table 1: Likelihood contributions of all possible combinations of outcomes of T_1, T_2 and D assuming binary dependence structure (Values in brackets are unknown under verification bias. f_i: number of individuals for each combination of results)

				Contribution to likelihood
D	T_1	T_2	f_i	Binary dependence
1	1	1	a	$p[S_1S_2 + \psi_D]$
1	1	0	b	$p[S_1(1 - S_2) - \psi_D]$
1	0	1	c	$p[(1 - S_1)S_2 - \psi_D]$
1	0	0	$[d]$	$p[(1 - S_1)(1 - S_2) + \psi_D]$
0	1	1	e	$(1 - p)[(1 - E_1)(1 - E_2) + \psi_{ND}]$
0	1	0	f	$(1 - p)[(1 - E_1)E_2 - \psi_{ND}]$
0	0	1	g	$(1 - p)[E_1(1 - E_2) - \psi_{ND}]$
0	0	0	$[h]$	$(1 - p)[E_1E_2 + \psi_{ND}]$

2.2. Modelling Dependence using Copula Functions

Let us assume that the test outcomes are realizations of the random variables V_1 and V_2 measured in a positive continuous scale, that is, $V_1 > 0$ and $V_2 > 0$. 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 measure the degree of the dependence structure between the random variables V_1 and V_2, let us consider the use of copula functions (For details about this topic, (Nelsen 1999) is a good reference). For specified univariate marginal distribution functions $F_1(v_1), \dots, F_m(v_m)$, the function $C(F_1(v_1), \dots, F_m(v_m))$ which is defined using a copula function C, results in a multivariate distribution function with univariate marginal distributions specified as $F_1(v_1), \dots, F_m(v_m)$. Any multivariate distribution function 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, $F(v_1, \dots, v_m) = C(F_1(v_1), \dots, F_m(v_m))$. 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 can be made of each marginal variable in such a way that each transformed marginal variable has an uniform distribution. Once this is done, the dependence structure can be expressed as a multivariate distribution on the obtained uniforms and a copula is precisely a multivariate distribution on marginally uniform random variables. In this way, there are many families of copulas which differ in the detail of the dependence they represent. In the bivariate case, let V_1 and V_2 be two random variables with continuous distribution functions F_1 and F_2. The probability integral transformation is applied separately for the two random variables to define $U = F_1(V_1)$ and $W = F_2(V_2)$ where U and W have uniform $(0, 1)$ distributions, but are usually dependent if V_1 and V_2 are dependent (V_1 and V_2 independent implies that U and W are independent). 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. In this paper, we use two copula functions to study the dependence between two diagnostic tests namely: the Farlie Gumbel Morgerstern (FGM) copula and the Gumbel copula.

The FGM copula is defined by,

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

where $u = F_1(v_1)$, $w = F_2(v_2)$ and $-1 \leq \varphi \leq 1$. As, φ measures the dependence between the two marginals, then, if $\varphi = 0$, we have independent random variables. We assume two dependence parameters φ_D and φ_{ND} with the same value for diseased and non-diseased individuals, respectively. From (1), the cumulative joint distribution and the joint survival distribution functions for the random variables V_1 and V_2 conditional on the diseased status (D subscript and superscript) are given respectively by,

$$\begin{aligned} F(v_1, v_2) = C(F_1(v_1), F_2(v_2)) = F_1(v_1)F_2(v_2) \\ [1 + \varphi(1 - F_1(v_1))(1 - F_2(v_2))] \end{aligned} \tag{2}$$

$$S(v_1, v_2) = P(V_1 > v_1, V_2 > v_2) = 1 - F_1(v_1) - F_2(v_2) + F(v_1, v_2) \tag{3}$$

To obtain the likelihood function contributions within the diseased individuals group we have;

$$P(T_1 = 1, T_2 = 1 \mid D = 1) = P(V_1 > \xi_1, V_2 > \xi_2 \mid D = 1) = S_D(\xi_1, \xi_2),$$

$$P(T_1 = 1 \mid D = 1) = P(V_1 > \xi_1 \mid D = 1) = S_1,$$
$$P(T_2 = 1 \mid D = 1) = P(V_2 > \xi_2 \mid D = 1) = S_2,$$
$$F_1^D(\xi_1) = P(V_1 \leq \xi_1 \mid D = 1) = 1 - S_1,$$

and

$$F_2^D(\xi_2) = P(V_2 \leq \xi_2 \mid D = 1) = 1 - S_2$$

Using (2) we get,

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

and,

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

That is,

$$P(T_1 = 1, T_2 = 1 | D = 1) = S_1 S_2 (1 + \varphi_D (1 - S_1)(1 - S_2)) \tag{4}$$

and

$$P(T_1 = 1, T_2 = 1, D = 1) = p S_1 S_2 (1 + \varphi_D (1 - S_1)(1 - S_2)) \tag{5}$$

Similarly, we get all likelihood contributions with diseased and non-diseased individuals (see Table 2).

The Gumbel copula, developed by Gumbel (1960) is defined as,

$$C(u, w) = u + w - 1 + (1 - u)(1 - w) \exp\{-\phi \ln(1 - u) \ln(1 - w)\} \tag{6}$$

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

$$\begin{aligned} F(v_1, v_2) = F_1(v_1) + F_2(v_2) - 1 + \\ (1 - F_1(v_1))(1 - F_2(v_2)) \exp\{-\phi \ln(1 - F_1(v_1)) \ln(1 - F_2(v_2))\} \end{aligned} \tag{7}$$

The dependence parameter of the Gumbel copula does not model positive linear correlations and when the two variables are independents, we have $\phi = 0$.

Employing the same arguments considered in the FGM copula to find the joint probabilities of all combinations with D, T_1 and T_2 and using (7) we obtain all the contributions for the likelihood function using the Gumbel copula (See Table 2).

Table 2: 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. (Values in brackets are unknown under verification bias. f_i: number of individuals for each combination of results)

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

$Q_1 = \exp(-\phi_D \ln S_1 \ln S_2), \quad Q_2 = \exp(-\phi_{ND} \ln(1-E_1)\ln(1-E_2))$

3. Indexes Developed by Böhning and Patilea

Bohning & Patilea (2008), developed two association indexes to study the case of two dependent diagnostic tests in situations where it is not possible to verify the true disease status in individuals with negative outcome in both screening tests. The authors proposed computation of the indexes using the observed probabilities in the likelihood function. The Böhning and Patilea's indexes θ_i and α_i ($i = D$ denotes diseased individuals and $i = ND$ denotes non-diseased individuals) are defined as:

$$\begin{aligned}\theta_i &= \frac{P(T_1=1 \mid T_2=1, D=i)}{P(T_1=1, D=i)} \\ &= \frac{P(T_1=1, T_2=1, D=i)}{P(T_1=1, D=i)P(T_2=1, D=i)} \quad \theta_i \in (0,\infty)\end{aligned} \tag{8}$$

If $\theta_i = 1$ the tests results are independent; if $\theta_i < 1$ there is negative association between tests and if $\theta_i > 1$, the association between tests is positive.

$$\alpha_i = \frac{P(T_1=1, T_2=1, D=i)P(T_1=0, T_2=0, D=i)}{P(T_1=1, T_2=0, D=i)P(T_1=0, T_2=1, D=i)} \quad \alpha_i \in (0,\infty) \tag{9}$$

Thus, α_i is defined as the odds ratio in the ith diseased state, and when $\alpha_i = 1$ we have independence between tests; negative dependence is expressed by $\alpha_i < 1$ and positive dependence by $\alpha_i > 1$.

In spite of the fact that, both indexes measure dependence and they are within of the same range of values, they are different in nature. To establish the relationship between them, the authors considered a reparametrization given by: $a_i = \theta_i P(T_1 = 1 \mid D = i) = P(T_1 = 1 \mid T_2 = 1, D = i)$, $\quad b_i = \theta_i P(T_2 = 1 \mid D = i) = P(T_2 = 1 \mid T_1 = 1, D = i)$ and $\eta_i = \frac{1}{\theta_i}$.

Let us rewrite the cell probabilities in the cross-tabulation as:

$$P(T_1=1, T_2=1, D=i) = \eta a_i b_i,$$

$$P(T_1 = 0, T_2 = 1 = \eta a_i(1 - b_i)),$$

$$P(T_1 = 1, T_2 = 0, D = i) = \eta(1 - a_i)b_i$$

and

$$P(T_1 = 0, T_2 = 0, D = i) = \eta(1 - a_i)(1 - b_i) + 1 - \eta$$

In this way, the α_i parameter can be expressed in terms of the parameter η_i, by,

$$\alpha_i = 1 + \frac{1 - \eta_i}{\eta_i(1 - a_i)(1 - b_i)} \tag{10}$$

The BP indexes were developed assuming that the tests have the same dependence within the disease and non-disease populations ($\theta_D = \theta_{ND}$ and $\alpha_D = \alpha_{ND}$) and they are useful when the design of the study does not include the verification with "gold standard" of those individuals with negative outcome in both screening tests. So, using the θ index, we can to estimate the unknown quantities of disease and non-disease individuals,

$$n_D = a + b + c + [d] \quad \text{and} \quad n_{ND} = e + f + g + [h]$$

as follows:

$$\begin{aligned} \widehat{n}_D &= \widehat{\theta}\left\{\frac{(a+b+1)(a+c+1)}{(a+1)} - 1\right\} = \widehat{\theta}q_1 \\ \widehat{n}_{ND} &= \widehat{\theta}\left\{\frac{(e+f+1)(e+g+1)}{(e+1)} - 1\right\} = \widehat{\theta}q_2 \end{aligned} \tag{11}$$

On the other hand, with the α index we can to estimate the unknown quantities d and h, as follows:

$$\begin{aligned} \widehat{d} &= \widehat{\alpha}\left\{\frac{(b+1)(c+1)}{(a+1)} - 1\right\} = \widehat{\alpha}(r_1 - 1) \\ \widehat{h} &= \widehat{\alpha}\left\{\frac{(f+1)(g+1)}{(e+1)} - 1\right\} = \widehat{\alpha}(r_2 - 1) \end{aligned} \tag{12}$$

where

$$\widehat{\theta} = \frac{n}{q_1 + q_2}, \qquad \widehat{\alpha} = \frac{u+2}{r_1 + r_2}$$

u is the quantity of individuals not verified by the "gold standard" and n is the total of participants in the screening study.

Assuming the three dependence structures for the two diagnostic tests, using the results showed in Tables 2 and 3 and the equations (8), (9), we obtained the analytic relationship between θ_i and α_i with ψ_i φ_i and ϕ_i.

- Diseased individuals population:

Binary Covariance

$$\theta_D = 1 + \frac{\psi_D}{S_1 S_2},$$

$$\alpha_D = \frac{[S_1 S_2 + \psi_D][(1 - S_1)(1 - S_2) + \psi_D]}{[S_1(1 - S_2) - \psi_D][S_2(1 - S_1) - \psi_D]},$$

$$\varphi_D = \psi_D[S_1 S_2(1 - S_1)(1 - S_2)]^{-1},$$

$$\phi_D = -[\ln S_1 \ln S_2]^{-1} \ln[\psi_D S^{-1} S_2^{-1} + 1]$$

FGM copula

$$\theta_D = 1 + \varphi_D(1 - S_1)(1 - S_2);$$

$$\alpha_D = \frac{S_1 S_2[1 + \varphi_D(1 - S_1)(1 - S_2)](1 - S_1)(1 - S_2)[1 + \varphi_D S_1 S_2]}{S_1(1 - S_2)[1 - \varphi_D S_2(1 - S_1)]S_2(1 - S_1)[1 - \varphi_D S_1(1 - S_2)]}$$

Gumbel copula

$$\theta_D = \exp\{-\phi_D \ln S_1 \ln S_2\},$$

$$\alpha_D = \frac{\exp\{-\phi_D \ln S_1 \ln S_2\}[S_1 S_2(1 - S_1)(1 - S_2)]}{[S_1 - S_1 S_2 \exp\{-\phi_D \ln S_1 \ln S_2\}][S_2 - S_1 S_2 \exp\{-\phi_D \ln S_1 \ln S_2\}]}$$

- Non-diseased individuals population:

Binary Covariance,

$$\theta_{ND} = 1 + \frac{\psi_{ND}}{(1 - E_1)(1 - E_2)},$$

$$\alpha_{ND} = \frac{[(1 - E_1)(1 - E_2) + \psi_{ND}][E_1 E_2 + \psi_{ND}]}{[(1 - E_1)E_2 - \psi_{ND}][E_1(1 - E_2) - \psi_{ND}]},$$

$$\varphi_{ND} = \psi_D[E_1 E2(1 - E_1)(1 - E_2)]^{-1},$$

$$\phi_{ND} = -[\ln(1 - E_1)\ln(1 - E_2)]^{-1} \ln[\psi_{ND}(1 - E_1)^{-1}(1 - E_2)^{-1} + 1]$$

FGM copula,

$$\theta_{ND} = 1 + \varphi_{ND} E_1 E_2,$$

$$\alpha_{ND} = \frac{(1 - E_1)(1 - E_2)[1 + \varphi_{ND} E_1 E_2]E_1 E_2[1 + \varphi_{ND}(1 - E_1)(1 - E_2)]}{E_1(1 - E_2)[1 - \varphi_{ND} E_2(1 - E_1)]E_2(1 - E_1)[1 - \varphi_{ND} E_1(1 - E_2)]}$$

Gumbel copula,

$$\theta_{ND} = \exp\{-\phi_{ND} \ln(1 - E_1)\ln(1 - E_2)\},$$

$$\alpha_{ND} = \frac{[(1 - E_1)(1 - E_2)\exp\{-k\}][E_1 + E_2 - 1 + (1 - E_1)(1 - E_2)\exp\{-k\}]}{[(1 - E_1) - (1 - E_1)(1 - E_2)\exp\{-k\}][(1 - E_2) - (1 - E_1)(1 - E_2)\exp\{-k\}]}$$

where, $k = \phi_{ND} \ln(1 - E_1)\ln(1 - E_2)\}$.

For two independent tests, we obtain $\lambda_i = 1$ and $\delta_i = 1$ when $\psi_i = 0$, $\varphi_i = 0$ and $\phi_i = 0$, regardless of the performance test values.

In all cases, the Böhning and Patilea's association indexes (BP indexes) are functions of the performance characteristics of the tests, and when the performance parameters are going to one or zero, the BP indexes could go to infinity or to be indeterminate. We have in Table 3, the limit values of the BP indexes when the test parameters are going to zero or one and the dependence coefficients are fixed at their extreme values.

Table 3: Limits of θ and α indexes when the performance test parameters (PTP), are going to zero or one and $\psi_i = \pm 1$, $\varphi_i = \pm 1$ and $\phi_i = 1$. (for diseased individuals, PTP are S_1 and S_2; for non-diseased individuals, PTP are E_1 and E_2)

				Limit values of θ and α				
Coefficient	Population	Limit values of PTP		$\psi_i = -1$	$\psi_i = 1$	$\varphi_i = -1$	$\varphi_i = 1$	$\phi_i = 1$
θ_D	Diseased individuals	0	0	$-\infty$	$+\infty$	0	2	0
		1	1	0	2	1	1	1
		0	1	$-\infty$	$+\infty$	1	1	1
		1	0	$-\infty$	$+\infty$	1	1	1
θ_{ND}	Non-diseased individuals	0	0	0	2	1	1	1
		1	1	$-\infty$	$+\infty$	0	2	0
		0	1	$-\infty$	$+\infty$	1	1	1
		1	0	$-\infty$	$+\infty$	1	1	1
α_D	Diseased individuals	0	0	0	2	0	2	0
		1	1	0	2	0	2	0
		0	1	1/2	$+\infty$	1/2	$+\infty$	0
		1	0	1/2	$+\infty$	1/2	$+\infty$	0
α_{ND}	Non-diseased individuals	0	0	0	2	0	2	0
		1	1	0	2	0	2	0
		0	1	1/2	$+\infty$	1/2	$+\infty$	0
		1	0	1/2	$+\infty$	1/2	$+\infty$	0

The relationship between covariance and copula parameters is not shown in Table 3, given that the covariance has zero as limit value in all combinations of extreme values made with the copula parameters and performance test parameters.

Observe that, under the hypothetical situation where we have two binary tests with the same perfect association within each individuals group ($\psi_D = -1$ and $\psi_{ND} = -1$ or $\psi_D = 1$ and $\psi_{ND} = 1$) if the tests have perfect negative association, we need for both tests to have absolutely perfect sensitivities ($S_1 = S_2 = 1$) and absolutely imperfect specificities ($E_1 = E_2 = 0$), to model association using θ_i; otherwise, we can not use it. If the tests have perfect positive association, we can model associations with θ_i values belonging to the interval $[2, \infty)$ provided that the performance parameters belong to the interval $(0, 1)$. In this way, values of θ_i very close to 2, will be related with ψ_i values close of zero.

On the other hand, if we have two tests, whose perfectly negative or positive dependence structure can be modeled using the FGM copula, we only can have agreement between the copula parameter and the BP indexes, when both sensitivities are equal to zero and both specificities are equal to one; then, under those conditions, we can only model associations when θ_i belongs to the interval $[0, 2]$. When the test parameters take values inside the within $(0, 1)$, the θ_i parameter would be indicating independence between tests while the FGM is indicating strong dependence between them.

For the Gumbel copula, we evaluated the extreme value 1 because this model is only applicable for positive dependence and 0 indicates independence. When the Gumbel dependence parameter indicates perfect positive dependence between two tests, both with absolutely imperfect sensitivities $S_1 = S_2 = 0$ or both with absolutely perfect specificities $E_1 = E_2 = 1$, the θ_i index takes the zero value indicating perfect negative association. When the tests have performance parameters belonging to interval $(0, 1)$, the θ_i parameter indicates independence between tests when those have perfect Gumbel dependence.

When the diagnostic tests have perfect FGM dependence, the θ index indicates independence and only when both tests have perfect specificities and absolutely imperfect sensitivity ($S_j = 0$), the index expresses a very weak association between tests ($\theta \in [0, 2]$).

Based on these facts, we observe that, the α_i parameter has a performance better than the θ_i parameter in their relations with the other parameters of association. For all combinations of sensitivity and specificity, α_i takes values within the range allowed by definition. When we have two binary tests negatively or positively associated with extreme values in their tests parameters, α_i takes values in the interval $[0, 2]$ for both populations, whereas, when the two tests have performance parameters belonging to the interval $(0, 1)$, the α_i parameter takes values in the interval $[1/2, \infty)$ for diseased and non-diseased individuals. For tests with dependence structure modeled by the FGM copula, the behaviour of α_i within groups of individuals is similar to that observed when we have two binary tests. When de dependence structure responds to the perfectly dependent Gumbel copula, in both populations, α_i indicates independence between tests regardless of the values of their performance parameters.

4. Examples

To illustrate the performance of the indexes, we show two examples, one of them with simulated data and the other one with a data set used by Bohning & Patilea (2008) to illustrate their methodology.

4.1. Example with Simulated Data

As a first example, we simulated 10000 pairs of observations with binary dependence structure and the same number of pairs of data for each copula structure (1000 diseased individuals and 9000 non-diseased individuals), considering the following conditions:

- Three dependence levels: weak (0.2), moderate (0.5) and strong (0.9),
- The dependence is the same in both populations ($\psi_D = \psi_{ND}, \quad \varphi_D = \varphi_{ND}$ and $\phi_D = \phi_{ND}$)

- The specificities of the dependent tests are the same ($E_2 = E_3 = 0.95$) and the prevalence is relatively lower ($p = 0.10$)
- Stage 1: the dependent tests have the same relatively high sensitivities ($S_1 = S_2 = 0.85$)
- Stage 2: the dependent tests have the same relatively low sensitivities ($S_1 = S_2 = 0.45$)

We wrote a program in R to simulate pairs of variates with the different dependence forms. To simulate outcomes of the correlated binary variables Z_1, Z_2 we implemented the algorithm developed by Park, Park & Shin (1996) and to simulate the variables T_1, T_2 with FGM structure and the variables V_1,V_2 with Gumbel structure, we implemented algorithms introduced by Johnson (1987) as follows:

1. Binary data (ψ is the correlation coefficient)
   - Initialize p_1, p_2, $q_1 = 1 - p_1$, $q_2 = 1 - p_2$ and ψ_{12}
   - Let $\lambda_{11} = log\left\{1 + q_1 p_1^{-1}\right\}$, $\lambda_{22} = log\left\{1 + q_2 p_2^{-1}\right\}$ and $\lambda_{12} = \left\{1 + \psi_{12}\sqrt{\frac{q_1 q_2}{p_1 p_2}}\right\}$
   - Generate $X_1 \sim \text{Poisson}(\lambda_{11} - \lambda_{12})$, $X_2 \sim \text{Poisson}(\lambda_{22} - \alpha_{12})$ and $X_3 \sim \text{Poisson}(\lambda_{12})$
   - Set $Y_1 = X_1 + X_3$ and $Y_2 = X_2 + X_3$
   - Set $Z_1 = 1$ if $Y_1 = 0$, else $Z_1 = 0$ and $Z_2 = 1$ if $Y_2 = 0$, else $Z_2 = 0$
   - Then, $Z_j \sim$ Bernoulli (p_j); $j = 1, 2$ and ψ_{12} is the correlation coefficient.
2. FGM data (φ is the dependence parameter)
   - Initialize φ
   - Generate variates $U_1 \sim U(0,1)$, and $U_2 \sim U(0,1)$
   - Set

$$\begin{aligned} T_1 &= U_1 \\ A &= \varphi(2U_1 - 1) - 1 \\ B &= [1 - 2\varphi(2U_1 - 1) + \varphi^2(2U_1 - 1)^2 + 4\varphi U_2(2U_1 - 1)]^{1/2} \\ T_2 &= 2U_2/(B - A) \end{aligned}$$

3. Gumbel data (ϕ is the dependence parameter)
   - Initialize ϕ
   - Generate $U_1 \sim U(0,1)$, $U_2 \sim U(0,1)$ and $U_3 \sim U(0,1)$
   - Set $W_1 = -\ln(U_1)$ and $Y = -\ln(U_2)$
   - Compute $\beta = 1 + \phi W_1$ and $q = (\beta - \phi)/\beta$

- If $U_3 < q$, set $W_2 = \beta Y$ stop
- If $U_3 \geq q$, generate $U_4 \sim U(0,1)$, set $X_2 = \beta(Y - \ln U_4)$ and stop
- Let $V_1 = 1 - e^{-W_1}$ and $V_2 = 1 - e^{-W_2}$

As our data resulted from simulation, so we know all frequencies of individuals, but for the data analysis, we assume that we only have the total number of individuals with negative results in both tests.

The data analysis was made using the Bayesian paradigm, for that, we assumed that the screening tests have positive dependence ($P(\psi < 0) = P(\varphi < 0) = 0$) and we used the Beta(17,122), Beta(39.5, 39.5) and Beta(122, 17) as informative prior distributions for the weak, moderate and strong dependences respectively. To obtain the estimates we used a code in Winbugs software and we simulate $60,000$ Gibbs samples from the conditional distribution of each parameter. From these generated samples, we discarded the first $10,000$ samples to eliminate the effect of the initial values. The results obtained are showed in Table 4

Table 4: Simulated data with three different dependence structures in two scenarios and BP indexes estimates.

Scenary	f_i	$\psi = 0.2$	$\psi = 0.5$	$\psi = 0.9$	$\varphi = 0.2$	$\varphi = 0.5$	$\varphi = 0.9$	$\phi = 0.2$	$\phi = 0.5$	$\phi = 0.9$
	a	745	800	832	712	725	721	652	551	462
	b	116	68	13	131	137	133	152	175	190
	c	93	59	12	133	116	123	153	195	188
$S_1 = 0.85$	d	46	73	143	24	22	23	43	79	160
$S_2 = 0.85$	e	95	241	405	20	22	17	23	14	7
$E_1 = 0.95$	f	332	207	53	430	438	417	354	294	249
$E_2 = 0.95$	g	352	213	46	427	420	451	360	308	266
	h	8221	8339	8496	8123	8120	8115	8306	8384	8478
	$\hat{\theta}$	3.37	5.63	7.31	0.94	1.01	0.81	1.22	1.33	1.74
	$\hat{\alpha}$	6.67	44.5	1336	0.93	1.01	0.77	0.94	1.43	2.00
	a	238	350	433	198	209	201	275	334	406
	b	179	119	21	245	254	261	246	220	214
	c	214	105	33	239	239	238	223	246	209
$S_1 = 0.45$	d	369	426	513	298	298	300	256	200	171
$S_2 = 0.45$	e	95	241	405	20	22	17	19	17	15
$E_1 = 0.95$	f	332	207	53	430	438	417	353	329	255
$E_2 = 0.95$	g	352	213	46	427	420	451	365	320	264
	h	8221	8339	8496	8123	8120	8115	8263	8332	8466
	$\hat{\theta}$	3.59	6.88	10	0.94	0.94	0.81	1.23	1.34	1.76
	$\hat{\alpha}$	6.20	39.8	1130	0.93	0.93	0.78	1.28	1.41	1.99

The results presented in Table 4, confirm those shown in Table 3. When the data have a binary structure with linear dependence, both BP indexes tend to have high values. It is important to point out that, the sensibility has little effect on the index estimates but with low sensitivities the θ index shows a slight increase and the α index shows a opposite behaviour. If the data have a low or moderate FGM dependence both indexes express independence while for high FGM dependency both indexes indicate negative association in the data, that behaviour remains independent of the sensitivity of the tests. With low or moderate type Gumbel dependences, the BP indexes indicate independence between tests, while with high Gumbel dependences the indexes express low dependence.

Table 5: Estimates of BP indexes and unknown quantities of diseased and non-diseased individuals within group with negative outcome in both screening tests, using data with binary and copula dependence. (n_D = 1,000; n_{ND} = 9,000 and $n_{(D+ND)}$ = 10,000)

Scenary	Dependence	BP Index	$\widehat{n}_D$	$\widehat{n}_{ND}$	$\widehat{d}$	$\widehat{h}$	$\widehat{n}_{(D+ND)}$
	$\psi = 0.2$	$\widehat{\theta} = 3.37$	3,264	6,728	2,310	5,949	9,991
		$\widehat{\alpha} = 6.67$	1,046	8,943	92	8,164	9,989
$S_1 = S_2 = 0.85$	$\psi = 0.5$	$\widehat{\theta} = 5.63$	5,247	4,747	4,320	4,086	9,994
$E_1 = E_2 = 0,95$		$\widehat{\alpha} = 44.5$	1,113	8,801	186	8,104	9,913
	$\psi = 0.9$	$\widehat{\theta} = 7.31$	6,266	3,728	5,409	3,224	9,994
		$\widehat{\alpha} = 1336$	1,149	7,518	292	7,014	8,666
	$\varphi = 0.2$	$\widehat{\theta} = 0.94$	940	9,043	-36	8,166	9,984
		$\widehat{\alpha} = 0.93$	998	9,002	22	8,125	10,000
$S_1 = S_2 = 0.85$	$\varphi = 0.5$	$\widehat{\theta} = 1.01$	1,010	8,967	32	8,087	9,977
$E_1 = E_2 = 0,95$		$\widehat{\alpha} = 1.01$	1,000	9,000	22	8,121	10,000
	$\varphi = 0.9$	$\widehat{\theta} = 0.81$	810	9,180	-167	8,295	9,990
		$\widehat{\alpha} = 0.77$	994	9,006	17	8,121	10,000
	$\phi = 0.2$	$\widehat{\theta} = 1.22$	1,222	8,759	239	8,022	9,981
		$\widehat{\alpha} = 0.94$	1,000	6,971	17	6,094	7,971
$S_1 = S_2 = 0.85$	$\phi = 0.5$	$\widehat{\theta} = 1.33$	1,328	8,665	341	7,999	9,992
$E_1 = E_2 = 0,95$		$\widehat{\alpha} = 1.42$	1,002	8,997	15	8,331	9,999
	$\phi = 0.9$	$\widehat{\theta} = 1.74$	1,731	8,250	745	7,716	9,981
		$\widehat{\alpha} = 2.00$	1,002	8,996	16	8,462	9,998
	$\psi = 0.2$	$\widehat{\theta} = 3.59$	2,841	7,167	2,210	6,388	10,008
		$\widehat{\alpha} = 6.20$	1,635	8,361	1,004	7,582	9,996
$S_1 = S_2 = 0.45$	$\psi = 0.5$	$\widehat{\theta} = 6.88$	4,194	5,801	3,620	5,140	9,995
$E_1 = E_2 = 0,95$		$\widehat{\alpha} = 39.8$	2,017	7,945	1,443	7,284	9,962
	$\psi = 0.9$	$\widehat{\theta} = 10.0$	4,886	5,100	4,596	4,399	9,986
		$\widehat{\alpha} = 1130$	2,435	6,438	1,948	5,934	8,872
	$\varphi = 0.2$	$\widehat{\theta} = 0.94$	939	9,062	248	8,185	10,001
		$\widehat{\alpha} = 0.93$	976	9,025	285	8,148	10,000
$S_1 = S_2 = 0.45$	$\varphi = 0.5$	$\widehat{\theta} = 0.94$	933	9,069	231	8,191	10,002
$E_1 = E_2 = 0,95$		$\widehat{\alpha} = 0.93$	971	9,027	269	8,149	9,998
	$\varphi = 0.9$	$\widehat{\theta} = 0.81$	816	9,180	116	8,295	9,996
		$\widehat{\alpha} = 0.78$	941	9,060	241	8,175	10,000
	$\phi = 0.2$	$\widehat{\theta} = 1.23$	1,160	8,831	416	8,094	9,990
		$\widehat{\alpha} = 1.28$	999	9,001	255	8,264	9,999
$S_1 = S_2 = 0.45$	$\phi = 0.5$	$\widehat{\theta} = 1.34$	1,289	8,730	489	8,064	10,000
$E_1 = E_2 = 0,95$		$\widehat{\alpha} = 1.41$	1,029	8,971	229	8,305	9,999
	$\phi = 0.9$	$\widehat{\theta} = 1.76$	1,652	8,345	823	7,811	10,000
		$\widehat{\alpha} = 1.99$	1,047	8,951	218	8,417	9,998

Using the index estimate, we computed the estimated unknown quantities of diseased and non-diseased individuals within group with negative outcome in both tests (d and h in Tables 1 and 2), using equations 11 and 12. We observed that, when the data have linear binary dependence the θ index overestimate d and n_D and underestimates h and n_{ND}, that effect is more evident when the covariance level is increased. With weak and moderate linear binary dependences, the α index tends to overestimates all quantities but the observed bias is very little, if the dependence is strong, the observed behaviour is similar to the observed with the other index and it remains regardless of the test sensitivities. See Table 5.

When the diagnostic tests show a weak or strong FGM dependence structure and the sensitivities are higher than 0.5, the θ index takes a value lower than one indicating negative dependence which underestimates n_D and the estimate value for d is negative. For moderate FGM dependences the θ index expresses independence. If the test sensitivities are lower than 0,5, the n_D and d quantities are underestimated while n_{ND} and h are overestimated but the estimation biases are lower than those observed in data with linear binary dependence. For this type of dependences with high sensitivities the α index shows estimates very close to the true quantities but if the tests have sensitivities lower than 0.5, the behaviour is similar to that observed with binary data however the estimation biases are lower. See Table 5.

The obtained results using data within dependence type Gumbel, have a behaviour very similar with the at observed in binary linear dependent data, but in this case, the estimation bias is lower in all cases. Table 5

4.2. Example with Published Data

Bohning & Patilea (2008) used a published data set to illustrate the performance of their two indexes. The authors took the subset of data of serum cholesterol and body mass index as risk factors for cardiovascular disease considered in the Framingham Heart Study (Shurtleff 1974). In agreement with these authors, for that data, conditional on the disease status, the risk factors are positively and significantly associated as when measured by the Mantel-Haenszel odds ratio with the summary taken over disease status. With the estimate values of the indexes, they estimated the quantities of diseased and non-diseased individuals within group with negative outcome in both tests, using for that the equations 11 and 12.

We fit models assuming covariance, FGM dependence and Gumbel dependence and we obtained prevalence, performance test and dependence estimates under Bayesian paradigm. As we have six observed frequencies in the cross table and we have seven parameters to estimate, we have a non-identifiable model. So, in the same manner as Joseph, Gyorkos & Coupal (1995) we fitted models using Beta(a,b) as informative prior distributions on dependence parameters and Beta(1/2, 1/2) as non-informative prior distribution on prevalence and test parameters. Since we have no prior information on dependence parameters, we used the three prior distributions employed in the example with simulated data and we used the Deviance Information Criteria (DIC) to select the better fit. With our estimates, we computed the values of the BP indexes and we assumed that d and h are not known, so we use the estimators $\widehat{d}$ and $\widehat{h}$ to predicting the known n_D and n_{ND}. The results are showed in Table 6.

Table 6: Estimates of indexes and disease classes in a study with completely known disease status. (Observed frequencies: $a = 51$, $b = 19$, $c = 70$, $d = \mathbf{86}$, $e = 69$, $f = 20$, $g = 38$, $h = \mathbf{60}$, $n_{(D+ND)} = 413$)

Model	Index	$\widehat{n}_D$	$\widehat{n}_{ND}$	$\widehat{d}$	$\widehat{h}$	$\widehat{n}_{(D+ND)}$
Böhning*	$\widehat{\theta}$ =1,36	225	188	85	61	413
	$\widehat{\alpha}$ =3.79	242	171	104	44	414
FGM copula*	$\widehat{\theta}$ =1.32	219	183	79	56	402
	$\widehat{\alpha}$ =2.48	205	154	65	27	359
Covariance	$\widehat{\theta}_D$ =1.57	261	-	121	-	-
	$\widehat{\theta}_{ND}$ =1.08	-	202	-	75	463
	$\widehat{\alpha}_D$ =8.32	366	-	227	-	-
	$\widehat{\alpha}_{ND}$ =3.14	-	161	-	34	527
Gumbel copula	$\widehat{\theta}_D$ =0.45	75	-	(-65)	-	-
	$\widehat{\theta}_{ND}$ =0.78	-	129	-	(-11)	204
	$\widehat{\alpha}_D$ =0.23	146	-	6	-	-
	$\widehat{\alpha}_{ND}$ =0.27	-	130	-	3	276

*Models with homogeneous dependence

With this data set, the index values obtained after of fitting a model with Gumbel parameter show negative dependence between test outcomes and the estimates of the unknown quantities are negative which makes no sense, indicating the data do not fit well with the Gumbel copula. The model assuming binary dependence eliminates the assumption of homogeneity retained by the Böhning and FGM models. That model overestimates the numbers of individuals not verified by "gold standard" expressing that dependence between the tests do not have linear binary structure. The results obtained using the model with FGM dependence shows a better fit despite, it tends to underestimate both indexes a little and underestimate the unknown quantities. This implies tending a contradiction because in agreement with Bohning & Patilea (2008) when $\theta > 1$, the expected value of n_i, $(i = D, ND)$ will be below the true value of n and the amount of underestimation is determined by the value of θ; the higher value of θ, the higher the underestimation; therefore, if the index values obtained with FGM fitted model are lower than Böhning index, the estimate quantities for n_i should be higher than those observed with the Böhning model. In both models, the θ index estimate shows better behaviour than the other index.

5. Conclusions

In many clinical diagnostic procedures, it is necessary to use two or more (observable or not) biological traits expressed on a continuous scale in designs that includes verification with gold standard only for those individuals with at least one positive outcome in the screening tests. To obtain the diagnostic, those measures are dichotomized using a cut point, in this way, the final result is one of two values (positive or negative). The continuous traits measured can be correlated in some way (not necessarily linear dependence) but when performing data analysis, can occur dependence is assumed with binary structure and not on the continuous structure. Given that the study planning has verification bias, some values in cross table are unknown so it is very complex to estimate the prevalence and performance test parameters using the maximum likelihood procedure. Many authors have considered the estimation problem using models with latent variables to complete the data set, others as Bohning & Patilea (2008) have developed reparametrizations using the observed incomplete data under binary structure assumption. In this paper, we studied the performance of models developed by Bohning & Patilea (2008) and we compared them with the performance of models that use covariance and copula functions to obtain information on the dependence between diagnostic tests.

Despite the covariance and the θ index have different parametric spaces, within the diseased population, we observed that, regardless of the population (diseased and non-diseased individuals) it exists a perfect linear relation between them, whenever there is the diagnostic tests have binary dependence structure, it is possible that the θ index to take values lower than zero and this range of values is not considered within the construction of the index. To have $\theta < 0$ indicates that we have at least one of the tests with sensitivity zero or at least one test without specificity and both situations are unacceptable in practical terms. The α index does not identify covariances lower than -0.5; and when tests with perfect sensitivities ($S_j = 1$) have a strong dependence expressed by a covariance close to unity, the α index takes values in a very constrained range [0,2] indicating very weak dependence; therefore, in cases where tests with perfect performance have perfect binary dependence structure, the BP index either may indicate values not allowed by construction or may underestimate the true dependence. It is obvious that tests with absolutely perfect or imperfect performance is a hypothetical situation very unlikely to occur in reality. In our simulation study with more realistic conditions, we observed that the BP indexes take values within range $(0, \infty)$, the relationship between covariance and α index grows more exponentially and the same do not show strong changes with the differences in the test sensitivities.

It is totally wrong to use the BP indexes with dichotomized data that initially have some of the two copula dependence structures studied, therefore, to use some of those indexes developed to binary data with dichotomized data, leads to erroneous conclusions regarding the dependence between tests which modify the final diagnostic result. In this work, we used two copula functions that model weak non linear dependences and the BP indexes failed to model them, whenever there are many other copula families which the BP indexes relationship could be studied. On the other hand, when the continuous traits are perfectly dependent with some of the copula structures studied, and we evaluate the dependence hypothesis using the dichotomized results and assume binary covariance as parameter, the estimation leads to conclude that test outcomes are independent from each other what directly affects the estimation of the test performance parameters and the prevalence.

Acknowledgments

The authors thank the editor and reviewers for their constructive suggestions. The first author was supported in part by a grant of the Coordenação de Aperfeiçõamento de Pessoal de Nivel Superior (CAPES); the second author was supported by grant of Conselho Nacional de Pesquisas (CNPq) Brazil. This paper is a part of the doctoral in statistics project of the first author.

[Recibido: enero de 2011 — Aceptado: junio de 2011]

References

Bohning, 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.

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

Enoe, C., Georgiadis, M. P. & Johnson, W. O. (2000), 'Estimation of sensitivity and specificity of two diagnostic tests', *Preventive Veterinary Medicine* **45**, 61–81.

Georgiadis, M. P., Johnson, W. O. & Gardner, I. A. (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. J. (1960), 'Bivariate exponential distributions', *Journal of the American Statistical Association* **55**, 698–707.

Johnson, M. E. (1987), *Multivariate Statistical Simulation*, Wiley and Sons, New York.

Joseph, L., Gyorkos, T. W. & 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.

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

Park, C. G., Park, T. & Shin, D. W. (1996), 'A simple method for generating correlated binary variates', *The American Statistician* **50**(4), 306–310.

Shurtleff, D. (1974), Some characteristics related to the incidence of cardiovascular disease and death: 18-year follow-up, *in* W. B. Kannel & T. Gordon, eds, 'An Epidemiological Investigation of Cardiovascular Disease. The Framinham Study', National Institute of Health, Washington, D. C.

Thibodeau, L. A. (1981), 'Evaluating diagnostic tests', *Biometrics* **37**, 801–804.

Torrance-Rynard, V. L. & Walter, S. D. (1997), 'Effects of dependent errors in the assessment of diagnostic tests performance', *Statistics in Medicine* **16**, 2157–2175.

Vacek, P. M. (1985), 'The effect of conditional dependence on the evaluation of diagnostic tests', *Biometrics* **41**, 959–968.

Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 451 a 460

The Generalized Logistic Regression Estimator in a Finite Population Sampling without Replacement Setting with Randomized Response

El estimador de regresión logística generalizado en un muestreo sin reemplazo con respuesta aleatorizada en poblaciones finitas

Víctor Hugo Soberanis-Cruz[a], Víctor Miranda-Soberanis[b]

División de Ciencias e Ingeniería, Universidad de Quintana Roo, Quintana Roo, México

Abstract

The randomized response technique (RR), introduced by Warner (1965) was designed to avoid non-answers to questions about sensitive issues and protect the privacy of the interviewee. Some other randomized response techniques have been developed as the Mortons technique which was developed based on a finite population sampling without replacement. In this paper we are presenting an estimation of the population (total of individuals N) based on Mortons technique assisted for a logistic regression model and considering a specific sensitive characteristic A, with an auxiliary variable associated to the sensitive variable. Analyses were conducted assuming finite population sampling and based on the p-estimators theory through a model assisted estimator. In addition, we propose an estimator of the variance of the estimator, as well as the results of simulations showing that the model assisted estimator of the variance decreases compared with an estimator which depends of the sampling design.

***Key words*:** Model assisted inference, Randomized response, Sampling design, Sensitive question.

Resumen

La técnica de respuesta aleatorizada (RA) introducida por Warner (1965), fue diseñada para disminuir la no-respuesta sobre aspectos sensibles y para proteger la confidencialidad del entrevistado en muestreos con reemplazo. Otras técnicas RA para muestreos sin reemplazo en poblaciones finitas, como la de Morton, han sido desarrolladas y comparadas. En este trabajo se exponen los resultados de la estimación del total de individuos de una población

[a] Professor. E-mail: vsobera@uqroo.mx
[b] Professor. E-mail: vmsoberanis@gmail.com

finita con la técnica de Morton, considerando una característica específica sensitiva A en un muestreo sin reemplazo y asistido por un modelo de regresión logística, con una variable auxiliar asociada a la variable sensitiva. Se desarrolla en el contexto de poblaciones finitas y en el marco de la teoría de los estimadores-π, a través de un estimador asistido por el modelo. Asimismo, se propone un estimador para la varianza del estimador y se muestra, vía simulación, que este estimador para la varianza disminuye, comparado con otro estimador que depende únicamente del diseño de muestreo.

***Palabras clave*:** diseño de muestreo, inferencia asistida por modelo, pregunta sensitiva, respuesta aleatorizada.

1. Introduction

Surveys represent procedures used by researchers to obtain information about a sample of individuals. Sometimes surveys include one or more questions related to personal information that could be considered as "private" and cause the individual to feel at risk (Méndez, Eslava & Romero 2004), and therefore, the individual refuses to participate or provides untrue responses.

When interviewers try to obtain honest responses, in studies where some sensitive issues, such as drug use, tax evasion, or sexual preferences, through survey sampling, they may face difficulties that intrinsically belong to the interviewee: attitude, time available, a different way of thinking, among others. (Sánchez 1985).

Strategies to minimize resistance from the interviewee to provide the real response, when the topic might represent an invasion of privacy are classified into two types. The first strategy is based on the phrasing of the questions that contain the characteristic that wants to be measure in such a way that indirect questions are used to obtain the real response. The second strategy refers to the method of randomized response (RR), introduced by Warner (1965). The RR is a specially designed method to ensure the privacy of the interviewee when sensitive, delicate, or embarrassing topics are studied. With these two strategies, the researcher avoids that the behavior of those surveyed gets skewed toward socially desirable responses. In this regard, real responses are obtained about sensitive issues (true/false) while assuring the confidentiality of the responses. In other words, the interviewer will not know the real answer.

Warner (1965) developed a technique called Randomized Response that guarantees the anonymity of the interviewee. It consists of a random mechanism that selects one of two complementary questions, as follows: Question (1): "do you have a specific characteristic A?" whereas Question (2) is "do you have the complementary characteristic?" where A represents the sensitive characteristic of interest and, the absence of such characteristic (Estevao, Hidiroglou & Sarndal 1999). The interviewee will provide the answer (yes or no), however, the interviewer will not know which question is answered. In this way, the anonymity of the interviewee is protected. This technique is also known as the Complementary Question Model.

As an alternative to Warner's Model of Complementary Questions Greenberg, Abul-Ela & Horvitz (1969) proposed a Model of Randomized Responses that contain two unrelated questions. One question addresses the sensitive issue of interest and the second one is innocuous. In other words, the second question is non-sensitive.

Horvitz, Greenberg & Abernathy (1976) attributed to R. Morton the idea about the Randomized Response Technique in which the random selection of unrelated questions are made among three options: (1) the sensitive question itself, (2) an instruction that indicates "yes", and (3) an instruction that indicates "no", that can be chosen with their respective probabilities P_1,P_2,P_3, where $P_1 + P_2 + P_3 = 1$. In this paper, we will call this model MU (Morton Unrelated), in allusion to R. Morton who had the original idea.

In order to understand the objective of this paper, we will consider a finite population of N units with a sensitive characteristic y_k, $(k = 1, 2, \dots, N)$ in which the total, $T_y = y_1 + y_2 + \cdots + y_N$ wants to be estimated. The main objective of this paper is to obtain better estimations of T_y (Soberanis, Ramírez, Pérez & González 2008) through Sampling without Replacement and the design of Morton Unrelated (MU), exploring estimators assisted by the Generalized Logistic Regression Estimator (GLRE) proposed by Lehtonen & Veijanen (1998*a*). In addition, the standard deviations of the estimators will be compared through simulation and recommendations will be provided. Thus, in this paper, an estimator of T_y will be proposed using a super population model: The Model of Logistic Regression.

2. The Generalized Logistic Regression Model

Let $U = 1, \dots, k, \dots, N$ be a finite population of participants. The subset $A \subset U$ is defined by a sensitive characteristic A; therefore, the RR technique is used to protect the anonymity of the sample of individuals (Soberanis et al. 2008). We will estimate $T_A = \sum\limits_U y_k$ where $y_k = 1$ if $k \in A$, and $y_k = 0$ if $k \notin A$. The selection of sample S is conducted under the sampling design $p(s)$ with positive inclusion probabilities π_k and π_{kl}, where

$$\pi_k = Pr(S \ni k) = \sum_{S \ni k} p(s) \text{ y } \pi_{kl} = Pr(S \ni \, k \,\&\, l) = \sum_{S \ni \, k \,\&\, l} p(s)$$

The estimator GLRE is generated with the predicted values obtained through the adjustment of the following model (Estevao, Hidiroglou & Sarndal 1995): We need to estimate the total population $T = \sum_U y_k$. A sample S is obtained assigning to unit k the sampling weight $a_k = \frac{1}{\pi_k}$. $\underline{x}$ represents an auxiliary vector of dimension $J \geq 1$, and $\underline{x}_k$ represents the a *priori* value for unit $k \in U$. In this method, data $\{(y_k, \underline{x}_k) : k \in s\}$ are observed. For those units that are not included in the sample, $y - k$ is unknown but it is possible to obtain a value μ_k that approximates y_k for all population units even though the approximation is not the most precise. Now, let

$$T_A = \sum_U \mu_k + \sum_U (y_k - \mu_k)$$

where the sum $\sum_U \mu_k$ is the dominante term and the residual sum $\sum_U (y_k - \mu_k)$, even though it is small, it needs to be estimated. Let us assume that $\underline{x}_k$ is situated in the sampling frame for all $k \in s$. The predicted values $\widehat{y}_k$ are obtained from the supplementary information adjusting a model in such a way that $E_\xi(Y_k \mid \underline{x}_k; \underline{\beta}) = \mu(\underline{x}_k \mid \underline{\beta})$, where E_ξ is the expectation operator under the theoretical model $\xi, \mu(\underline{x}_k \mid \underline{\beta})$. In addition $\xi, \mu(\underline{x}_k \mid \underline{\beta})$ is a specific function, and $\underline{\beta}$ is an unknown vector of parameters in the model. The function of model ξ is to describe the elements in the population in a "reasonable" way as if they would have been generated by the model itself.

However, it is not expected that the population was generated by the model ξ, therefore, conclusions about population parameters, including $\widehat{T}_A$, are independent from assumptions underlying the model.

In this manner, using the data from sampling$\{(y_k, \underline{x}_k : k \in s\}$, $\underline{\widehat{\beta}}$ is obtained as the maximum likelihood pseudo-estimator (MLEP) of $\underline{\beta}$, as it includes the sampling weights. In addition, the predicted values $\widehat{y}_k = \mu(\underline{x}_k \mid \underline{\widehat{\beta}}) = \widehat{\mu}_k$ are calculated, for each $k \in U$. Further, using $\widehat{\mu}$ and the Horvitz-Thompson estimator (HT) for the residual sum, we obtain:

$$\widehat{T}_{LGREG} = \sum_U \widehat{\mu}_k + \sum_s \pi_k^{-1}(y_k - \widehat{\mu}_k) \tag{1}$$

Equation (1) is the GLRE from Lehtonen & Veijanen (1998*a*).

In practice, model ξ works as a way to find $\widehat{\beta}$, in order to use it in the estimation functions.

2.1. The LGREG for Model MU

The Generalized Logistic Regression Estimator $\widehat{T}_{A,LGREG}$ for $\widehat{T}_A$, proposed previously, is an application of the Lehtonen & Veijanen (1998*a*), which, as we have mentioned before, is an estimator assisted for a Logistic Regression Model. In other words, for $\underline{y} = (y_1, \dots, y_k, \dots, y_N)^t$, the following model is suggested:

$$Pr\{Y_k = 1 \mid \underline{x}_k; \underline{\beta}\} = \frac{e^{\underline{x}_k^t \underline{\beta}}}{1 + e^{\underline{x}_k^t \underline{\beta}}},\ k = 1, 2, \dots, N \tag{2}$$

From now on, this super population model will be referred as model ξ. In addition, for Morton's (MU) Random Mechanism (RR), it is defined:

$$Z_k = \begin{cases} y_k & \text{with probability } P_1 \\ 1 & \text{with probability } P_2 \\ 0 & \text{with probability } 1 - P_1 - P_2 \end{cases} \tag{3}$$

$k = 1, 2, \dots, N$. Thus,

$$\begin{aligned} E(Z_k) &= E_\xi E_{RC}(Z_k) \\ &= E_\xi[P_1 y_k + P_2] \\ &= P_1 E_\xi(y_k) + P_2 \\ &= P_1 \mu_k + P_2 \end{aligned} \tag{4}$$

where

$$\mu_k = E_\xi(Y_k) = Pr\{Y_k = 1 \mid \underline{x}_k; \underline{\beta}\} = \frac{e^{\underline{x}_k^t \underline{\beta}}}{1 + e^{\underline{x}_k^t \underline{\beta}}} \tag{5}$$

Therefore, if $\lambda_k = E(Z_k) = Pr\{Z_k = 1 \mid \underline{x}_k; \underline{\beta}\}$, we obtain:

$$\lambda_k = P_1 \mu_k + p_2 \tag{6}$$

Now,

$$t_A = \sum_U y_k = \sum_U \mu_k + \sum_U (y_k - \mu_k) \tag{7}$$

and,

$$\widehat{T}_A = \sum_U \widehat{\mu}_k + \sum_s \frac{(y_k - \widehat{\mu}_k)}{\pi_k} \tag{8}$$

where

$$\widehat{\mu}_k = \mu(\underline{x}_k^t \underline{\widehat{B}}) = \frac{e^{\underline{x}_k^t \underline{\widehat{B}}}}{1 + e^{\underline{x}_k^t \underline{\widehat{B}}}} \tag{9}$$

Thus, $\underline{\widehat{\beta}}$, satisfies the following equation

$$\sum_S \left[y_k - \mu(\underline{x}_k^t \underline{B} \right] \frac{\underline{x}_k}{\pi_k} = \underline{0}$$

where $\underline{\beta}$ represents the population parameter defined for the likelihood equation:

$$\partial \log L(\underline{\beta}) / \partial \underline{\beta} = \underline{0}$$

which is equivalent to the following equation

$$\sum_U \left[y_k - \mu\big(\underline{x}_k^t \underline{B}\big) \right] \underline{x}_k = \underline{0}$$

For practical purposes, we will use either $\underline{\beta}$ or $\underline{B}$.

For the open sampling, the likelihood function $L(\underline{\beta})$ is given by:

$$L(\underline{\beta} \mid \underline{y}) = \prod_{k \in A} \mu\big(\underline{x}_k^t \underline{\beta}\big) \prod_{k \in U-A} \left[1 - \mu\big(\underline{x}_k^t \underline{\beta}\big) \right]$$

Regarding the Random Responses problem, the vector of observations is $\underline{Z}_s = (Z_k)_{k \in s}$ its parameter, $\underline{\lambda} = (\lambda_k)_{k \in s}$, and the likelihood function is given by

$$\begin{aligned} L(\underline{\beta} \mid \underline{z}) &= \prod_U Pr\{Z_k = z_k\} \\ &= \prod_U \lambda_k^{z_k} (1 - \lambda_k)^{1 - z_k} I_{\{0,1\}}(z_k) \end{aligned}$$

as if $(Z_k, \underline{x}_k)$ was observed for each $k \in U$, as in a census. Thus,

$$\begin{aligned} l(\underline{\beta} \mid \underline{z}) &= \ln L(\underline{\beta} \mid \underline{z}) \\ &= \sum_U \Big[z_k \ln \lambda_k + (1 - z_k) \ln \{(1 - \lambda_k)\Big] \end{aligned}$$

It should be noted that the function $l(B \mid z)$ reaches a maximum in $\underline{B}$, and is defined and characterized by the following equation:

$$\frac{\partial l(\underline{B} \mid \underline{Z})}{\partial \underline{B}} = \underline{0} \Leftrightarrow \sum_U \left[\Big(z_k - (P_1 \mu_k + P_2) \Big) \frac{\mu_k (1 - \mu_k)}{P_2(1 - P_2) + P_1 \mu_k (1 - P_1 \mu_k - 2P_2)} \right] \underline{x}_k \\ = \underline{0} \tag{10}$$

Thus, $\underline{B}$ is defined implicitly as the one parameter that maximizes $l(B \mid Z)$.

Also, the estimator π of $l(\underline{B} \mid \underline{z})$ is given by:

$$\widehat{I}_\pi(\underline{\beta} \mid \underline{z}) = \sum_s \pi_k^{-1} \Big[z_k \ln \lambda_k + (1 - z_k) \ln (1 - \lambda_k) \Big]$$

Where λ_k is given by (6).

In addition,

$$\frac{\partial \lambda_k}{\partial \underline{\beta}} = P_1 \frac{\partial \mu_k}{\partial \underline{\beta}} = P_1 \mu_k (1 - \mu_k) \underline{x}_k$$

thus,

$$\begin{aligned} \frac{\partial \widehat{I}_\pi}{\partial \underline{\beta}} &= \sum_s \pi_k^{-1} \left[\frac{z_k}{\lambda_k} \frac{\partial \lambda_k}{\partial \underline{\beta}} - \Big(\frac{1 - z_k}{1 - \lambda_k} \Big) \frac{\partial \lambda_k}{\partial \underline{\beta}} \right] \\ &= P_1 \sum_s \pi_k^{-1} \left[\frac{(z_k - \lambda_k) \mu_k (1 - \mu_k)}{\lambda_k (1 - \lambda_k)} \right] \underline{x}_k \end{aligned}$$

but

$$\frac{\mu_k (1 - \mu_k)}{\lambda_k (1 - \lambda_k)} = \frac{\mu_k (1 - \mu_k)}{P_2(1 - P_2) + P_1 \mu_k (1 - P_1 \mu_k - 2P_2)}$$

then,

$$\begin{aligned} \frac{\partial \widehat{I}_\pi}{\partial \underline{\beta}} &= P_1 \sum_s \pi_k^{-1} \left[(z_k - \lambda_k) \frac{\mu_k (1 - \mu_k)}{P_2(1 - P_2) + P_1 \mu_k (1 - P_1 \mu_k - 2P_2)} \right] \underline{x}_k \\ &= P_1 \sum_s \pi_k^{-1} \left[\Big(z_k - (P_1 \mu_k + P_2) \Big) \frac{\mu_k (1 - \mu_k)}{P_2(1 - P_2) + P_1 \mu_k (1 - P_1 \mu_k - 2P_2)} \right] \underline{x}_k \end{aligned}$$

Therefore, by solving the following equation $\dfrac{\partial \widehat{I}_\pi}{\partial \underline{\beta}} = \underline{0}$, we obtained $\widehat{\underline{\beta}}$.

Once $\widehat{\underline{\beta}}$ is obtained, the estimator proposed for T_A is

$$\widehat{T}_{A,LGREG} = \sum_U \widehat{\mu}_k + \frac{1}{P_1} \sum_s \frac{Z_k - (P_1 \widehat{\mu_k} + P_2)}{\pi_k} \tag{11}$$

where $\widehat{\mu}_k$ is given by (9).

2.2. Estimation of the Estimator Variance

Base on the π estimators theory (Sarndal, Swensson & Wretman 1992, Wretman, Sarndal & Cassel 1977, Lehtonen & Veijanen 1998a), the following estimator is proposed for $Var(\widehat{T}_{A,LGREG})$:

$$\widehat{V}(\widehat{T}_{A,LGREG}) = \left(\frac{1}{P_1^2}\right)\sum\sum_{s}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{Z_k - \widehat{\lambda}_k}{\pi_k}\right)\left(\frac{Z_1 - \widehat{\lambda}_1}{\pi_1}\right) \tag{12}$$

3. An Estimator that only Depends on the Sampling Design

Based on Morton's random response technique and a sampling design $p(s)$, a π-estimador for t_A, is given by Soberanis et al. (2008):

$$\widehat{T}_{A,\pi} = \frac{1}{P_1}\sum_{S}\frac{Z_k}{\pi_k} - \frac{NP_2}{P_1} \tag{13}$$

Its variance is given by

$$V(\widehat{T}_{A,\pi}) = \frac{1}{P_1^2}\left\{\sum\sum_{U}\Delta_{kl}\widehat{\lambda_k}\widehat{\lambda_1} + \sum_{U}\frac{\lambda_k(1-\lambda_k)}{\pi_k}\right\}$$

Where $\widehat{\lambda} = \lambda_k/\pi_k$. So, the estimator proposed for the variance of the estimator is:

$$\widehat{V}(\widehat{T}_{A,\pi}) =$$

$$\frac{1}{P_1^2}\left\{\sum\sum_{s}\breve{\Delta}_{kl}\left(\frac{Z_k}{\pi_k}\right)\left(\frac{Z_1}{\pi_l}\right) + P_1(1-P_1-2P_2)\sum_{s}\frac{\widehat{Z_k}}{\pi_k^2} + P_2(1-P_2)\sum_{U}\frac{1}{\pi_k}\right\}$$

Where $\widehat{Z}_k = \dfrac{Z_k - P_2}{P_1}$ y $\breve{\Delta}_{kl} = \dfrac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}$

4. Simulations using Simple Random Sampling with No-Replacement

This section analyzes the properties of the estimator (11) in the specific case of Simple Random Sampling with No-replacement (SRSN). If the sampling design, $p(s)$, is the SRSN, then,

$$\pi_k = \frac{n}{N} = f; \quad \pi_{kl} = \frac{n(n-1)}{N(N-1)},\ k \neq 1; \quad \pi_{kk} = \pi_k$$

$$\Delta_{kl} = \pi_{kl} - \pi_k\pi_l = \frac{-f(1-f)}{N-1}, k \neq 1; \quad \Delta_{kk} = \pi_k(1-\pi_k) = f(1-f) \tag{14}$$

5. Comparison of the Estimators $\widehat{T}_{A,\pi}$ and $\widehat{T}_{LGREG,\pi}$

In order to compare the estimators $\widehat{T}_{A,\pi}$ and $\widehat{T}_{LGREG,\pi}$, a population with $B = (B_0, B_1)' = (-3, 0.1)'$ with $A = 490$ "successes" using the Accept-Reject algorithm to generate random variables.

5.1. Simulations Results

Table 1: Mean, minimum, maximum and percentiles of estimators $\widehat{T}_{A,\pi}$ and $\widehat{T}_{A,LGREG}$, using $N = 700$, $A = 490$, $n = 140$ and $N = 800$ (Number of simulations).

					Percentiles		
Estimator	Mean	DE	Minimun	25%	50%	75%	Maximum
$\widehat{T}_{A,\pi}$	490.164	39.04	342.857	464.285	492.857	514.285	621.428
$\widehat{T}_{A,LGREG}$	489.818	36.22	383.185	466.997	491.953	513.498	594.419

6. Benefits of the Estimator's Variance of $\widehat{T}_{A,LGREG}$

For the simulated population, Section 6.1, shows that

$$\widehat{V}(\widehat{T}_{A,LGREG}) = \left(\frac{1}{P_1^2}\right) \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{Z_k - \widehat{\lambda}_k}{\pi_k}\right)\left(\frac{Z_l - \widehat{\lambda}_l}{\pi_l}\right)$$

is an excellent estimator of the variance of $\widehat{T}_{A,LGREG}$.

6.1. Simulation Results for the Estimators Variance of $\widehat{T}_{A,LGREG}$

According to Table 2, standard deviation of $\widehat{T}_{A,GLRE}$ is 36.22684, whereas our estimator given by (12) is, on average, 33.601760 and standard deviation 1.344364.

Table 2: Mean, minimum, maximum, and percentiles of the variance of the estimator $\widehat{T}_{A,LGREG}$ ($N = 700$).

					Percentiles		
Estimator	Mean	DE	Minimum	25%	50%	75%	Maximum
$\widehat{V}(\widehat{T}_{A,LGREG})$	33.601	1.344	29.431	32.717	33.660	34.591	38.248

7. Discussion and Conclusions

7.1. Discussion

This paper focuses on the use of auxiliary variables as well as on models for the random response sampling (RR) in finite populations, i.e. in populations where all observation units are identifiable. Furthermore, the sampling designs used were sampling designs with no replacement due to the sensitivity of the variable of interest.

The proposal is based on the work of Lehtonen & Veijanen (1998*a*), Lehtonen & Veijanen (1998*b*), Estevao et al. (1999), which was developed for sampling in finite populations. In other words, for a type of sampling where the researcher is focused on exhaustive data collection until the process to delimit and define the fundamental variables occurs constantly (Pandit 1996, Goulding 2002).

The use of auxiliary variables in a conventional way, i.e. when the variable of interest is correlated with the auxiliary variable in the context of RR, does not necessarily improve the Simple Random Sampling, at least for the Rao-Hartley-Cochran scheme. This happens because the auxiliary information is not used properly, as the variable of interest is a discrete variable. In fact, the proper use of the variable of interest is through a model that assists in the problem of estimating the population total, hence the use of the Generalized Logistic Regression Model.

7.2. Conclusions

If the logistic regression model describes the population adequately, then the estimators GLRE and MU should be used. The results suggest that by using this method, we will obtain a significant reduction in the estimator's variance. It should be noted that it is not necessary that the model is "true", as it represents a process in which the population being studied is generated. However, additional simulations under different conditions should be done in order to compare them to the results obtained in this paper. Only then specific recommendations on the most appropriate approach can be provided.

[Recibido: octubre de 2010 — Aceptado: julio de 2011]

References

Estevao, V., Hidiroglou, M. A. & Sarndal, C. E. (1995), 'Methodological principles for a generalized estimation system at Statistics Canada', *Journal of Official Statistics* **11**, 181–204.

Estevao, V., Hidiroglou, M. A. & Sarndal, C. E. (1999), 'The use of auxiliary information in design-based estimation for domains', *Survey Methodology* **2**, 213–221.

Goulding, C. (2002), *Grounded Theory, A Practical Guide for Management, Business and Market Researchers*, SAGE Publications Ltd.

Greenberg, B. G., Abul-Ela, Abdel-Latif, A. S. W. R. & Horvitz, D. C. (1969), 'The unrelated question RR model: Theoretical framework', *Journal of the American Statistical Association* **64**, 520–539.

Horvitz, D. C., Greenberg, B. G. & Abernathy, J. R. (1976), 'RR: a data gathering device for sensitive questions', *International Statistical Review* **44**, 181–196.

Lehtonen, R. & Veijanen, A. (1998*a*), 'Logistic generalized regression estimators', *Survey Methodology* **24**, 51–55.

Lehtonen, R. & Veijanen, A. (1998*b*), 'On multinomial logistic generalized regression estimators', *Survey Methodology* **24**(1), 51–55.

Méndez, I., Eslava, G. & Romero, P. (2004), 'Conceptos básicos de muestreo', *Monografías* **12**(27), 10–58.

Pandit, N. R. (1996), 'The creation of theory: A recent application of the grounded theory method', *The Qualitative Report* **2**(4).

Sarndal, C. E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer Verlag, New York.

Sánchez, J. L. (1985), 'El tratamiento de preguntas de carácter íntimo: modelo de respuesta aleatorizada', *Revista Estadística Española* **65**, 5–12.

Soberanis, V. H., Ramírez, G., Pérez, S. & González, F. V. (2008), 'Muestreo de respuestas aleatorizadas en poblaciones finitas: Un enfoque unificador', *Agrociencia* **42**, 537–549.

Warner, S. L. (1965), 'Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias', *Journal of the American Statistical Association* **60**, 63–69.

Wretman, J. K., Sarndal, C. E. & Cassel, C. M. (1977), *Foundations of Inference in Survey Sampling*, John Wiley and Sons, Inc., New York.

Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 461 a 476

Pseudo Stochastic Dominance. Applications

Cuasi dominancia estocástica. Aplicaciones

Elena Almaraz-Luengo[a]

Departamento de Estadística e Investigación Operativa, Facultad de Ciencias Matemáticas, Universidad Complutense de Madrid, Madrid, España

Abstract

The aim of this work is to show that on certain ocasions classic decision rules used in the context of options (Stochastic Dominance criteria and Mean-Variance rules) do not provide a selection of one specific option over the other, therefore, the need of working with other criteria that can help us in our choice. We place special interest in economic and financial applications.

***Key words*:** Mean, Variance, Stochastic dominance.

Resumen

El objetivo de este trabajo es mostrar que en ocasiones las reglas clásicas de decisión sobre inversiones (reglas de Dominancia Estocástica y reglas de Media-Varianza) no siempre conducen a una selección de una inversión sobre otra, surgiendo la necesidad de trabajar con otros criterios que ayudan en dicha elección cuando los clásicos no conducen a ninguna selección concreta. Se pone principal interés en las aplicaciones de carácter económico-financiero.

***Palabras clave*:** dominancia estocástica, media, varianza.

1. Introducción

The use of Mean-Variance rules (MV) or Stochastic Dominance rules (SD) may not be as useful as desired, since it might be the case that these criteria do not lead to selection of an investment over another. For example, suppose that there are two investments X and Y, with the following characteristics:

$$E(X) = 20000, \sigma_X = 20.2$$

$$E(Y) = 1, \sigma_Y = 20$$

[a] Professor. E-mail: ealmarazluengo@mat.ucm.es

Where $E(X)$ and $E(Y)$ denote the expectations of X and Y, respectively and σ_X and σ_Y their standard deviations. Note that neither is preferred over the other (X is not preferred over Y, and Y is not preferred over X) using MV criteria, this is because $E(X) > E(Y)$ but $\sigma_X > \sigma_Y$. But there is no doubt that almost all investment decision-makers would select X. That is, MV rules have not been capable of choosing one investment over another even though most decision makers would have selected X.

This problem is not new, for example Baumol (1963) noticed this and suggested a different approach to selecting investments known as "Expected Gain-Confidence Limit Criterion" as a replacement for the MV decision rules. Baumol argued that an investment with a high standard deviation σ will be relatively safe if its expected value μ is large enough. He proposed the following index of risk: $RI = \mu - k\sigma$, where k is a positive constant that represents the level of risk aversion of the investor. Another measure to evaluate an investment is known as Sharpe ratio, which measures the profitability of a title independent from the market, that is, it measures the fluctuation of the investment compared to the market.

Let us now propose the following example in which the SD rules will be applied. Let X be the asset which provides 1 euro with probability 0.01 and provides the value 1000000 euros with probability 0.99; and let Y be the asset which provides 2 euros with probability 1. It would not be strange to expect that nearly 100% of investors would prefer asset X over asset Y, but the SD rules are not conclusive in this case. For example, assume utility function:

$$U(x) = \begin{cases} x, & if \quad x \leq 1 \\ 1, & if \quad x > 1 \end{cases}$$

In this case, it is easily verified that, investors who have this utility function will prefer Y over X. From this, it can be deducted that, these investors who have an "extreme utility" do not represent the majority of investors.

For the reasons discussed above, it has been necessary to establish alternative decision rules to help decide in cases where the above rules (SD and MV) do not allow selection of an investment over another. These rules are known as "Almost Stochastic Dominance rules" (ASD). With ASD rules it is possible that, given two assets X and Y, whose distribution functions do not have any preference using SD rules, but with a " minor change" in the expression of the distribution functions, reveal a preference, and it is possible to select one over another. This small change in the distributions removes extreme preferences (profits), considering the profits that are more common among investors. The utility function above example is a case of extreme utility.

The advantages of ASD over SD and MV are:

1. ASD is able to rank investments in cases where SD and MV are inconclusive.
2. ASD remove from the SD efficient set, alternatives which may be worse for most investors.

3. ASD shed light on the efficient portfolio selection problem and the horizon of the investment. It is possible to establish a relationship between the percentage of equity in the efficient portfolio and the investment horizon. That is, ASD can help investors in choosing their investment portfolio.

Let us continue with the previous example with assets X and Y described above. Let F be the distribution function of X defined as:

$$F(x) = \begin{cases} 0, & if \quad x < 1 \\ 1/100, & if \quad 1 \leq x < 1000000 \\ 1, & if \quad x \geq 1000000 \end{cases}$$

and let G be the distribution function of Y defined as:

$$G(y) = \begin{cases} 0, & if \quad y < 2 \\ 1, & if \quad y \geq 2 \end{cases}$$

Their representation is given in the next figure, in that, it is possible to see how the distributions intersect, also it is representing the area between these two distributions:

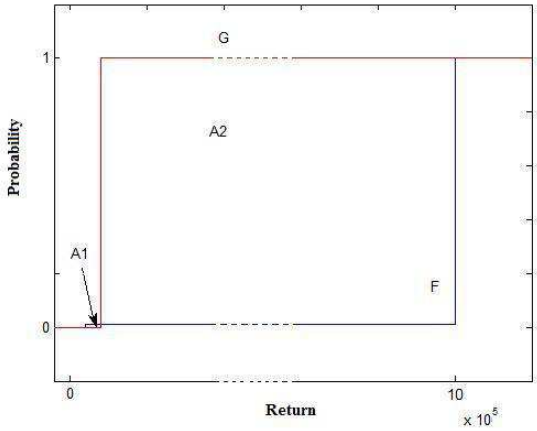

Figure 1: Distributions F and G and area between them.

Although as noted, most investors prefer would F (X) over G (Y), technically, and using the definition of FSD[1], there is no dominance in that sense, because the distributions intersect. Previously, this fact was shown noticing that there are some extreme preferences (profits) which made G preferable (better) to F.

[1] FSD: First order Stochastic Dominance. It is said that random variable X with distribution F dominates random variable Y with distribution G in the first order degree stochastic dominance, if $F(x) \leq G(x)$ for all x and with at least one point in which the inequality is strict.

Moreover, in this example, there is no SSD[2] or MV (for more information about SD or MV see Shaked & Shanthikumar (2007), Almaraz (2009), Almaraz (2010) o Steinbach (2001)). ASD criteria, have come up as an extension of SD criteria to help in these situations. Intuitively, if the area between the two distributions which causes the violation of the FSD criterion (area A_1 in the example) is very small relative to the total area between them (area $A_1 + A_2$ in the figure), then there is dominance of one over another for almost all investors (that is, those with reasonable preference). Hence the name of ASD criteria.

Formally, let S be the range of possible values that both assets can take (or in general two random variables) and S_1 is defined as the range of values in which the FSD rule is violated:

$$S_1(F, G) = \{t : G(t) < F(t)\} \tag{1}$$

where F and G are the distribution functions of the assets (or random variables) under comparison. ε is defined as the quotient between the area in which FSD criterion is violated and the total area between F and G, that is:

$$\varepsilon = \frac{\int_{S_1}(F(t) - G(t))dt}{\int_S |F(t) - G(t)|dt} \tag{2}$$

Another way to write this:

$$\varepsilon = \frac{\int_{S_1}(F(t) - G(t))dt}{\int_{S_1}(F(t) - G(t))dt + \int_{\bar{S}_1}(G(t) - F(t))dt} = \frac{A_1}{A_1 + A_2} \tag{3}$$

where $\bar{S}_1$ denotes the complementary set of S_1 and A_i, $i = 1, 2$ are the areas described previously.

For $0 < \varepsilon < 0.5$, it is said that F dominates G by $\varepsilon - AFSD$. The lower the value of ε the higher degree of dominance. Almost First degree Stochastic Dominance criterion (AFSD) is:

Definition 1. Let F and G be two distribution functions with values in the range of S. It is said that F dominates G by AFSD (for a particular ε, or also ε-AFSD) and it is denoted $F \geq_{AFSD} G$, if and only if:

$$\int_{S_1} [F(t) - G(t)]dt \leq \varepsilon \int_S |F(t) - G(t)|dt \tag{4}$$

where $0 < \varepsilon < 0.5$.

And the definition of Almost Second degree Stochastic Dominance criterion (ASSD) is:

[2] SSD: Second order Stochastic Dominance. It is said that the random variable X with distribution F dominates random variable Y with distribution G in the SSD sense if $\int_{-\infty}^{x}(G(t) - F(t)) \geq 0$ for all x and with at least one point in which the inequality is strict.

Definition 2. Let F and G be two distribution functions with values in the range of S. It is said that F dominates G by ASSD (for a particular ε, or also ε-ASSD) and it is denoted $F \geq_{ASSD} G$, if and only if:

$$\int_{S_2} [F(t) - G(t)]dt \leq \varepsilon \int_S |F(t) - G(t)|dt \tag{5}$$

and $E_F(X) \geq E_G(Y)$, where $0 < \varepsilon < 0.5$ y $S_2(F,G) = \{t \in S_1(F,G) : \int_{\inf S}^t G(x)dx < \int_{\inf S}^t F(x)dx\}$.

It can be shown that AFSD implies condition $E_F(X) \geq E_G(Y)$, but in (5) this implication is not true and therefore must appear in the ASSD definition.

The paper is organized as follows: in first Section, the decision problem will be introduced; in Section 2, principal results in the literature about Almost Stochastic Dominance will be shown, in Section 3, examples of ASD criteria applications in the economic context will be explained (laboratory and real examples, which constitute the main practical contribution of the paper). Finally, in Section 4, main conclusions of this work will be presented.

2. Main Results

In this section, the most noteworthy results about ASD will be described.

Proposition 1. *Let X and Y, be two random variables with distributions F and G respectively. Then:*

1. *F dominates G in the AFSD sense, if and only if, there exists a distribution $\widetilde{F}$ such that $\widetilde{F} \geq_{FSD} G$, and it happens that:*

$$\int_S |F(t) - \widetilde{F}(t)|dt \leq \varepsilon \int_S |F(t) - G(t)|dt \tag{6}$$

2. *F dominates G in the ASSD sense, if and only if, there exists a distribution $\widetilde{F}$ such that $\widetilde{F} \geq_{SSD} G$, and it happens that:*

$$\int_S |F(t) - \widetilde{F}(t)|dt \leq \varepsilon \int_S |F(t) - G(t)|dt \tag{7}$$

That is, the difference between F and $\widetilde{F}$ must be relatively small ($0 < \varepsilon < 0.5$). Having condition $\varepsilon < 0.5$ ensures that it is impossible than both distributions F and G to dominate each other according to AFSD, because if F dominates G in AFSD sense, then $E_F(X) > E_G(Y)$ (see proposition 2).

***Proof*.** See Leshno & Levy (2002). □

Proposition 2. *Let X and Y be two random variables with distribution functions F and G, respectively. If F dominates G in the ε-AFSD sense and F and G are not identical, then $E_F(X) > E_G(Y)$. So, it is impossible that F dominates G in the ε-AFSD sense and that G dominates F in the ε-AFSD sense.*

***Proof*.** See Leshno & Levy (2002). □

As in the case of SD, there is also a characterization of the ASD criteria by utility functions. To address this issue, it is necessary to define the following sets:

Definition 3. Let S be the support of the random variables X and Y, the following sets are defined:

- Let U_1 be the set of all non-decreasing and differentiable utility functions, $U_1 = \{u : u' \geq 0\}$.
- Let U_2 be the set of all concave and two time differentiable utility functions, $U_2 = \{u : u' \geq 0, u'' \leq 0\}$.
- $U_1^*(\varepsilon) = \{u \in U_1 : u' \leq \inf\{u'(x)\}[\frac{1}{\varepsilon} - 1], \forall x \in S\}$.
- $U_2^*(\varepsilon) = \{u \in U_2 : -u'' \leq \inf\{-u''(x)\}[\frac{1}{\varepsilon} - 1], \forall x \in S\}$.

Theorem 1. *Let X and Y be two random variables with distribution functions F and G respectively.*

1. *F dominates G in the ε-AFSD sense, if and only if, for all function $u \in U_1^*(\varepsilon)$ it happens that $E_F(u) \geq E_G(u)$.*
2. *F dominates G in the ε-ASSD sense, if and only if, for all function $u \in U_2^*(\varepsilon)$ it happens that $E_F(u) \geq E_G(u)$.*

***Proof*.** See Leshno & Levy (2002). □

Proposition 3. *Let X and Y be two random variables with distribution functions F and G respectively.*

1. *F dominates G in the FSD sense, if and only if, for all $0 < \varepsilon < 0.5$, F dominates G in the ε-AFSD sense.*
2. *F dominates G in the SSD sense, if and only if, for all $0 < \varepsilon < 0.5$, F dominates G in the ε-ASSD sense.*

***Proof*.** The first part of the proposition will be proven.

Let us assume that F dominates G in the FSD sense, then for all t it happens that $S_1(F, G) = \emptyset$, in this way, for all $0 < \varepsilon < 0.5$:

$$\int_{S_1} [F(t) - G(t)]dt = 0 \leq \varepsilon \int_S |F(t) - G(t)|dt,$$

and F dominates G in the $\varepsilon - AFSD$ sense. Let us now assume that for all $0 < \varepsilon < 0.5$, F dominates G in the $\varepsilon - AFSD$ sense. If $\mu(S_1) = 0$, where μ denotes the Lebesgue's measure over $\mathbb{R}$, then as F and G are non-decreasing and continuous on the right functions, for all t, $F(t) \leq G(t)$, that is, F dominates G in the FSD sense. If $\mu(S_1) > 0$ and there is no FSD, it will be proven that there is no AFSD for some $\varepsilon > 0$.

It will be denoted by $\varepsilon_0 = \int_{S_1} [F(t) - G(t)]dt > 0$. For $\varepsilon = \frac{\varepsilon_0}{2\int_S |F(t)-G(t)|dt}$, we have $\varepsilon_0 = 2\varepsilon \int_S |F(t) - G(t)|dt > \varepsilon \int_S |F(t) - G(t)|dt$. That is, F does not dominate G for any ε, as intended to prove.

Part 2 is analogous. □

3. Financial Applications of Almost Stochastic Dominance

Many authors argue that as the investment horizon increases, an investment portfolio with a higher proportion of assets will dominate, or will be preferred over a portfolio of predominantly government bonds, although this is not in accordance with SD rules, that is, in this case there is some type of dominance, ASD. Therefore, investors prefer long-term assets over bonds, moreover, as the investment horizon increases, the set of "almost all" investors becomes the set of "all" the investors. (See Bernstein (1976), Leshno & Levy (2002) and Bali, Demirtas, Levy & Wolf (2009)).[3]

Examples of this fact will be proposed.

Example 1. Let us consider two simple investments: one bond which has an annual return of 9% with probability 1, and one asset which annual return of -5% with probability 0.5, and 35% with probability 0.5. The target is defining what type of investment is more attractive for investors. The fact mentioned above, will be confirmed, as the horizon of the investment advances, the asset will be more clearly preferred over bonds.

Let X be the random variable which represents the annual return of the asset and let Y be the random variable which represents the annual return of the bond. Let F be the distribution function of X, and G the distribution function of Y. The return of the asset in the first year is $X^{(1)} = 1 + X_0$, being X_0 the initial capital destined to the investment in assets and for the case of the bonds, this will be $Y^{(1)} = 1 + Y_0$ with Y_0 the initial capital destined to the investment in bonds. The return after n periods (years, in this case) will be $X^{(n)} = \prod_{i=1}^{n}[1 + X^{(i)}]$ and $Y^{(n)} = \prod_{i=1}^{n}[1 + Y^{(i)}]$ in assets and bonds, respectively.

For this example, it will be assumed, without loss of generality, that $X_0 = 1 = Y_0$. The procedure that will be followed is to calculate, for each year n, the possible returns of the investment in assets and bonds; this will provide a series of values for random variables with their respective probabilities. After, the

[3] This will be clarified later.

associated distributions will be calculated, they will be denoted as $F^{(n)}$ and $G^{(n)}$ for the assets and bonds, respectively.

For example, for the first year, the returns obtained for the assets are:

$$1 \quad u.m. \begin{cases} 1-0.05*1=0.95 & u.m. \\ 1+0.35*1=1.35 & u.m. \end{cases}$$

where u.m. denotes monetary units, and for the bonds:

$$1 \quad u.m. \longrightarrow 1+0.09*1=1.09 \quad u.m.$$

In this way:

$$F^{(1)}(x)=\begin{cases} 0, \quad if \quad x<0.95 \\ 0.5, \quad if \quad 0.95\leq x<1.35 \\ 1, \quad if \quad x\geq 1.35 \end{cases}$$

and

$$G^{(1)}(x)=\begin{cases} 0, \quad if \quad x<1.09 \\ 1, \quad if \quad x\geq 1.09 \end{cases}$$

These distributions do not verify the FSD criterion because they intercept, as shown in the graphic:

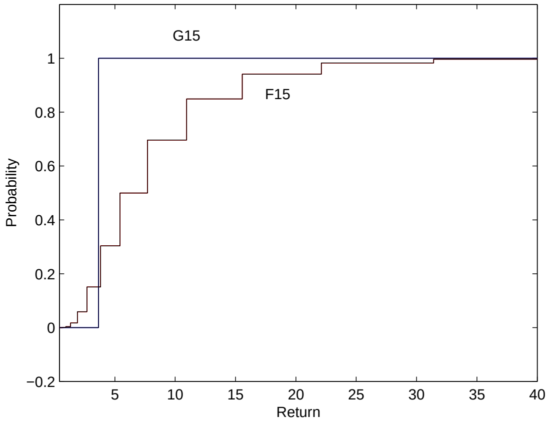

Figure 2: Distributions $F^{(1)}$ and $G^{(1)}$.

For the second year, the returns on the investment in assets are:

$$\begin{cases} 0.95 \quad u.m. \begin{cases} 0.9025 & u.m. \\ 1.2825 & u.m. \end{cases} \\ 1.35 \quad u.m. \begin{cases} 1.2825 & u.m. \\ 1.8225 & u.m. \end{cases} \end{cases}$$

and for bonds:

$$1.09 \quad u.m. \longrightarrow 1.1881 \quad u.m.$$

Then:

$$F^{(2)}(x) = \begin{cases} 0, & if \quad x < 0.9025 \\ 1/4, & if \quad 0.9025 \leq x < 1.2825 \\ 3/4, & if \quad 1.2825 \leq x < 1.8225 \\ 1, & if \quad x \geq 1.8225 \end{cases}$$

and

$$G^{(2)}(x) = \begin{cases} 0, & if \quad x < 1.1881 \\ 1, & if \quad x \geq 1.1881 \end{cases}$$

In this case the graphic is:

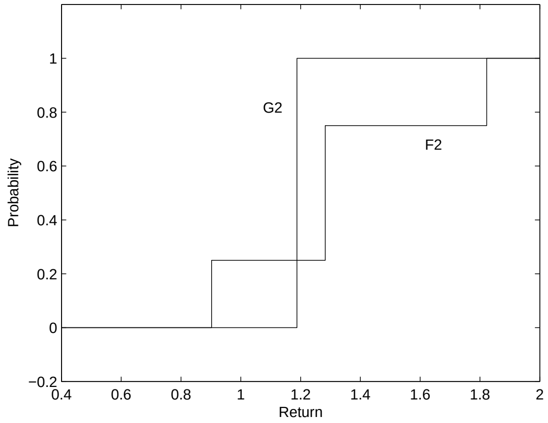

Figure 3: *Distributions* $F^{(2)}$ *and* $G^{(2)}$.

and so on.

Horizons of $1, 2, \ldots, 10, 15$ and 20 years will be considered, and it will be assumed that the investment began in the first of these years. For each year, the value ε will be calculated and it will be proven that this value decreases with the time, reason for which investors will prefer assets to bonds.

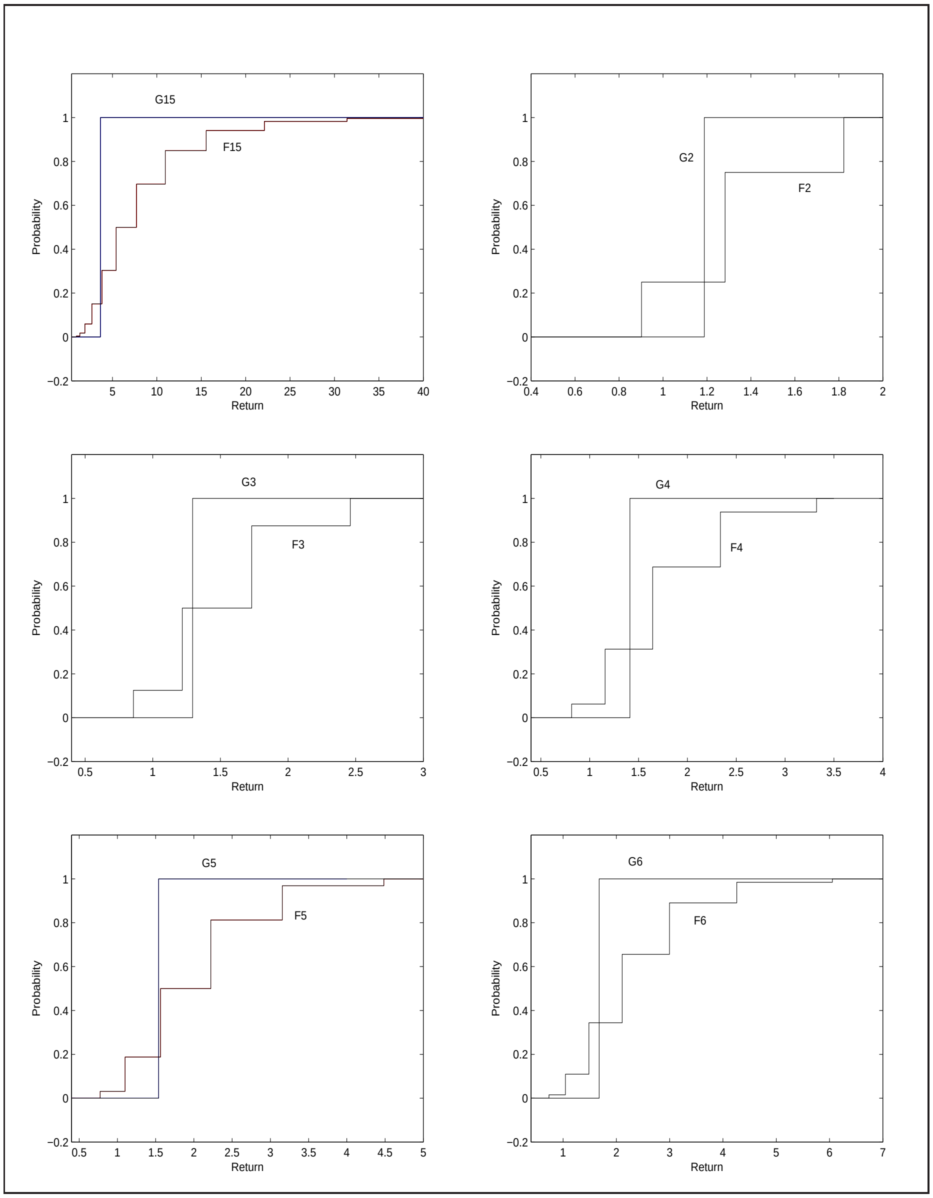

FIGURE 4: Distributions $F^{(n)}$ and $G^{(n)}$, with $n = 1, \dots, 10, 15$ and 20. As observed, the area of violation of the FSD criterion, namely, the area in which $F^{(n)}$ is above $G^{(n)}$-A_1 of the ε definition-, decreases to the extent that the horizon of the investment increases, the value of ε also decreases, that is, as time increases, investors will prefer assets to bonds.

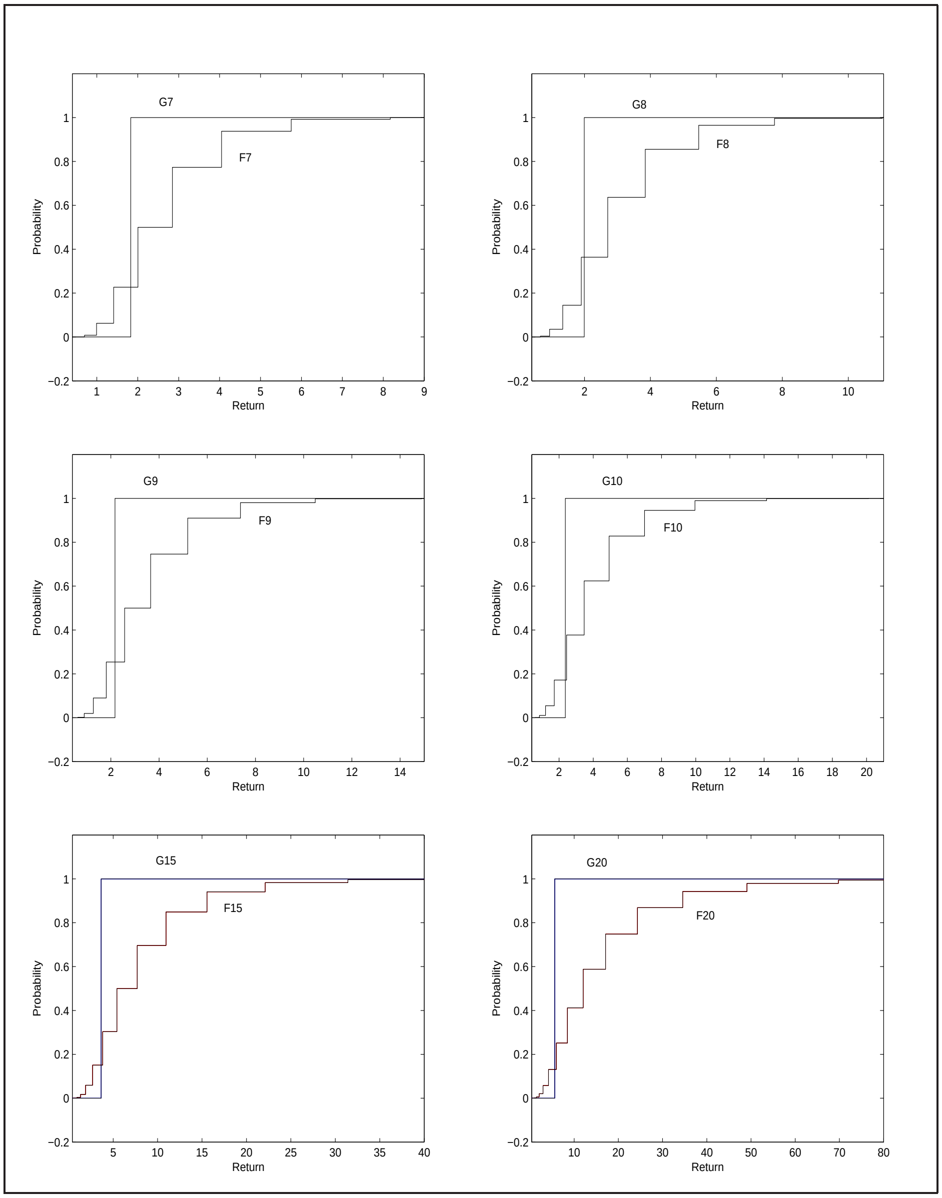

Figure 4: Continuation

Next, ε values will be shown for each horizon of the investment. As shown in the Table 1, these values decrease with time:

Table 1: ε values for each horizon of the investment.

Number of years	ε
1	0.3500
2	0.2576
3	0.2125
4	0.1856
5	0.1406
6	0.1363
7	0.1132
8	0.0972
9	0.0919
10	0.0464
15	0.0414
20	0.0247

Comments at the beginning of this subsection will be explained. As verified, ASD criteria have been used to establish a strong argument in favor of assets over bonds. Let us consider an investor who maximizes expected profits in a period T. Returns are supposed to be independent and identically distributed (i.i.d.) and the investments are supposed to be constant along through time. It is well known that, given different investments with i.i.d. returns and a large enough investment planning horizon, the investment which has higher geometric mean in returns (per period) almost certainly provides a greater benefit than those with lower geometric mean. In the long run, the distribution function of the investment that has a higher geometric mean is almost entirely to the right of the other distributions that represent alternatives, that is, ε decreases with time, as discussed throughout this section. However, there is some controversy in the economic meaning of this fact. Latané (1959), Markowitz (1976) and Leshno & Levy (2004), argue that the decrease in the value of ε is tied to an increase in the range of investor preferences ($U_1^*(\varepsilon)$), that is, they argue that in the long term, all reasonable preferences (profits) are considered. Levy (2009), highlights this fact, saying that, really as time goes by ε decreases (it has been shown in example 1), but the set $U_1^*(\varepsilon)$ does not increase. What happens is that as the periods of the investment increase, the set of all possible values of the random variables also increase, that is, set S is not a fixed set. Of course, if the set S is fixed, the set $U_1^*(\varepsilon)$ increases, but the fact is that S is not fixed. If the last example is observed, set S for the first year is: $[0.95, 1.35]$, for the second year is $[0.9025, 1.8225]$, for the third year is $[0.857375, 2.460375]$, etc.

In summary, there are two facts as time progresses: first ε decreases and this causes an increase of set $U_1^*(\varepsilon)$, and on the other hand, S increases, causing that for a given ε, the set $U_1^*(\varepsilon)$ decreases. The total effect over set $U_1^*(\varepsilon)$ is a mix between these two effects and this depends, on the kind of utility functions that are used.

Definition 4. Given a set S, ε_u is defined as the higher value of ε for which the utility function u still belongs to the set $U_1^*(\varepsilon)$, that is:

$$\varepsilon_u = \left[1 + \frac{\sup\{u'(x), x \in S\}}{\inf\{u'(x), x \in S\}}\right]^{-1} \tag{8}$$

As S increases with time, coefficient $\frac{\sup\{u'(x), x\in S\}}{\inf\{u'(x), x\in S\}}$ increases, and therefore, ε_u decreases. Observe that ε_u shows the higher value of the area alloweb to violate stochastic dominance criteria, for a given utility u such that u still belongs to the set $U_1^*(\varepsilon)$. If $\varepsilon > \varepsilon_u$ then $u \notin U_1^*(\varepsilon)$, otherwise $u \in U_1^*(\varepsilon)$. To be part or not of the set $U_1^*(\varepsilon)$ depends on the speed of decrease of ε and ε_u, that is, the fact that ε decreases is not enough to choose in the long term, it also depends on the utility function.

Let us continue with the last example. Values of ε_u will be calculated for different utility functions u.

Example 2. Let us continue with example 1. Utility functions u will be considered and the associated values of ε_u, will be calculated.

Table 2: Values of ε and ε_u for each horizon of the investment.

Number		ε_u	ε_u	ε_u	ε_u
of	ε	$u(x) = -\exp^{-x}$	$u(x) = \ln(x)$	$u(x) = \frac{x^{1-\alpha}}{1-\alpha}$	$u(x) = \frac{(x-0.2)^{1-\alpha}}{1-\alpha}$
years				$\alpha = 4$	$\alpha = 2$
1	0.3500	0.4013	0.4130	0.1969	0.2984
2	0.2576	0.2849	0.3312	0.0567	0.1579
3	0.2125	0.1676	0.2584	0.0145	0.0780
4	0.1856	0.0754	0.1969	$3.6030 * 10^{-3}$	0.0373
5	0.1406	0.02389	0.1472	$8.8595 * 10^{-4}$	0.0176
6	0.1363	$4.8769 * 10^{-3}$	0.1083	$2.1740 * 10^{-4}$	$8.2874 * 10^{-3}$
7	0.1132	$5.6743 * 10^{-4}$	0.0787	$5.3320 * 10^{-5}$	$3.8923 * 10^{-3}$
8	0.0972	$3.1390 * 10^{-5}$	0.0567	$1.3076 * 10^{-5}$	$1.8269 * 10^{-3}$
9	0.0919	$6.4004 * 10^{-7}$	0.0406	$3.2059 * 10^{-6}$	$8.5653 * 10^{-4}$
10	0.0464	$3.3717 * 10^{-9}$	0.0289	$7.8616 * 10^{-7}$	$4.0100 * 10^{-4}$
15	0.0414	$1.1114 * 10^{-39}$	$5.1121 * 10^{-3}$	$3.2858 * 10^{-7}$	$8.5647 * 10^{-6}$
20	0.0247	$3.8185 * 10^{-176}$	$8.8591 * 10^{-4}$	$6.1816 * 10^{-13}$	$1.5380 * 10^{-7}$

For each representative column of values of ε and ε_u, the decrease mentioned above may be observed. Now, if columns 2 and 3 are compared, it can be shown that ε_u decreases faster than ε and for periods of 1 or 2 years, $\varepsilon < \varepsilon_u$, so $u(x) = -\exp(-x) \in U_1^*(\varepsilon)$, whereas periods strictly exceeding 2 years $u(x) = -\exp(-x) \notin U_1^*(\varepsilon)$. In this case, it is evidenced that the set $U_1^*(\varepsilon)$ does not necessarily increase with time. For this type of utility functions, it is not possible to reason as the authors previously mentioned. In case of working with log-utilities the reasoning is analogous, but for horizons of 5 or less than 5 years, and more than 5 years. In the case of columns 5 and 6, it is verified that $\varepsilon > \varepsilon_u$ for the analized periods, in these cases $u(x) \notin U_1^*(\varepsilon)$ for each studied period.

Example 3. In this case, two financial data series will be considered, in particular series of Ibex 35[4] and Nasdaq Composite indexes[5] corresponding to years from 1926 to 2008. A similar construction as that in the previous example will be performed. In this case, ϵ value is 0.3053, concluding that the Nasdaq series dominates Ibex 35 in a AFSD sense. The illustrative graphic is:

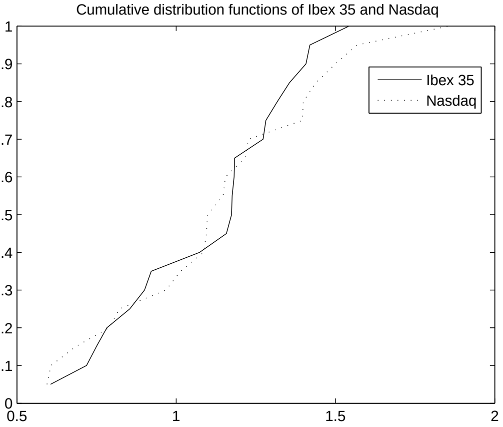

FIGURE 5: Distributions $F^{(1)}$ and $G^{(1)}$ for Ibex and Nasdaq Composite.

4. Conclusions

There are different rules in the literature for comparing investments, for example, Stochastic Dominance rules (SD), Mean-Variance (MV) and Almost Stochastic Dominance (ASD).

SD rules are useful in different areas of knowledge and they arise in a natural way from the need to make comparisons between different choices, using more information available in some situations (distribution functions, density functions, failure rate, etc.) than the mere comparison of averages or other numerical single data.

However, in is situations it may be useful to compare certain functional relationships dependent on means, variances or other measures of uncertainty (for example in the efficient portfolio selection or the scope of the study of the utility). In these cases, MV rules are used.

But sometimes, the use of SD or MV rules is not conducive to a specific selection of an investment over another, consequently, other rules (ASD) arise in response

[4] The official index of the Spanish Continuous Market, which is comprised of the 35 most liquid stocks traded on the market.

[5] A market-capitalization weighted index of the more than 3,000 common equities listed on the Nasdaq stock exchange. The types of securities in the index include American depositary receipts, common stocks, real estate investment trusts (REITs) and tracking stocks. The index includes all Nasdaq listed stocks that are not derivatives, preferred shares, funds, exchange-traded funds (ETFs) or debentures.

to this need for selection. These rules (ASD) are intended to be an extension of SD rules in cases where SD does not respond and they are defined in such manner as to be a useful guide for selection for almost all decision-makers, hence its name.

This paper presents a review of the different classical rules for investment decisions and the importance of ASD concepts selecting some investments over others has been highlighted in cases where there was no clear relationship according to SD and/or MV rules. Likewise, several examples have been proposed, in which applying ASD rules, it has been able to make a clear selection of some investments over others. It is important to note the selection made of Nasdaq Composite Index over the Ibex 35, for an annual series from 1926 to 2008.

[Recibido: septiembre de 2010 — Aceptado: julio de 2011]

References

Almaraz, E. (2009), Cuestiones notables de ordenación estocástica en optimación financiera, Tesis de Doctorado, Universidad Complutense de Madrid, Facultad de Ciencias Matemáticas. Departamento de Estadística e Investigación Operativa, Madrid.

Almaraz, E. (2010), Reglas de decisión en ambiente de riesgo, Tesis de Master, Universidad Nacional de Eduación a Distancia, Facultad de Ciencias Matemáticas. Departamento de Estadística e Investigación Operativa, Madrid.

Bali, T., Demirtas, K., Levy, H. & Wolf, A. (2009), 'Bond versus Stock: Investors' Age and Risk Taking', *http://ssrn.com/abstract=936648* .

Baumol, W. (1963), 'An expected gain-confidence limit criterion for portfolio selection', *Management Science* **10**, 174–182.

Bernstein, P. (1976), 'The time of your life', *Journal of Portfolio Management* **2**(4), 4–7.

Latané, H. (1959), 'Criteria for choice among risky ventures', *Journal of Political Economy* **67**(2), 144–155.

Leshno, M. & Levy, H. (2002), 'Preferred by all and preferred by most decision makers: almost stochastic dominance', *Management Science* **48**(8), 1074–1085.

Leshno, M. & Levy, H. (2004), 'Stochastic dominance and medical decision making', *Health Care Management Science* **7**, 207–2215.

Levy, M. (2009), 'Almost stochastic dominance and stocks for the long run', *European Journal of Operational Research* **194**, 250–257.

Markowitz, H. (1976), 'Investment for the long run: New evidence for an old rule', *Journal of Finance* **31**, 1273–1286.

Shaked, M. & Shanthikumar, G. (2007), *Stochastic Orders*, Springer Series in Statistics, Springer Series in Statistics.

Steinbach, M. C. (2001), 'Markowitz revisited: Mean-variance models in financial portfolio analysis', *SIAM Review* **43**(1), 31–85.

Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 477 a 495

An Application of Semi-Markovian Models to the Ruin Problem

Una aplicación de los modelos semi-markovianos al problema de la ruina

Elena Almaraz-Luengo[a]

Departamento de Estadística e Investigación Operativa, Facultad de Ciencias Matemáticas, Universidad Complutense de Madrid, Madrid, España

Abstract

We consider the classical ruin problem due to Cramér and Lundberg and we generalize it. Ruin times of the considered models are studied and sufficient conditions to usual stochastic dominance between ruin times are established. In addition an algorithm to simulate processes verifying the conditions under consideration is proposed.

***Key words*:** Coupling, Markov chains, Semi-Markov process, Simulation, Stochastic ordering.

Resumen

Se considera el problema clásico de ruina de Cramér y Lundberg y se generaliza. Se estudian los tiempos hasta la ruina de los modelos considerados y se establecen condiciones suficientes para la dominancia estocástica en el sentido usual entre los tiempos de ruina. Por otro lado, se establecen algoritmos de simulación de los procesos bajo estudio y de obtención de estimadores para las probabilidades involucradas.

***Palabras clave*:** cadenas de Markov, dominancia estocástica, emparejamiento, proceso semi-markovianos, simulación.

1. Introduction

The main purpose of the Ruin Theory is to obtain exact formulas or approximations of ruin probabilities in different risk models, see Seal (1969), Gerber (1995) and Ramsay (1992). Some of the most popular approximations are due to Beekman (1969), in which a Gamma distribution is used to approximate the

[a] Professor. E-mail: ealmarazluengo@mat.ucm.es

distribution of the claims, or the approximation due to De Vylder (1996), who approximates the ruin process using a simply process in which the ruin probability is an exponential type. A relatively recent approach to estimate the probability of ruin is presented by Goovaerts (1990), where bounds are established through the ordering of the risks. Another kind of approach arises from the use of nonparametric techniques such as resampling (see Frees 1986) or Monte Carlo simulation (see Beard, Pentikäinen & Pesonen 1984).

Many authors have studied the ruin problem, for example Reinhard (1984) and Asmussen (1989). Reinhard (1984) considers a class of risk models in which the frequency of claims and the quantities to be paid are influenced by an external Markovian process (or environmental process), Reinhard & Snoussi (2001, 2002) have analized the severity of ruin and the distribution of surplus prior to ruin in a discrete semi-Markovian risk model. For more information about risk theory see Beard et al. (1984), Latorre (1992) or Daykin (1994).

In what follows, times to ruin in certain risk models will be ordered without an explicit expression for the probability of ruin and without the use of approximations thereof, as was classically done by Ferreira & Pacheco (2005) and Ferreira & Pacheco (2007).

Many authors have studied these processes int he context of the Queuing Theory. However, they also have applicability for dynamic solvency models and survival analysis.

This paper is organized as follows: in Section 2; the classical Cramér and Lundberg risk model is described; in Section 3, the principal concepts and notation being used in the rest of the paper are defined; in Section 4, the generalized model is described and the principal results are shown; finally, in Section 5 algorithms of simulation of the processes considered in Section 4, will be proposed.

2. Classical ruin model

The Cramér-Lundberg's classical risk model has its origin in Filip-Lundberg's doctoral thesis in 1903. In this work, Lundbery studied the collective reinsurance problem and used compounded homogenous Poisson process. In 1930, Harald Cramér re-examined Lundberg's original ideas and formalize them in the stochastic processes context.

The original model is:

$$X(t) = X(0) + ct - \sum_{n=1}^{N_t} Y_n \tag{1}$$

with $c > 0, X(0) \geq 0$ and $X(0)$ Being the initial capital, c the premium density, which is assumed to be constant, Y_j the amount of the j-th claim and N_t is an homogeneous Poisson process which represents the number of claims up to time t (independent of the interval position and the history of the process). Claims Y_j are supposed to positive independent random variables which are independent of the process N_t, with distribution F such that $F(0) = 0$ and whose mean μ is finite.

If the arrival of the n-th claim is denoted by S_n, then:

$$N_t = \sup\left\{n \geq 1 : S_n \leq t\right\}, \quad t \geq 0$$

***Note* 1.** The number of claims that have occurred up to time t can be approximated, in the Cramér and Lundberg's model, from other distribution functions.

The intervals between claims $T_k = S_k - S_{k-1}, k = 2, 3, \dots$ are independent and identically distributed random variables with an exponential distribution with parameter λ and finite mean and $T_1 = S_1$.

The aggregate claims until instant t are given by the random variable

$$S(t) = \sum_{n=1}^{N_t} Y_n$$

known as compound Poisson. Its distribution is followed by

$$G_t(x) = P\left[S(t) \leq x\right] = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} F^{(n)}(x)$$

with $x, t \geq 0$ and $F^{(n)}$ the n-th convolution of F with $F^{(0)}$ the distribution function of the measure of Dirac in 0.

The time to ruin is defined as:

$$T = \inf\left\{t > 0 : X(t) \leq 0\right\} \tag{2}$$

where $\inf \emptyset = \infty$.

The probability of ruin in the interval $[0, t]$ or probability of ruin in a finite horizon is defined as:

$$\psi(u, t) = P[T \leq t | X(0) = u] \tag{3}$$

and the probability of ruin in an infinite horizon or simply probability of ruin is:

$$\psi(u) = \lim_{t \to \infty} \psi(u, t) = \lim_{t \to \infty} P[T \leq t | X(0) = u] = P[T < \infty | X(0) = u] \tag{4}$$

***Note* 2.** In this case, the probability in an infinite horizon is usually approximated by the Normal-Power.

Definition 1. The basic Cramér-Lundberg's process is described as

$$X(t) = X(0) + (1 + \upsilon)\lambda\mu t - S(t)$$

where $\lambda\mu t = E\left[S(t)\right]$ and $\upsilon = \frac{c}{\lambda\mu} - 1 > 0$ is referred to as "solvency or safety margin", in order to guarantee survival (defined as the set of free capital whose purpose is to address those risks that may threaten the solvency of the company, the latter being the capacity to face obligations).

3. Preliminaries

In this section we introduce notation that is used throughout the paper and we set up some definitions. The introduced definitions are general and they can be found in several texts, for example in Müller & Stoyan (2002), Shaked (2007) or Almaraz (2009) among others.

We let the following sets $\mathbb{N} = \{0, 1, 2, \ldots\}$, $\mathbb{N}_+ = \{1, 2, \ldots\}$ and $\mathbb{R} = (-\infty, \infty)$.

Definition 2. Given two random variables X and Y taking values in a countable ordered state space I, then Y is stochastically smaller than X in the usual sense, and it is denoted as $Y \leq_{st} X$, if $P(Y \leq i) \geq P(X \leq i)$ for all $i \in I$.

Definition 3. A subset U of $\mathbb{R}^n$ is regarded to be as increasing if $y \in U$ when $y \geq x$ and $x \in U$.

Definition 4. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random vectors such that $P[\boldsymbol{X} \in U] \leq P[\boldsymbol{Y} \in U]$ for all the increasing subsets $U \subseteq \mathbb{R}^n$. Then $\boldsymbol{X}$ is stochastically smaller than $\boldsymbol{Y}$ in the usual sense and it is denoted as $\boldsymbol{X} \leq_{st} \boldsymbol{Y}$.

Definition 5. Let $X = \{X(t), t \in T\}$ and $Y = \{Y(t), t \in T\}$, be two stochastic processes with state space $I \subseteq \mathbb{R}$ and time parameter space T (usually $T = [0, \infty)$ or $T = \mathbb{N}_+$). Suppose that, for all choices of an integer m and $t_1 < t_2 < \cdots < t_m$ en T, it happens that:

$$(X(t_1), X(t_2), \ldots, X(t_m)) \leq_{st} (Y(t_1), Y(t_2), \ldots, Y(t_m))$$

Then $X = \{X(t), t \in T\}$ is said to be stochastically smaller than $Y = \{Y(t), t \in T\}$ in the usual sense and it is denoted as $X = \{X(t), t \in T\} \leq_{st} Y = \{Y(t), t \in T\}$.

Definition 6. A finite measure matrix is a matrix with non-negative entries whose lines are finite measure vectors.

Definition 7. Let I and J be two countable ordered sets and let $A = (a_{ij})_{i \in I, j \in J}$ and $B = (b_{ij})_{i \in I, j \in J}$ be two finite measure matrix with common indices on $I \times J$. Then the matrix A is said to be smaller than B in the Kalmykov sense, and it is denoted as $A \leq_K B$, if and only if:

$$\sum_{m \geq n} a_{im} \leq \sum_{m \geq n} b_{jm}, \forall i \leq j \quad \forall n$$

Also the following concepts will be necessary.

Definition 8. The counting process $N = (N_t)_t$ is an homogeneous Poisson process with rate $\lambda > 0$ if:

1. $N_0 = 0$, almost sure.

2. N has independent stationary increments.

3. $\forall \quad 0 \leq s < t < \infty$, $N_t - N_s \sim P(\lambda(t-s))$, that is,

$$P\left[N_t - N_s = k\right] = e^{-\lambda(t-s)}\frac{(\lambda(t-s))^k}{k!}, \quad k \in \mathbb{N}$$

Definition 9. (Markov process (MP)). A stochastic process $\{X_t, t \in T\}$, is said to be a Markov process (or Markovian process) if

$$P[X_{t_{n+1}} = x_{n+1} \mid X_{t_1} = x_1, X_{t_2} = x_2, \dots, X_{t_n} = x_n] =$$

$$= P[X_{t_{n+1}} = x_{n+1} \mid X_{t_n} = x_n]$$

for each $n \in \mathbb{N}$ y $t_1 < t_2 < \cdots < t_n < t_{n+1}$.

This condition is known as the Markovian condition.

A Markovian process with finite state space is known as the Markov Chain and it can be in discrete time (DTMC) or continuous time (CTMC).

Definition 10. (Markovian Renewal process (MRP)). A bivariate process $(Z,S) = (Z_n, S_n)_{n\in\mathbb{N}}$ is a Markovian Renewal process with phase states (countable) I and kernel $\boldsymbol{Q} = (\boldsymbol{Q}(t))_{t\in\mathbb{R}_+}$ where $\boldsymbol{Q}(t) = (Q_{ij}(t))_{i,j\in I}$ is a family of sub-distribution functions such that $\sum_{j\in I} Q_{ij}(t)$ is a distribution function, for each $i \in I$, if it is a Markov process in $I \times \mathbb{R}_+$ such that $S_0 = 0$ and

$$Q_{ij}(t) = P[Z_{n+1} = j, S_{n+1} - S_n \leq t \mid Z_n = i, S_n = s]$$

for each $n \in \mathbb{N}, i, j \in I$ and $s, t \in \mathbb{R}_+$

Definition 11. (Semi-Markovian process (SMP)) A process $W = (W_t)_{t\in\mathbb{R}_+}$ is a semi-Markovian process with state space I and kernel $\boldsymbol{Q}$ (or admitting an embedded kernel $(\boldsymbol{P}, \boldsymbol{F})$) if

$$W_t = Z_n, \quad S_n \leq t < S_{n+1}$$

for some MRP (Z,S) with phase space I and kernel $\boldsymbol{Q}$ (embedded kernel $(\boldsymbol{P}, \boldsymbol{F})$)

4. Stochastic dominance of ruin times in semi-Markov modulated risk processes

Let us consider the following generalization of the classic model:

$$X(t) = X(0) + \int_0^t c_{J(s)} ds - \sum_{n=1}^{N_t} Y_n \tag{5}$$

where $c_j > 0$ for all j, and $X(0) \geq 0$.

Where $X(0)$ is a random variable that represents the initial capital; $J(s)$ a semi-Markovian process; c_j the Premium density when the process $J(s)$ is in the state j; Y_n the size of the n-th claim and N_t a counting process associated to J that represents the number of claims up to time t.

Let (S_n, K_n) a Markovian sequence associated to the process J, where

$$S_n = \inf\{t \geq 0 : N_t \geq n\}, n \in \mathbb{N}$$

represents a sequence of events and

$$K_n = J(S_n), n \in \mathbb{N}$$

is an irreductible and discrete Markov chain with state space I, a countable subset of $\mathbb{R}$, transition matrix $P = (P_{ij})_{i,j\in I}$ and representing the state visited in the n-th tansistion, where

$$J_t = K_n, S_n \leq t < S_{n+1}$$

Let H_n be the time between the $(n-1)$-th and the n-th claim:

$$H_n = S_n - S_{n-1}, n \geq 1 \tag{6}$$

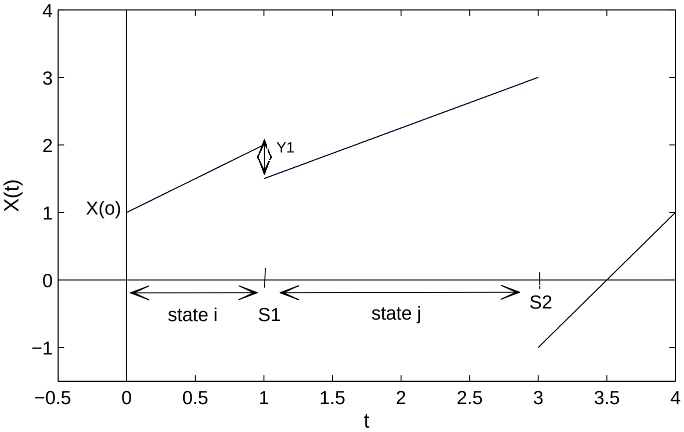

Figure 1: Path of the process $X(t)$.

In classical literature, this graphical representation (Figure 1) is known as a surplus process of ruin or process (see Bowers, Gerber, Hickman, Jones & Nesbitt 1997).

In this way, the process may be written as:

$$X(t) = X(0) + \sum_{n=0}^{N_t-1} c_{K_n} H_{n+1} + c_{K_{N_t}}(t - S_{N_t}) - \sum_{n=1}^{N_t} Y_n \tag{7}$$

where $\quad c_j > 0, \quad \forall j \in I.$

The semi-Markovian dependence structure under consideration is of the following type:

$$P\left[H_{n+1} \leq x, Y_{n+1} \leq y, K_{n+1} = j | K_n = i, (H_r, Y_r, K_r), 0 \leq r \leq n\right] =$$

$$= P\left[H_1 \leq x, Y_1 \leq y, K_1 = j | K_0 = i\right] = Q_{ij}(x, y) \tag{8}$$

The sequence $Q = (Q_{ij})_{i,j \in I}$ is the kernel of this process.

The purpose of this paper is to establish sufficient conditions for the first order stochastic dominance between the times of ruin of two processes like described in (5).

4.1. Stochastic processes in which the amount of the claims depends on the environment

Let us consider the following structure of the kernel:

$$Q_{ij}(x, y) = p_{ij} F_{ij}(x) G_{ij}(y) \tag{9}$$

where

- $p_{ij} = Q_{ij}(\infty, \infty) = P[K_{n+1} = j | K_n = i], n \in \mathbb{N}_+, i, j \in I$
- F_{ij} is the distribution function of $H_n | (K_{n-1} = i, K_n = j), n \in \mathbb{N}_+, i, j \in I$
- G_{ij} is the distribution function of $Y_n | (K_n = i, K_{n+1} = j), n \in \mathbb{N}_+, i, j \in I$

This implies that $(Y_1, Y_2, \ldots)$ and $(H_1, H_2, \ldots)$ are conditionally independent given $(K_0, K_1, \ldots)$ That is, the are conditionally independent given the evolution of the process J.

The parametrization of this process is (c, P, F, G), where $c = (c_i)_{i \in I}$, $P = (p_{ij})_{i,j \in I}$, $F = (F_{ij})_{i,j \in I}$ y $G = (G_{ij})_{i,j \in I}$.

We will denote the time to ruin of this process by:

$$T_{ab}^{(l)} = \inf\left\{t > 0 : X^{(l)}(t) \leq 0\right\} \mid \left(X^{(l)}(0) = b, K_0^{(l)} = a\right)$$

Theorem 1. *Let $X^{(1)} = (X^{(1)}(t))_{t \geq 0}$ and $X^{(2)} = (X^{(2)}(t))_{t \geq 0}$ be two stochastic processes with parametrizations $(c^{(1)}, P^{(1)}, F^{(1)}, G^{(1)})$ and $(c^{(2)}, P^{(2)}, F^{(2)}, G^{(2)})$ respectively, as described in (5) and (9). Let $J^{(1)}(0) \leq J^{(2)}(0)$ and $X^{(1)}(0) \leq X^{(2)}(0)$.*

If

$$c_i^{(1)} \leq c_k^{(2)}, \quad \forall i \leq k \tag{10}$$

$$P^{(1)} \leq_K P^{(2)} \tag{11}$$

$$F_{ij}^{(1)} \leq_{st} F_{kl}^{(2)}, \quad \forall i \leq k, j \leq l \tag{12}$$

$$G_{kl}^{(2)} \leq_{st} G_{ij}^{(1)}, \quad \forall i \leq k, j \leq l \tag{13}$$

then $T_{iu}^{(1)} \leq_{st} T_{jv}^{(2)} \quad \forall i \leq j, u \leq v.$

***Proof*.** Let $X^{(1)}$ and $X^{(2)}$ as in the statement. We must prove that $T_{iu}^{(1)} \leq_{st} T_{jv}^{(2)}$ for each $i \leq j, u \leq v$.

The Markovian renewal sequence associated to $J^{(l)}$, $l = 1, 2$ will be denoted as $\left(S_n^{(l)}, K_n^{(l)}\right)$.

It is defined:

$$T_{iu}^{*(l)} = \inf\left\{S_n^{(l)} : X_{S_n^{(l)}}^{(l)} \leq 0\right\} |(X_{S_0^{(l)}}^{(l)} = u, K_0^{(l)} = i), \quad l = 1, 2. \tag{14}$$

Note that the fact of $X^{(l)}$, $l = 1, 2$, being a non-decreasing sequence in $\left[S_n^{(l)}, S_{n+1}^{(l)}\right)$, denotes that:

$$T_{iu}^{(l)} = T_{iu}^{*(l)} \tag{15}$$

so it is enough to prove that $T_{iu}^{*(1)} \leq_{st} T_{jv}^{*(2)} \quad \forall u \leq v, i \leq j$.

Let

$$\left(\widetilde{X}_n^{(1)}, \widetilde{X}_n^{(2)}\right)$$

be a couple of

$$X_{S_n^{(1)}}^{(1)} | X_{S_0^{(1)}}^{(1)} = u \quad and \quad X_{S_n^{(2)}}^{(2)} | X_{S_0^{(2)}}^{(2)} = v$$

on a common product probability space

$$\Lambda = \Lambda_1 \times \Lambda_2 = (\Omega, \mathcal{F}, P) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$$

such that

$$\widetilde{X}_n^{(1)}(\omega) \leq \widetilde{X}_n^{(2)}(\omega), \forall \omega \in \Omega$$

and

$$\widetilde{S}_n^1(\omega) \leq \widetilde{S}_n^2(\omega), \forall \omega \in \Omega$$

being $\widetilde{S}_n^{(l)}$, a copy of the process $S_n^{(l)}$, $l = 1, 2$.

To do that, the following independent sequences of independent uniform random variables on the interval $(0, 1)$ will be used: $(U_n)_{n \in \mathbb{N}_+}$ in Λ_1, $(V_n)_{n \in \mathbb{N}_+}$ and $(W_n)_{n \in \mathbb{N}_+}$ in Λ_2.

Let $\widetilde{K}_0^{(1)}(\omega_1) = i$ and $\widetilde{K}_0^{(2)}(\omega_1) = j$.

In detail, for $l = 1, 2$:

$$\widetilde{K}_n^{(l)}(\omega_1) = \left[P_{\widetilde{K}_{n-1}^{(l)}, \cdot}^{(l)}\right]^{-1} (U_n(\omega_1))), \quad n \in \mathbb{N}_+, \omega_1 \in \Omega_1 \tag{16}$$

$$\widetilde{H}_n^{(l)}(\omega) = \left[F^{(l)}_{(\widetilde{K}_{n-1}^{(l)}(\omega_1),\widetilde{K}_n^{(l)}(\omega_1))}\right]^{-1}(V_n(\omega_2)), \quad n \in \mathbb{N}_+, \omega = (\omega_1, \omega_2) \in \Omega \tag{17}$$

$$\widetilde{Y}_n^{(l)}(\omega) = \left[G^{(l)}_{(\widetilde{K}_{n-1}^{(l)}(\omega_1),\widetilde{K}_n^{(l)}(\omega_1))}\right]^{-1}(W_n(\omega_2)), \quad n \in \mathbb{N}, \omega = (\omega_1, \omega_2) \in \Omega \tag{18}$$

Let $\widetilde{X}_0^{(1)} = u$ and $\widetilde{X}_0^{(2)} = v$ and:

$$\widetilde{X}_n^{(l)}(\omega) = \widetilde{X}_0^{(l)} + \sum_{m=0}^{n-1} c^{(l)}_{\widetilde{K}_m^{(l)}} \widetilde{H}_{m+1}^{(l)}(\omega) - \sum_{m=1}^{n} \widetilde{Y}_m^{(l)}(\omega), \tag{19}$$

$n \in \mathbb{N}$, $l = 1, 2$, $\omega = (\omega_1, \omega_2) \in \Omega$, be the embedded Markov Process of $(X^{(l)}_{S_n^{(l)}})_{n\geq 0}$, $l = 1, 2$.

Using (11), (12) and (13) we have respectively by construction:

$$\widetilde{K}_n^{(1)}(\omega_1) \leq \widetilde{K}_n^{(2)}(\omega_1), \quad \forall \omega_1 \in \Omega_1, n \in \mathbb{N} \tag{20}$$

$$\widetilde{H}_n^{(1)}(\omega) \leq \widetilde{H}_n^{(2)}(\omega), \quad \forall \omega = (\omega_1, \omega_2) \in \Omega, n \in \mathbb{N} \tag{21}$$

and

$$\widetilde{Y}_n^{(1)}(\omega) \geq \widetilde{Y}_n^{(2)}(\omega), \quad \forall \omega = (\omega_1, \omega_2) \in \Omega, n \in \mathbb{N} \tag{22}$$

On the other hand, using (10), (21) and (22), we have that:

$$\sum_{m=1}^{n} \widetilde{Y}_m^{(1)}(\omega) \geq \sum_{m=1}^{n} \widetilde{Y}_m^{(2)}(\omega), \quad \forall \omega = (\omega_1, \omega_2) \in \Omega, n \in \mathbb{N} \tag{23}$$

$$\widetilde{S}_n^{(1)}(\omega) = \sum_{m=1}^{n} \widetilde{H}_m^{(1)}(\omega) \leq \sum_{m=1}^{n} \widetilde{H}_m^{(2)}(\omega) = \widetilde{S}_n^{(2)}(\omega), \quad \forall \omega = (\omega_1, \omega_2) \in \Omega, n \in \mathbb{N} \tag{24}$$

$$\sum_{m=0}^{n-1} c^{(1)}_{\widetilde{K}_m^{(1)}} \widetilde{H}_{m+1}^{(1)}(\omega) \leq \sum_{m=0}^{n-1} c^{(2)}_{\widetilde{K}_m^{(2)}} \widetilde{H}_{m+1}^{(2)}(\omega), \quad \forall \omega = (\omega_1, \omega_2) \in \Omega, n \in \mathbb{N} \tag{25}$$

thus, leading to

$$\widetilde{X}_n^{(1)}(\omega) \leq \widetilde{X}_n^{(2)}(\omega) \quad \forall \omega = (\omega_1, \omega_2) \in \Omega, n \in \mathbb{N} \tag{26}$$

If denoted by:

$$\widetilde{T}_{iu}^{*(l)} = \inf\left\{\widetilde{S}_n^{(l)} : \widetilde{X}_n^{(l)} \leq 0\right\} \mid (\widetilde{X}_0 = u, \widetilde{K}_0^{(l)} = i), \quad l = 1, 2 \tag{27}$$

we have that

$$\widetilde{T}_{iu}^{*(1)}(\omega) \leq \widetilde{T}_{jv}^{*(2)}(\omega), \quad \forall i \leq j, u \leq v, \omega \in \Omega \tag{28}$$

intended to prove. □

In the special case in which the environment J is Markovian, particulary a continuous time Markov chain (CTMC), the parametrization of the process (5) is (c, P, q, G), where $P = (p_{ij})_{i,j \in I}$, q is the vector of transition rates from states of J and $G = (G_{ij})_{i,j \in I}$. In this case, the previous theorem has an immediate application, it is enough to see that the condition $F^{(1)}_{ij} \leq_{st} F^{(2)}_{kl}, \quad \forall i \leq k, j \leq l$ is translated in a condition to the vectors of transition rates from the states $q^{(2)}_k \leq q^{(1)}_i \quad \forall i \leq k$, that is, the distribution function of $H^{(l)}_n | \left(K^{(l)}_{n-1} = i, K^{(l)}_n = j\right), l = 1, 2$, which was denoted as $F^{(l)}_{ij}$, $l = 1, 2$ has in this case the following expression: $F^{(l)}_{ij}(x) = q^{(l)}_i e^{-q^{(l)}_i x}$ for $l = 1, 2$.[1]

In this way, the following two processes will be considered $X^{(1)}$ and $X^{(2)}$ as describes in (5) with parametrizations $(c^{(1)}, P^{(1)}, q^{(1)}, G^{(1)})$ and $(c^{(2)}, P^{(2)}, q^{(2)}, G^{(2)})$ respectively and then, the result for this particular case is described in the next corollary.

Corollary 1. *Let $X^{(1)} = (X^{(1)}_t)_{t \geq 0}$ and $X^{(2)} = (X^{(2)}_t)_{t \geq 0}$ be two stochastic processes with parameterizations $(c^{(1)}, P^{(1)}, q^{(1)}, G^{(1)})$ and $(c^{(2)}, P^{(2)}, q^{(2)}, G^{(2)})$ respectively, as described in (5), with environments $J^{(1)}$ and $J^{(2)}$ beings CTMCs with state space I, embedded transition probability matrices $P^{(1)}$ and $P^{(2)}$ and vectors of transition rates from states $q^{(1)}$ and $q^{(2)}$, respectively.*

Let $J^{(1)}(0) \leq J^{(2)}(0)$, and $X^{(1)}(0) \leq X^{(2)}(0)$. If

$$c^{(1)}_i \leq c^{(2)}_k, \quad \forall i \leq k \tag{29}$$

$$P^{(1)} \leq_K P^{(2)} \tag{30}$$

$$q^{(2)}_k \leq q^{(1)}_i \quad \forall i \leq k \tag{31}$$

$$G^{(2)}_{kl} \leq_{st} G^{(1)}_{ij}, \quad \forall i \leq k, j \leq l \tag{32}$$

then $T^{(1)}_{iu} \leq_{st} T^{(2)}_{jv} \quad \forall i \leq j, u \leq v$.

Proof. It is a direct consequence of Theorem 1. □

For the particular case in which the processes have the same transition matrix P we may relax the conditions of the Theorem 1 to conditions involving only one pair of states (i, j) such that $p_{ij} > 0$, in the following way.

Corollary 2. *Let $X^{(1)} = (X^{(1)}_t)_{t \geq 0}$ and $X^{(2)} = (X^{(2)}_t)_{t \geq 0}$ be two stochastic processes with parameterizations $(c^{(1)}, P, F^{(1)}, G^{(1)})$ and $(c^{(2)}, P, F^{(2)}, G^{(2)})$ respectively, as described in (5) and (9). If*

$$c^{(1)}_j \leq c^{(2)}_j, \quad \forall j \tag{33}$$

[1] Exponential distribution with intensity $q^{(l)}_i$ and the distribution decreases stochastically in the usual sense with this intensity

and for each pair (i,j) such that $P_{ij} > 0$

$$F_{ij}^{(1)} \leq_{st} F_{ij}^{(2)} \tag{34}$$

$$G_{ij}^{(2)} \leq_{st} G_{ij}^{(1)} \tag{35}$$

then

$$T_{iu}^{(1)} \leq_{st} T_{iv}^{(2)} \quad \forall u \leq v.$$

***Proof*.** The result follows is derived from the construction used in the proof of Theorem 1 since this construction now leads to $\widetilde{K}_n^{(1)}(\omega_1) = \widetilde{K}_n^{(2)}(\omega_1)$, $\forall n \in \mathbb{N}, \omega_1 \in \Omega_1$. This fact allows to conclude (21) and (22) using (34) and (35) despite of (12) and (13). The rest of the proof is analogous. □

4.1.1. A simple application

In this section, a case from the counting process N_t, identified as a semi-Markovian process J whose state space is $\{0, 1, 2 \ldots\}$ and whose transition probability matrix P is deterministic, in which case the probability to go from state n to $n+1$ is 1, is studied.

Let $(S_n^{(l)})_{n\geq 0}$ be a stochastic process with

$$0 = S_0^{(l)} < S_1^{(l)} < \cdots$$

such that

$$H_n^{(l)} = S_n^{(l)} - S_{n-1}^{(l)}, \quad n \in \mathbb{N}_+, \quad l = 1, 2$$

are independent random variables with distribution function

$$F_n^{(l)}, l = 1, 2$$

and let

$$N_t^{(l)} = \sup\left\{n \geq 0 : S_n^{(l)} \leq t\right\}, t \geq 0, l = 1, 2$$

the counting process.

Let $(Y_j^{(l)}), j \in \mathbb{N}, l = 1, 2$ be a sequence of independent random variables with distribution function $(G_j^{(l)}), j \in \mathbb{N}_+, l = 1, 2$. Let $H_n^{(l)}, G_n^{(l)}$ be independent $\forall n \in \mathbb{N}_+$.

Let us consider the process $(X^{(l)}(t))_{t\geq 0}$ with parametrization $(c^{(l)}, F^{(l)}, G^{(l)})$ con $l = 1, 2$ defined as follows:

$$X^{(l)}(t) = X^{(l)}(0) + c^{(l)}t - \sum_{j=1}^{N_t^{(l)}} Y_j^{(l)} \tag{36}$$

with $c^{(l)} > 0, X^{(l)}(0) \geq 0$, and it is defined

$$T_u^{(l)} = \inf\left\{t \geq 0 : X^{(l)}(t) \leq 0\right\} | X^{(l)}(0) = u \tag{37}$$

Theorem 2. *Let $X^{(1)} = (X^{(1)}(t))_{t\geq 0}$ and $X^{(2)} = (X^{(2)}(t))_{t\geq 0}$ be two stachastic processes with parametrizations $(c^{(1)}, F^{(1)}, G^{(1)})$ and $(c^{(2)}, F^{(2)}, G^{(2)})$ as described in (36), with $X^{(1)}(0) \leq X^{(2)}(0)$. If*

$$c^{(1)} \leq c^{(2)} \tag{38}$$

$$F_n^{(1)} \leq_{st} F_n^{(2)}, \quad \forall n \in \mathbb{N}_+ \tag{39}$$

$$G_n^{(2)} \leq_{st} G_n^{(1)}, \quad \forall n \in \mathbb{N}_+ \tag{40}$$

then

$$T_u^{(1)} \leq_{st} T_v^{(2)}, \quad \forall u \leq v$$

***Proof*.** Let $X^{(1)}$ and $X^{(2)}$ as stated. It must be proved that $T_u^{(1)} \leq_{st} T_u^{(2)}$ for all $u \leq v$.

It is defined:

$$T_u^{*(l)} = \inf\left\{S_n^{(l)} : X_{S_n^{(l)}}^{(l)} \leq 0\right\} | X_{S_0^{(l)}}^{(l)} = u, \quad l = 1, 2 \tag{41}$$

Note that the fact of $X^{(l)}$, $l = 1, 2$, being a non-decreasing sequence in $\left[S_n^{(l)}, S_{n+1}^{(l)}\right)$, gives that:

$$T_u^{(l)} = T_u^{*(l)} \tag{42}$$

therefore is enough to prove that $T_u^{*(1)} \leq_{st} T_v^{*(2)} \quad \forall u \leq v$.

For that, couplings will be build

$$\left(\widetilde{S}_n^{(1)}, \widetilde{S}_n^{(2)}\right) \quad \text{and} \quad \left(\widetilde{X}_n^{(1)}, \widetilde{X}_n^{(2)}\right)$$

of

$$\left(S_n^{(1)}, S_n^{(2)}\right) \quad \text{and} \quad \left(X_{S_n^{(1)}}^{(1)}, X_{S_n^{(2)}}^{(2)}\right)$$

given $\left(X_{S_n^{(1)}}^{(1)}, X_{S_n^{(2)}}^{(2)}\right) = (u, v)$ such that

$$\widetilde{X}_n^{(1)}(\omega) \leq \widetilde{X}_n^{(2)}(\omega), \forall \omega \in \Omega, \quad n \in \mathbb{N}$$

and

$$\widetilde{S}_n^{(1)}(\omega) \leq \widetilde{S}_n^{(2)}(\omega), \forall \omega \in \Omega, \quad n \in \mathbb{N}$$

To do that, independent sequences of independent and identically distributed $U(0,1)$ random variables will be used $(U_n)_{n\in\mathbb{N}_+}$ and $(V_n)_{n\in\mathbb{N}_+}$, defined on a common probability space $(\Omega, \mathcal{F}, P)$.

Let for $\omega \in \Omega$ and $l = 1, 2$:

$$\widetilde{H}_n^{(l)}(\omega) = \left[F_n^{(l)}\right]^{-1}(U_n(\omega))), \quad n \in \mathbb{N}_+ \tag{43}$$

$$\widetilde{Y}_n^{(l)}(\omega) = \left[G_n^{(l)}\right]^{-1}(V_n(\omega))), \quad n \in \mathbb{N}_+ \tag{44}$$

Let for $\omega \in \Omega$ and $l = 1, 2$:

$$\widetilde{X}_n^{(l)}(\omega) = \widetilde{X}_0^{(l)}(\omega) + c^{(l)} \sum_{m=1}^{n} \widetilde{H}_m^{(l)}(\omega) - \sum_{m=1}^{n} \widetilde{Y}_m^{(l)}(\omega), \quad n \in \mathbb{N} \tag{45}$$

with $\widetilde{X}_0^{(1)}(\omega) = u$ and $\widetilde{X}_0^{(2)}(\omega) = v$, be the embedded Markov process of $(X_t^{(l)})_{t\geq 0}$, coupling of $X_{S_n^{(l)}}^{(l)} \quad l = 1, 2$.

Using (39) and (40) we have for construction that:

$$\widetilde{H}_n^{(1)}(\omega) \leq \widetilde{H}_n^{(2)}(\omega), \quad \forall \omega \in \Omega, n \in \mathbb{N} \tag{46}$$

$$\widetilde{Y}_n^{(1)}(\omega) \geq \widetilde{Y}_n^{(2)}(\omega), \quad \forall \omega \in \Omega, n \in \mathbb{N} \tag{47}$$

On the other hand, from (46) and (47):

$$\widetilde{S}_n^{(1)}(\omega) = \sum_{m=1}^{n} \widetilde{H}_m^{(1)}(\omega) \leq \sum_{m=1}^{n} \widetilde{H}_m^{(2)}(\omega) = \widetilde{S}_n^{(2)}(\omega), \quad \forall \omega \in \Omega, n \in \mathbb{N} \tag{48}$$

$$\sum_{m=1}^{n} \widetilde{Y}_m^{(1)}(\omega) \geq \sum_{m=1}^{n} \widetilde{Y}_m^{(2)}(\omega), \quad \forall \omega \in \Omega, n \in \mathbb{N} \tag{49}$$

which leads with condition (38) to:

$$\widetilde{X}_n^{(1)}(\omega) \leq \widetilde{X}_n^{(2)}(\omega) \quad \forall \omega \in \Omega, n \in \mathbb{N} \tag{50}$$

If we denote:

$$\widetilde{T}_u^{*(l)} = \inf \left\{ \widetilde{S}_n^{(l)} : \widetilde{X}_n^{(l)} \leq 0 \right\}, \quad l = 1, 2 \tag{51}$$

being $\widetilde{T}_u^{*(1)}$ and $\widetilde{T}_u^{*(2)}$ a coupling of $(T_u^{*(1)}, T_u^{*(2)})$ we have using (48) and (50) that

$$\widetilde{T}_u^{*(1)}(\omega) \leq \widetilde{T}_u^{*(2)}(\omega), \quad \forall \omega \in \Omega \tag{52}$$

as it was pretended. □

4.2. Comparisons of ruin probabilities

An algorithm which leads to simulate processes verifying the conditions of Theorem 1 will be described.

```
Input: Independent sequences of independent random variables $U(0,1)$: $(U_n)_{n\in\mathbb{N}_+}$, $(V_n)_{n\in\mathbb{N}_+}$, $(W_n)_{n\in\mathbb{N}_+}$. Values $x^{(1)}$ y $x^{(2)}$, $f^{(1)}$ y $f^{(2)}$ with $x^{(1)} \leq x^{(2)}$, $f^{(1)} \leq f^{(2)}$.
$\widetilde{X}_0^{(1)} = x^{(1)}$
$\widetilde{X}_0^{(2)} = x^{(2)}$
$\widetilde{K}_0^{(1)} = f^{(1)}$
$\widetilde{K}_0^{(2)} = f^{(2)}$
for $n = 0, ..., N$ do
  for $l = 1, 2$ do
    $\widetilde{K}_{n+1}^{(l)} = \left[P^{(l)}_{\widetilde{K}_n^{(l)},.}\right]^{-1}(U_{n+1})$
    $\widetilde{H}_{n+1}^{(l)} = \left[F^{(l)}_{(\widetilde{K}_n^{(l)},\widetilde{K}_{n+1}^{(l)})}\right]^{-1}(V_{n+1})$
    $\widetilde{Y}_{n+1}^{(l)} = \left[G^{(l)}_{(\widetilde{K}_n^{(l)},\widetilde{K}_{n+1}^{(l)})}\right]^{-1}(W_{n+1})$
    $\widetilde{X}_n^{(l)} = \widetilde{X}_0^{(l)} + \sum_{m=0}^{n-1} c^{(l)}_{\widetilde{K}_m^{(l)}} \widetilde{H}_{m+1}^{(l)} - \sum_{m=1}^{n} \widetilde{Y}_m^{(l)}$
  end for
end for
Output: Two sequences $\widetilde{X}^{(1)}$ and $\widetilde{X}^{(2)}$ such that
$T_{iu}^{(1)} \leq_{st} T_{jv}^{(2)} \quad \forall i \leq j, u \leq v$
```

FIGURE 2: Simulation of sequences of random variables as described in (5), under conditions of Theorem 1.

Next algorithm consists of showing a method which allows to estimate the difference between the ruin probabilities in a given period T, of two processes which satisfy conditions of Theorem 1, that is, $\psi^{(1)}(u,T) - \psi^{(2)}(u,T)$ is wanted to be estimated. For simplicity, $p^{(l)}$ will denote the ruin probability of the process l under the interval of consideration, so: $p^{(l)} = \psi^{(l)}(u,T)$, for $l = 1, 2$.

For the above purpose, M replicas of each process will be simulated. Let $X_r^{(l)}$ be, for $l = 1, 2$ and $r = 1, \dots, M$, the r-th replica of the process l. Let

$$T_r^{(l)} = \inf\left\{t \geq 0 : X_r^{(l)}(t) \leq 0\right\}$$

be the time to ruin of the r-th replica r of the process l, and $R_r^{(l)} = \mathbf{1}_{\left\{T_r^{(l)} \leq T\right\}}$ be a random variable which indicates if the process $X_r^{(l)}$ reaches ruin in the interval $[0, T]$.

The estimator of $p^{(l)}$, $l = 1, 2$, which will be denoted as $P^{(l)}$, is the proportion of replicas in which ruin has happened, that is:

$$P^{(l)} = \frac{\sum_{r=1}^{M} R_r^{(l)}}{M}$$

and the estimator for the difference of these probabilities is:

$$P = P^{(1)} - P^{(2)}$$

The method used in the proof of the Theorem 1 is based on independent simulations of random variables which gives a less variance for the estimator P:

$$Var(P) = \frac{P \cdot (1 - P)}{M}$$

in relation to an independent simulation of the some.

For the algorithm, a control variable $I_r^{(l)}$ is required, which has value 1 when the simulation must go on or 0 in other case; l denotes the process $l = 1, 2$ and $r = 1, \ldots, M$ the number of the replica.

For $l = 1, 2$, $P^{(l)}$ represents the proportion of replicas in which ruin occurs in process $X^{(l)}$ until time T. The number of replicas of the process $X^{(l)}$ with $l = 1, 2$ in which ruin happens up to time T, has a Binomial (Bi) distribution, that is, $M \cdot P^{(l)}$ is $Bi(M, p^{(l)})$. On the other hand, $R_r^{(1)} - R_r^{(2)}$ has a Bernoulli (Be) distribution $Be(p^{(1)} - p^{(2)})$ and so, $M \cdot P$ is $Bi(M, P)$.

As it was mentioned, the method used in Theorem 1 gives an estimator with less variance than the estimator obtained with independent simulations of $P^{(1)}$ and $P^{(2)}$. In fact, let

$$V_d = \frac{(p^{(1)} - p^{(2)})(1 - (p^{(1)} - p^{(2)}))}{M}$$

and

$$V_i = \frac{p^{(1)}(1 - p^{(1)}) + p^{(2)}(1 - p^{(2)})}{M}$$

the variances of the estimator in the case of dependent and independent simulations, respectively, and let

$$E = \frac{\sqrt{V_i} - \sqrt{V_d}}{\sqrt{V_i}}$$

The following table shows a numerical example of the reduction that is obtained by applying a dependent simulation method. In each entry a the table there the following three values are displayed: $\sqrt{M \cdot V_d}$, $\sqrt{M \cdot V_i}$ and E in percentage:

```
Input: Independent sequences of independent random variables identically distributed as $U(0,1)$: $(U_{r,n})_{n\in\mathbb{N}_+}$, $(V_{r,n})_{n\in\mathbb{N}_+}$
for $r = 1, \ldots, M$ do
  $(W_{r,n})_{n\in\mathbb{N}_+}$, $r = 1, \ldots, M$. Values $T$, $x$, $f^{(1)}$ and $f^{(2)}$ with $f^{(1)} \leq f^{(2)}$.
  $\widetilde{X}^{(1)}_{r,0} = x$, $\widetilde{X}^{(2)}_{r,0} = x$
  $\widetilde{K}^{(1)}_{r,0} = f^{(1)}$, $\widetilde{K}^{(2)}_{r,0} = f^{(2)}$
  $I^{(1)}_r = 1$, $I^{(2)}_r = 1$, $n_1 = 0$, $n_2 = 0$
  while $\max\left\{I^{(1)}_r, I^{(2)}_r\right\} = 1$ do
    for $l = 1, 2$ do
      if $I^{(l)}_r = 1$ then
        $\widetilde{K}^{(l)}_{r,n_l+1} = \left[P^{(l)}_{\widetilde{K}^{l}_{r,n_l},\cdot}\right]^{-1}(U_{r,n_l+1})$
        $\widetilde{H}^{(l)}_{r,n_l+1} = \left[F^{(l)}_{(\widetilde{K}^{(l)}_{r,n_l},\widetilde{K}^{(l)}_{r,n_l+1})}\right]^{-1}(V_{r,n_l+1})$
        $\widetilde{Y}^{(l)}_{r,n_l+1} = \left[G^{(l)}_{(\widetilde{K}^{(l)}_{r,n_l},\widetilde{K}^{(l)}_{r,n_l+1})}\right]^{-1}(W_{r,n_l+1})$
        $\widetilde{S}^{(l)}_{r,n_l+1} = \widetilde{S}^{(l)}_{r,n_l} + \widetilde{H}^{(l)}_{r,n_l+1}$
        $\widetilde{X}^{(l)}_{r,n_l+1} = \widetilde{X}^{(l)}_{r,n_l} + c^{(l)}_{\widetilde{K}^{(l)}_{r,n_l}}\widetilde{H}^{(l)}_{r,n_l+1} - \widetilde{Y}^{(l)}_{r,n_l+1}$
        $n_l = n_l + 1$
      end if
      if $\widetilde{S}^{(l)}_{r,n_l} \leq T$ and $\widetilde{X}^{(l)}_{r,n_l} \leq 0$ then $R^{(l)}_r = 1$ end if
      if $\widetilde{S}^{(l)}_{r,n_l} > T$ or $R^{(l)}_r = 1$ then $I^{(l)}_r = 0$ end if
    end for
  end while
end for
for $l = 1, 2$ do
$P^{(l)} = \frac{\sum_{r=1}^{M} R^{(l)}_r}{M}$
end for
$P = P^{(1)} - P^{(2)}$
$\hat{V}_d = \frac{P*(1-P)}{M}$
Output: Estimator $P$ of the difference between ruin probabilities of the two processes under consideration and its approximate variance $\hat{V}_d$
```

Figure 3: Algorithm to estimate the difference of ruin probabilities during a given time T, of two processes which satisfy conditions of Theorem 1.

Table 1: Reduction obtained by applying the dependent simulation method.

$p^{(1)}$ $p^{(2)}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.000 0.424 100%	0.300 0.500 40.0%	0.400 0.5477 27.0%	0.458 0.5745 20.2%	0.490 0.5831 16.0%	0.500 0.5745 13.0%	0.490 0.5477 10.6%	0.458 0.500 8.3%	0.400 0.4243 5.7%
0.2	- - -	0.000 0.566 100%	0.300 0.608 50.7%	0.400 0.632 36.7%	0.458 0.640 28.4%	0.490 0.632 22.5%	0.500 0.608 17.8%	0.490 0.566 13.4%	0.458 0.500 8.3%
0.3	- - -	- - -	0.000 0.648 100%	0.300 0.671 55.3%	0.400 0.678 41.0%	0.458 0.671 31.7%	0.490 0.648 24.4%	0.500 0.6083 17.8%	0.490 0.548 10.6%
0.4	- - -	- - -	- - -	0.000 0.693 100%	0.300 0.700 57.1%	0.400 0.693 42.3%	0.458 0.671 31.7%	0.490 0.632 22.5%	0.500 0.575 13.0%
0.5	- - -	- - -	- - -	- - -	0.000 0.707 100%	0.300 0.700 57.1%	0.400 0.678 41.0%	0.458 0.640 28.4%	0.490 0.583 16.0%
0.6	- - -	- - -	- - -	- - -	- - -	0.000 0.693 100%	0.300 0.671 55.3%	0.400 0.632 36.7%	0.458 0.574 20.2%
0.7	- - -	- - -	- - -	- - -	- - -	- - -	0.000 0.648 100%	0.300 0.608 50.7%	0.400 0.548 27.0%
0.8	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.000 0.566 100%	0.300 0.500 40.0%
0.9	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.000 0.424 100%

As it can be seen, values of $\sqrt{V_i}$ are higher than the correspondents $\sqrt{V_d}$ obtained with dependent simulation as in the proof of Theorem 1. In the particular case in which $p^{(1)} = p^{(2)}$, this method gives a big reduction, because the described method present a value $V_d = 0$, while values in the independent case are strictly positive.

With this method confidence intervals with lower amplitude can be built:

$$IC(1-\alpha) = \left(P \pm \Phi(1-\alpha/2) \cdot \sqrt{\frac{P*(1-P)}{M}}\right)$$

5. Conclusions

The problem of ruin was addressed from a different perspective to the traditional. Instead of setting expressions or quotations for the ruin probability of a particular model for the selection of each other, times to ruin have been ranked. This will allow to make a selection without knowing explicitly the expression of the probability of ruin or an approximation thereof.

On the other hand, simulation algorithms have been proposed for these processes and statistical inference methods to estimate differences between the probability of ruin of the models have been considered.

This paper is a reference tool which can be used to determine the actual level of risk assumed by insurers (sufficiency of financial resources, reserves and capital).

The problems of the minimum solvency margin and the probability of survival of the reserves can be approached from the perspective proposed, since it allows to model stochastic processes at groups, taking into account those risks that may occur at the group level and not necessarily at the level of companies considered individually.

[Recibido: septiembre de 2010 — Aceptado: julio de 2011]

References

Almaraz, E. (2009), Cuestiones notables de ordenación estocástica en optimación financiera, Tesis de Doctorado, Universidad Complutense de Madrid, Facultad de Ciencias Matemáticas. Departamento de Estadística e Investigación Operativa, Madrid.

Asmussen, S. (1989), 'Risk theory in a Markovian environment', *Scandinavian Actuarial Journal* **2**.

Beard, R. E., Pentikäinen, T. & Pesonen, E. (1984), *Risk Theory*, Chapman and Hall, London.

Beekman, J. (1969), 'A ruin function approximation', *Transactions of Society of Actuaries* **21**.

Bowers, N. L., Gerber, H. U., Hickman, J. C., Jones, D. A. & Nesbitt, C. J. (1997), *Actuarial Mathematics (2ł.Ed)*, The Society of Actuaries, USA.

Daykin, C. D. (1994), *Risk Theory of Actuarie*, Chapman & Hall, New York.

De Vylder, F. (1996), *Advanced Risk Theory*, Université de Bruxelles, Brussels.

Ferreira, F. & Pacheco, A. (2005), 'Level crossing ordering of semi-Markov processes and Markov chains', *Journal of Applied Probability* **42**(4), 989–1002.

Ferreira, F. & Pacheco, A. (2007), 'Comparision of level-crossing times for Markov and semi-Markov processes', *Statistics and Probability Letters* **77**(7), 151–157.

Frees, E. (1986), 'Nonparametric Estimation of the Probability of Ruin', *ASTIN Bulletin* **16**.

Gerber, H. (1995), *Life Insurance Mathematics*, second edn, Springer, Heidelberg.

Goovaerts, M. (1990), *Effective Actuarial Methods*, Elsevier Science Publishers B.V.

Latorre, L. (1992), *Teoría del Riesgo y sus Aplicaciones a la Empresa Aseguradora*, Editorial Mapfre.

Müller, A. & Stoyan, D. (2002), *Comparison Methods for Stochastic Models and Risks*, John Wiley & Sons.

Ramsay, C. (1992), 'A Practical Algorithm for Approximating the Probability of Ruin', *Transactions of the Society of Actuaries* **XLIV**.

Reinhard, J. (1984), 'On a class of semi-Markov risk models obtained as a classical risk models in a Markovian environment', *ASTIN Bulletin* **14**.

Reinhard, J. & Snoussi, M. (2001), 'On the distribution of the surplus prior to ruin in a discrete semi-Markov risk model', *ASTIN Bulletin* **31**.

Reinhard, J. & Snoussi, M. (2002), 'The severity of run in a discrete semi-Markov risk model', *Stochastic Models* **18**(1), 85–107.

Seal, H. (1969), *Stochastic Theory of a Risk Business*, John Wiley & Sons.

Shaked, M. y Shanthikumar, G. (2007), *Stochastic Orders*, Springer Series in Statistics.

Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 497 a 512

Bivariate Generalization of the Kummer-Beta Distribution

Generalización Bivariada de la Distribución Kummer-Beta

Paula Andrea Bran-Cardona[1,a],
Johanna Marcela Orozco-Castañeda[2,b], Daya Krishna Nagar[2,c]

[1]Departamento de Matemáticas, Facultad de Ciencias, Universidad del Valle, Cali, Colombia

[2]Departamento de Matemáticas, Facultad de Ciencias Naturales y Exactas, Universidad de Antioquía, Medellín, Colombia

Abstract

In this article, we study several properties such as marginal and conditional distributions, joint moments, and mixture representation of the bivariate generalization of the Kummer-Beta distribution. To show the behavior of the density function, we give some graphs of the density for different values of the parameters. Finally, we derive the exact and approximate distribution of the product of two random variables which are distributed jointly as bivariate Kummer-Beta. The exact distribution of the product is derived as an infinite series involving Gauss hypergeometric function, whereas the beta distribution has been used as an approximate distribution. Further, to show the closeness of the approximation, we have compared the exact distribution and the approximate distribution by using several graphs. An application of the results derived in this article is provided to visibility data from Colombia.

***Key words*:** Beta distribution, Bivariate distribution, Dirichlet distribution, Hypergeometric function, Moments, Transformation.

Resumen

En este artículo, definimos la función de densidad de la generalización bivariada de la distribución Kummer-Beta. Estudiamos algunas de sus propiedades y casos particulares, así como las distribuciones marginales y condicionales. Para ilustrar el comportamiento de la función de densidad, mostramos algunos gráficos para diferentes valores de los parámetros. Finalmente, encontramos la distribución del producto de dos variables cuya distribución conjunta es Kummer-Beta bivariada y utilizamos la distribución beta como

[a]Lecturer. E-mail: paula.bran@gmail.com
[b]Lecturer. E-mail: jmoc03@gmail.com
[c]Professor. E-mail: dayaknagar@yahoo.com

una aproximación. Además, con el fin de comparar la distribución exacta y la aproximada de este producto, mostramos algunos gráficos. Se presenta una aplicación a datos climáticos sobre niebla y neblina de Colombia.

***Palabras clave*:** distribución Beta, distribución bivariada, distribución Dirichlet, función hipergeométrica, momentos, transformación.

1. Introduction

The beta random variable is often used for representing processes with natural lower and upper limits. For example, refer to Hahn & Shapiro (1967). Indeed, due to a rich variety of its density shapes, the beta distribution plays a vital role in statistical modeling. The beta distribution arises from a transformation of the F distribution and is typically used to model the distribution of order statistics. The beta distribution is useful for modeling random probabilities and proportions, particularly in the context of Bayesian analysis. Varying within $(0, 1)$ the standard beta is usually taken as the prior distribution for the proportion p and forms a conjugate family within the beta prior-Bernoulli sampling scheme. A natural univariate extension of the beta distribution is the Kummer-Beta distribution defined by the density function (Gupta, Cardeño & Nagar 2001, Nagar & Gupta 2002, Ng & Kotz 1995),

$$\frac{\Gamma(a+c)}{\Gamma(a)\Gamma(c)} \frac{x^{a-1}(1-x)^{c-1}\exp(-\lambda x)}{{}_1F_1(a;a+c;-\lambda)} \tag{1}$$

where $a > 0, c > 0$, $0 < x < 1$, $-\infty < \lambda < \infty$ and ${}_1F_1$ is the confluent hypergeometric function defined by the integral (Luke 1969),

$$\begin{gathered} {}_1F_1(a;c;z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_0^1 t^{a-1}(1-t)^{c-a-1}\exp(zt)\,\mathrm{d}t, \\ \operatorname{Re}(c) > \operatorname{Re}(a) > 0 \end{gathered} \tag{2}$$

The Kummer-Beta distribution can be seen as bimodal extension of the Beta distribution (on a finite interval) and thus can help to describe real world phenomena possessing bimodal characteristics and varying within two finite bounds. The Kummer-Beta distribution is used in common value auctions where posterior distribution of “value of a single good” is Kummer-Beta (Gordy 1998). Recently, Nagar & Zarrazola (2005) derived distributions of product and ratio of two independent random variables when at least one of them is Kummer-Beta.

The random variables X and Y are said to have a bivariate Kummer-Beta distribution, denoted by $(X, Y) \sim KB(a, b; c; \lambda)$, if their joint density is given by

$$f(x, y; a, b; c; \lambda) = C(a, b; c; \lambda)x^{a-1}y^{b-1}\left(1-x-y\right)^{c-1}\exp[-\lambda(x+y)] \tag{3}$$

where $x > 0$, $y > 0$, $x + y < 1$, $a > 0, b > 0$, $c > 0$, $-\infty < \lambda < \infty$ and

$$C(a, b; c; \lambda) = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)}\{{}_1F_1(a+b;a+b+c;-\lambda)\}^{-1} \tag{4}$$

For $\lambda = 0$, the density (3) slides to a Dirichlet density with parameters a, b and c. In Bayesian analysis, the Dirichlet distribution is used as a conjugate prior distribution for the parameters of a multinomial distribution. However, the Dirichlet family is not sufficiently rich in scope to represent many important distributional assumptions, because the Dirichlet distribution has few number of parameters. We provide a generalization of the Dirichlet distribution with added number of parameters. Several other bivariate generalizations of Beta distribution are available in Mardia (1970), Barry, Castillo & Sarabia (1999), Kotz, Balakrishnan & Johnson (2000), Balakrishnan & Lai (2009), Hutchinson & Lai (1991), Nadarajah & Kotz (2005), and Gupta & Wong (1985).

The matrix variate generalization of Beta and Dirichlet distributions have been defined and studied extensively. For example, see Gupta & Nagar (2000).

It can also be observed that bivariate generalization of the Kummer-Beta distribution defined by the density (3), belongs to the Liouville family of distributions proposed by Marshall & Olkin (1979) and Sivazlian (1981), (also see Gupta & Song (1996), Gupta & Richards (2001) and Song & Gupta (1997)).

In this article we study several properties such as marginal and conditional distributions, joint moments, correlation, and mixture representation of the bivariate Kummer-Beta distribution defined by the density (3). We also derive the exact and approximate distribution of the product XY where $(X, Y) \sim KB(a, b; c; \lambda)$. Finally, an application of the results derived in this article is provided to visibility data about fog and mist from Colombia.

2. Properties

In this section we study several properties of the bivariate Kummer-Beta distribution defined in Section 1.

Using the Kummers relation,

$$ {}_1F_1(a; c; -z) = \exp(-z){}_1F_1(c - a; c; z) \tag{5}$$

the density given in (3) can be rewritten as

$$ C(a, b; c; \lambda) \exp(-\lambda) x^{a-1} y^{b-1} (1 - x - y)^{c-1} \exp[\lambda(1 - x - y)] \tag{6}$$

Expanding $\exp[\lambda(1 - x - y)]$ in power series and rearranging certain factors, the joint density of X and Y can also be expressed as

$$ \{{}_1F_1(c; a + b + c; \lambda)\}^{-1} \sum_{j=0}^{\infty} \frac{\Gamma(a + b + c)\Gamma(c + j)}{\Gamma(a + b + c + j)\Gamma(c)} \frac{\lambda^j}{j!} \frac{x^{a-1} y^{b-1} (1 - x - y)^{c+j-1}}{B(a, b, c + j)} $$

where

$$ B(\alpha, \beta, \gamma) = \frac{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}{\Gamma(\alpha + \beta + \gamma)} $$

Thus the bivariate Kummer-Beta distribution is an infinite mixture of Dirichlet distributions.

In Bayesian probability theory, if the posterior distributions are in the same family as the prior probability distribution, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior. In case of multinomial distribution, the usual conjugate prior is the Dirichlet distribution. If

$$P(r,s,f|x,y)=\binom{r+s+f}{r,s,f}x^r y^s(1-x-y)^f$$

and

$$p(x,y)=C(a,b;c;\lambda)x^{a-1}y^{b-1}\left(1-x-y\right)^{c-1}\exp[-\lambda(x+y)]$$

where $x>0$, $y>0$, and $x+y<1$, then

$$\begin{aligned}p(x,y\mid r,s,f)&=C(a+r,b+s;c+f;\lambda)\\&\qquad\times x^{a+r-1}y^{b+s-1}\left(1-x-y\right)^{c+f-1}\exp[-\lambda(x+y)]\end{aligned}$$

Thus, the bivariate family of distributions considered in this article is the conjugate prior for the multinomial distribution.

A distribution is said to be negatively likelihood ratio dependent if the density $f(x,y)$ satisfies

$$f(x_1,y_1)f(x_2,y_2)\leq f(x_1,y_2)f(x_2,y_1)$$

for all $x_1>x_2$ and $y_1>y_2$ (see Lehmann (1966)). In the case of bivariate generalization of the Kummer-Beta distribution the above inequality reduces to

$$(1-x_1-y_1)(1-x_2-y_2)<(1-x_1-y_2)(1-x_2-y_1)$$

which clearly holds. Hence, the bivariate distribution defined by the density (3) is negatively likelihood ratio dependent.

If $(X,Y)\sim KB(a,b;c;\lambda)$, then Ng & Kotz (1995) have shown that $Y/(X+Y)$ and $X+Y$ are mutually independent, $Y/(X+Y)\sim B(b,a)$ and $X+Y\sim KB(a+b;c;\lambda)$. Here we give a different proof of this result based on angular transformation.

Theorem 1. *Let* $(X,Y)\sim KB(a,b;c;\lambda)$ *and define* $X=R^2\cos^2\Theta$ *and* $Y=R^2\sin^2\Theta$*. Then,* R^2 *and* Θ *are independent,* $R^2\sim KB(a+b;c;\lambda)$ *and* $\sin^2\Theta\sim B(b,a)$*.*

***Proof*.** Using the transformation $X=R^2\cos^2\Theta$ and $Y=R^2\sin^2\Theta$ with the Jacobian $J(x,y\to r^2,\theta)=2r^2\cos\theta\sin\theta$, in the joint density of X and Y, we obtain the joint density of R and Θ as

$$C(a,b;c;\lambda)(r^2)^{a+b}(1-r^2)^{c-1}\exp(-\lambda r^2)(\cos\theta)^{2a-1}(\sin\theta)^{2b-1},\tag{7}$$

where $0<r^2<1$ and $0<\theta<\pi/2$. From (7), it is clear that R^2 and Θ are independent. Now, transforming $S=R^2$ and $U=\sin^2\Theta$ with the Jacobian $J(r^2,\theta\to s,u)=J(r^2\to s)J(\theta\to u)=(4s)^{-1}[u(1-u)]^{-1/2}$, above we get the desired result. □

We derive marginal and conditional distributions as follows.

Theorem 2. *If $(X, Y) \sim KB(a, b; c; \lambda)$, then the marginal density of X is given by*

$$C_1(a, b; c; \lambda) \exp(-\lambda x) x^{a-1} (1-x)^{b+c-1} {}_1F_1(b; b+c; -\lambda(1-x)) \tag{8}$$

where $0 < x < 1$ and

$$C_1(a, b; c; \lambda) = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b+c)} \{{}_1F_1(a+b; a+b+c; -\lambda)\}^{-1}$$

Proof. To find the marginal pdf of X, we integrate (3) with respect to y to get

$$C(a, b; c; \lambda) \exp(-\lambda x) x^{a-1} \int_0^{1-x} \exp(-\lambda y) y^{b-1} (1-x-y)^{c-1} \, \mathrm{d}y$$

Substituting $z = y/(1-x)$ with $\mathrm{d}y = (1-x)\,\mathrm{d}z$ above, one obtains

$$C(a, b; c; \lambda) x^{a-1} \exp(-\lambda x) (1-x)^{b+c-1} \int_0^1 \exp[-\lambda(1-x)z] z^{b-1} (1-z)^{c-1} \, \mathrm{d}z \tag{9}$$

Now, the desired result is obtained by using (2). □

Using the above theorem, the conditional density function of X given $Y = y$, $0 < y < 1$, is obtained as

$$\frac{\Gamma(a+c)}{\Gamma(a)\Gamma(c)} \frac{\exp(-\lambda x) x^{a-1} (1-x-y)^{c-1}}{(1-y)^{a+c-1} {}_1F_1(a; a+c; -\lambda(1-y))}, \quad 0 < x < 1-y$$

Graphs 1–6 of the density function for several values of a, b, c and λ corresponding to six rows of Table 1, depicted in Figure 1, show a wide range of densities. For example, large values of a, b, c give a density similar to a bivariate normal density, whereas for small values of a, b, c the density is close to a uniform density.

Table 1: Density functions for different values of a, b, c and λ.

Graph	a	b	c	λ
1	2	1	1.5	−5.0
2	2	2	5.0	−5.0
3	5	3	2.0	−5.0
4	2	1	2.0	−0.5
5	5	3	9.0	0.5
6	3	2	1.5	3.0

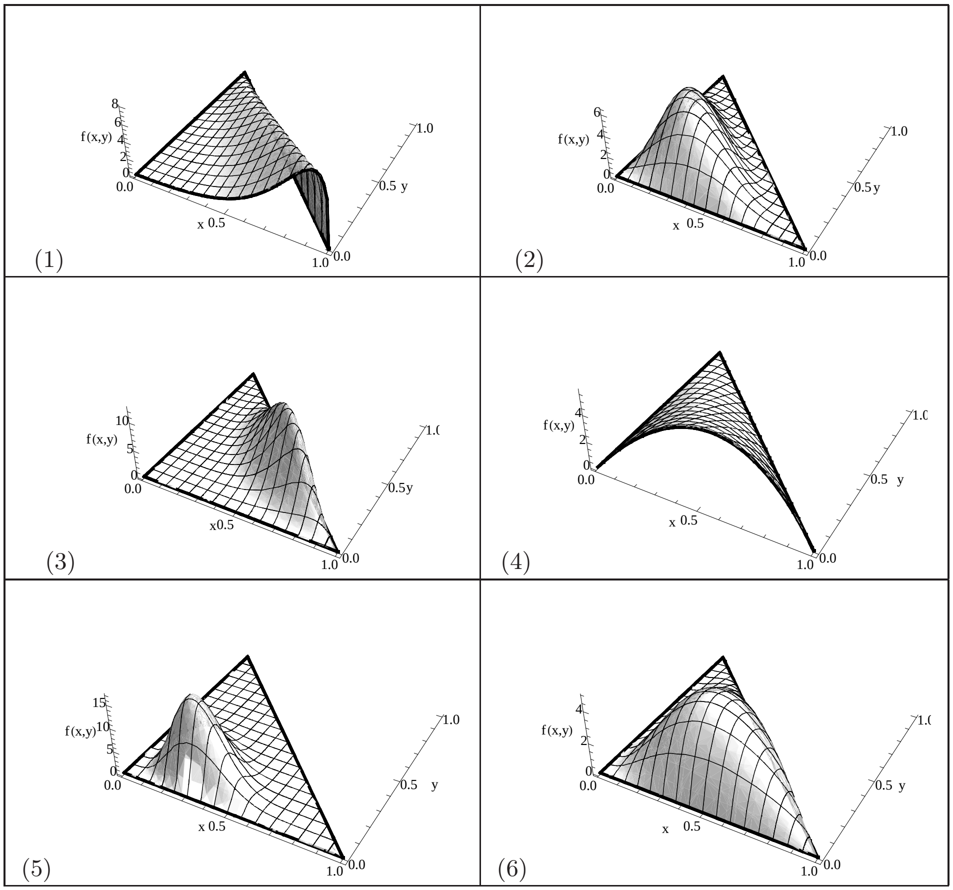

Figure 1: Density functions for different values of the parameters.

Further, using (3), the joint (r,s)-th moment is obtained as

$$\begin{aligned}\mathrm{E}(X^rY^s) &= C(a,b;c;\lambda)\int_0^1\int_0^{1-x}\exp[-\lambda(x+y)]x^{a+r-1}y^{b+s-1}\left(1-x-y\right)^{c-1}\,\mathrm{d}y\,\mathrm{d}x\\ &= \frac{C(a,b;c;\lambda)}{C(a+r,b+r;c;\lambda)}\\ &= \frac{\Gamma(a+r)\Gamma(b+s)\Gamma(d)}{\Gamma(a)\Gamma(b)\Gamma(d+r+s)}\frac{{}_1F_1(a+b+r+s;d+r+s;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}\end{aligned}$$

where $d=a+b+c$, $a+r>0$ and $b+s>0$. Now, substituting appropriately, we obtain

$$\mathrm{E}(X) = \frac{a}{d}\frac{{}_1F_1(a+b+1;d+1;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}$$

$$\mathrm{E}(Y) = \frac{b}{d}\frac{{}_1F_1(a+b+1;d+1;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}$$

$$\mathrm{E}(X^2)=\frac{a(a+1)}{d(d+1)}\frac{{}_1F_1(a+b+2;d+2;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}$$

$$\mathrm{E}(Y^2)=\frac{b(b+1)}{d(d+1)}\frac{{}_1F_1(a+b+2;d+2;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}$$

$$\mathrm{E}(XY)=\frac{ab}{d(d+1)}\frac{{}_1F_1(a+b+2;d+2;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}$$

$$\mathrm{E}(X^2Y^2)=\frac{ab(a+1)(b+1)}{d(d+1)(d+2)(d+3)}\frac{{}_1F_1(a+b+4;d+4;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}$$

$$\mathrm{Var}(X)=\frac{a}{d}\left[\frac{a+1}{d+1}\frac{{}_1F_1(a+b+2;d+2;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}-\frac{a}{d}\left\{\frac{{}_1F_1(a+b+1;d+1;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}\right\}^2\right]$$

$$\mathrm{Var}(Y)=\frac{b}{d}\left[\frac{b+1}{d+1}\frac{{}_1F_1(a+b+2;d+2;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}-\frac{b}{d}\left\{\frac{{}_1F_1(a+b+1;d+1;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}\right\}^2\right]$$

and

$$\mathrm{Cov}(X,Y)=\frac{ab}{d}\left[\frac{{}_1F_1(a+b+2;d+2;-\lambda)}{(d+1){}_1F_1(a+b;d;-\lambda)}-\frac{1}{d}\left\{\frac{{}_1F_1(a+b+1;d+1;-\lambda)}{{}_1F_1(a+b;d;-\lambda)}\right\}^2\right]$$

Notice that $\mathrm{E}(XY)$, $\mathrm{E}(X^2)$, $\mathrm{E}(Y^2)$, $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ involve ${}_1F_1(\alpha;\mu;-\lambda)$ which can be computed using Mathematica by providing values of α,μ and λ. Table 2 provides correlations between X and Y for different values of a,b,c and λ. All the tabulated values of correlation are negative because X and Y satisfy $x+y<1$. As can be seen, the choices of a,b small and c,λ large yield correlations close to zero, whereas large values of a or b and small values of c or λ give small correlations. Further, for fixed values of a,b and c, the correlation decreases as the value of λ increases. Likewise, for fixed values of a,b and λ, the correlation decreases as c increases.

3. Entropies

In this section, exact forms of Renyi and Shannon entropies are determined for the bivariate Kummer-Beta distribution defined in this article.

Let $(\mathcal{X},\mathcal{B},\mathcal{P})$ be a probability space. Consider a pdf f associated with $\mathcal{P}$, dominated by $\sigma-$finite measure μ on $\mathcal{X}$. Denote by $H_{SH}(f)$ the well-known Shannon entropy introduced in Shannon (1948). It is define by

$$H_{SH}(f)=-\int_{\mathcal{X}}f(x)\log f(x)\,d\mu \tag{10}$$

Table 2: Correlation for values of a, b, c and λ.

a	b	c	$\lambda = -5.000$	-2.000	-1.000	-0.500	0.000	0.500	1.000	2.000	5.000
3.0	2.0	0.5	−0.936	−0.888	−0.862	−0.846	−0.828	−0.808	−0.785	−0.731	−0.494
1.0	2.0	1.0	−0.848	−0.717	−0.653	−0.616	−0.577	−0.536	−0.493	−0.406	−0.172
3.0	2.0	1.5	−0.819	−0.716	−0.670	−0.644	−0.617	−0.589	−0.559	−0.497	−0.304
5.0	3.0	2.0	−0.799	−0.723	−0.690	−0.673	−0.655	−0.635	−0.616	−0.573	−0.433
0.5	1.0	1.5	−0.736	−0.499	−0.406	−0.360	−0.316	−0.275	−0.237	−0.171	−0.055
1.0	2.0	2.0	−0.712	−0.543	−0.477	−0.442	−0.408	−0.374	−0.341	−0.279	−0.135
0.5	1.0	2.0	−0.654	−0.414	−0.332	−0.294	−0.258	−0.225	−0.195	−0.144	−0.054
1.0	2.0	3.0	−0.598	−0.429	−0.371	−0.343	−0.316	−0.290	−0.265	−0.219	−0.118
2.0	4.0	5.0	−0.535	−0.428	−0.391	−0.374	−0.356	−0.339	−0.322	−0.290	−0.204
2.0	2.0	5.0	−0.494	−0.365	−0.324	−0.305	−0.286	−0.267	−0.250	−0.218	−0.141
1.0	0.5	5.0	−0.322	−0.185	−0.151	−0.136	−0.123	−0.111	−0.100	−0.082	−0.046

One of the main extensions of the Shannon entropy was defined by Rényi (1961). This generalized entropy measure is given by

$$H_R(\eta, f) = \frac{\log G(\eta)}{1-\eta} \qquad (\text{for } \eta > 0 \text{ and } \eta \neq 1) \tag{11}$$

where

$$G(\eta) = \int_{\mathcal{X}} f^{\eta} d\mu$$

The additional parameter η is used to describe complex behavior in probability models and the associated process under study. Rényi entropy is monotonically decreasing in η, while Shannon entropy (10) is obtained from (11) for $\eta \uparrow 1$. For details see Nadarajah & Zografos (2005), Zografos and Nadarajah (2005) and Zografos (1999).

First, we give the following lemma useful in deriving these entropies.

Lemma 1. *Let* $g(a, b, c, \lambda) = \lim_{\eta \to 1} h(\eta)$, *where*

$$h(\eta) = \frac{\mathrm{d}}{\mathrm{d}\eta} {}_1F_1(\eta(a+b-2)+2; \eta(a+b+c-3)+3; -\lambda\eta) \tag{12}$$

Then,

$$\begin{aligned} g(a,b,c,\lambda) = \sum_{j=1}^{\infty} \frac{\Gamma(a+b+j)\Gamma(a+b+c)}{\Gamma(a+b)\Gamma(a+b+c+j)} \frac{(-\lambda)^j}{j!} \Big[& j + (a+b-2)\psi(a+b+j) \\ & + (a+b+c-3)\psi(a+b+c) - (a+b-2)\psi(a+b) \\ & - (a+b+c-3)\psi(a+b+c+j) \Big] \end{aligned} \tag{13}$$

where $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ *is the digamma function.*

Proof. Expanding ${}_1F_1$ in series form, we write

$$h(\eta) = \frac{\mathrm{d}}{\mathrm{d}\eta}\sum_{j=0}^{\infty}\Delta_j(\eta)\frac{(-\lambda)^j}{j!} = \sum_{j=0}^{\infty}\left[\frac{\mathrm{d}}{\mathrm{d}\eta}\Delta_j(\eta)\right],\frac{(-\lambda)^j}{j!} \tag{14}$$

where

$$\Delta_j(\eta) = \frac{\Gamma[\eta(a+b-2)+2+j]\Gamma[\eta(a+b+c-3)+3]}{\Gamma[\eta(a+b-2)+2]\Gamma[\eta(a+b+c-3)+3+j]}\eta^j$$

Now, differentiating the logarithm of $\Delta_j(\eta)$ w.r.t. to η, one obtains

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\eta}\Delta_j(\eta) \quad = \quad & \Delta_j(\eta)\Big[\frac{j}{\eta} + (a+b-2)\psi(\eta(a+b-2)+2+j) \\ & +(a+b+c-3)\psi(\eta(a+b+c-3)+3) \\ & -(a+b-2)\psi(\eta(a+b-2)+2) \\ & -(a+b+c-3)\psi(\eta(a+b+c-3)+3+j)\Big]\end{aligned} \tag{15}$$

Finally, substituting (15) in (14) and taking $\eta \to 1$, one obtains the desired result. □

Theorem 3. *For the bivariate Kummer-Beta distribution defined by the pdf (3), the Rényi and the Shannon entropies are given by*

$$\begin{aligned}H_R(\eta, f) \quad = \quad & \frac{1}{1-\eta}\Bigg[\eta\log C(a,b;c;\lambda) + \log\Gamma[\eta(a-1)+1] \\ & +\log\Gamma[\eta(b-1)+1] + \log\Gamma[\eta(c-1)+1] \\ & -\log\Gamma[\eta(a+b+c-3)+3] \\ & +\log {}_1F_1(\eta(a+b-2)+2;\eta(a+b+c-3)+3;-\lambda\eta)\Bigg]\end{aligned} \tag{16}$$

and

$$\begin{aligned}H_{SH}(f) \quad = \quad & -\log C(a,b;c;\lambda) - [(a-1)\psi(a) + (b-1)\psi(b) + (c-1)\psi(c) \\ & -(a+b+c-3)\psi(a+b+c)] - \frac{g(a,b,c,\lambda)}{{}_1F_1(a+b;a+b+c;-\lambda)},\end{aligned} \tag{17}$$

respectively, where $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function and $g(a,b,c,\lambda)$ is given by (13).

Proof . For $\eta > 0$ and $\eta \neq 1$, using the joint density of X and Y given by (3), we have

$$\begin{aligned}
G(\eta) &= \int_0^1 \int_0^{1-x} f^\eta(x, y; a, b; c; \lambda)\, \mathrm{d}x\, \mathrm{d}y \\
&= [C(a, b; c; \lambda)]^\eta \int_0^1 \int_0^{1-x} x^{\eta(a-1)} y^{\eta(b-1)} \\
&\quad (1 - x - y)^{\eta(c-1)} \exp[-\eta\lambda(x + y)]\, \mathrm{d}x\, \mathrm{d}y \\
&= \frac{[C(a, b; c; \lambda)]^\eta}{C(\eta(a - 1) + 1, \eta(b - 1) + 1; \eta(c - 1) + 1; \lambda)} \\
&= \frac{\Gamma^\eta(a + b + c)\Gamma[\eta(a - 1) + 1]\Gamma[\eta(b - 1) + 1]\Gamma[\eta(c - 1) + 1]}{\Gamma^\eta(a)\Gamma^\eta(b)\Gamma^\eta(c)\Gamma[\eta(a + b + c - 3) + 3]} \\
&\quad \times \frac{{}_1F_1(\eta(a + b - 2) + 2; \eta(a + b + c - 3) + 3; -\lambda\eta)}{\{{}_1F_1(a + b; a + b + c; -\lambda)\}^\eta},
\end{aligned}$$

where the last line has been obtained by using (4). Now, taking logarithm of $G(\eta)$ and using (11) we get (16). The Shannon entropy is obtained from (16) by taking $\eta \uparrow 1$ and using L'Hopital's rule. □

4. Exact and Approximate Distribution of the Product

If $(X, Y) \sim KB(a, b; c; \lambda)$, then Ng & Kotz (1995) have shown that $X/(X+Y)$ and $X+Y$ are mutually independent, $X/(X+Y) \sim B(a, b)$ and $X+Y \sim KB(a+b; c; \lambda)$. In this section we derive the density of XY when $(X, Y) \sim KB(a, b; c; \lambda)$. The distribution of XY, where X and Y are independent random variables, $X \sim KB(a_1, b_1, \lambda_1)$ and $Y \sim KB(a_2, b_2, \lambda_2)$ has been derived in Nagar & Zarrazola (2005). In order to derive the density of the product we essentially need the integral representation of the Gauss hypergeometric function given by Luke (1969),

$$\begin{aligned}
{}_2F_1(a, b; c; z) &= \frac{\Gamma(c)}{\Gamma(a)\Gamma(c - a)} \int_0^1 t^{a-1}(1 - t)^{c-a-1}(1 - zt)^{-b}\, \mathrm{d}t, \\
&\qquad \mathrm{Re}(c) > \mathrm{Re}(a) > 0,\ |\arg(1 - z)| < \pi. \qquad (18)
\end{aligned}$$

Theorem 4. *If $(X, Y) \sim KB(a, b; c; \lambda)$, then the pdf of $W = XY$ is given by*

$$\begin{aligned}
&\frac{\sqrt{\pi}C(a, b; c; \lambda)\exp(-\lambda)}{2^{a+c-b-1}} \frac{w^{b-1}(1 - 4w)^{c-1/2}}{\left(1 + \sqrt{1 - 4w}\right)^{b+c-a}} \\
&\quad \times \sum_{i=0}^\infty \frac{\Gamma(c + i)}{\Gamma(c + 1/2 + i)\, 2^i\, i!} \left(\frac{1 - 4w}{1 + \sqrt{1 - 4w}}\right)^i \\
&\quad \times {}_2F_1\left(c + i, c + b - a + i; 2c + 2i; \frac{2\sqrt{1 - 4w}}{1 + \sqrt{1 - 4w}}\right),\ 0 < w < \frac{1}{4}. \qquad (19)
\end{aligned}$$

Proof. Making the transformation $W = XY$ with the Jacobian $J(x, y \rightarrow x, w) = x^{-1}$ in (3), we obtain the joint density of X and W as

$$C(a, b; c; \lambda) \exp(-\lambda) \frac{w^{b-1}(-x^2 + x - w)^{c-1}}{x^{b+c-a}} \exp\left[\frac{\lambda(-x^2 + x - w)}{x}\right]$$

where $p < x < q$ with

$$p = \frac{1 - \sqrt{1 - 4w}}{2}, \qquad q = \frac{1 + \sqrt{1 - 4w}}{2},$$

and $0 < w < 1/4$. Now, expanding $\exp\left[\lambda(-x^2 + x - w)/x\right]$ in power series and integrating x in the above expression, we obtain the marginal density of W as

$$\begin{aligned} &C(a, b; c; \lambda) \exp(-\lambda) w^{b-1} \int_p^q \frac{[(x - p)(q - x)]^{c-1}}{x^{b+c-a}} \exp\left(\frac{\lambda(x - p)(q - x)}{x}\right) \mathrm{d}x \\ &= C(a, b; c; \lambda) \exp(-\lambda) w^{b-1} \sum_{i=0}^{\infty} \frac{(q - p)^{2i+2c-1} \lambda^i}{q^{i+b+c-a}\, i!} \int_0^1 \frac{t^{c+i-1}(1 - t)^{c+i-1}\, \mathrm{d}t}{[1 - t\,(1 - p/q)]^{b+c-a+i}} \end{aligned}$$

where we have used the substitution $t = (q - x)/(q - p)$. Now, evaluating the above integral using (18) and simplifying the resulting expression, we get the desired result. □

In the rest of this section, we derive the approximate distribution of the product XY. It is clear from Theorem 4, that the random variable $4W = 4XY$ has support on $(0, 1)$. We, therefore, are motivated to use the Beta distribution of two parameters as an approximation to the exact distribution. Equating the first and the second moments of $4W$, with those of the Beta distribution with parameters α and β, it is easy to see that

$$\alpha = \frac{\mathrm{E}(W)[\mathrm{E}(W) - 4\mathrm{E}(W^2)]}{\mathrm{E}(W^2) - (\mathrm{E}(W))^2} \tag{20}$$

and

$$\beta = \frac{[\mathrm{E}(W) - 4\mathrm{E}(W^2)][1 - 4\mathrm{E}(W)]}{4[\mathrm{E}(W^2) - (\mathrm{E}(W))^2]}$$

The moments $\mathrm{E}(W)$ and $\mathrm{E}(W^2)$ are available in Section 2, and can be computed numerically for given values of a, b, c and λ. To demonstrate the closeness of the approximation we, in Figure 2, graphically compare the exact and approximated pdf of $4W$. First, for different values of the parameters (a, b, c, λ) we compute the corresponding estimates for (α, β), using (20) and (21). These estimates are given in Table 3, and corresponding graphics are given in Figure 2, showing comparison between exact and approximate densities. The exact pdf corresponds to the solid curve and approximate pdf corresponds to the broken curve. It is evident that the approximate density is quite close to the exact density.

Table 3: Estimated values of α and β.

Figure	a	b	c	λ	α	β
1	3.0	1.0	0.5	0.5	0.9567	1.0527
2	3.0	1.0	3.0	0.5	0.9514	3.7098
3	3.0	3.0	1.0	0.5	2.6239	1.5259
4	0.5	0.5	1.0	1.0	0.2646	1.8184
5	3.0	3.0	1.0	1.0	2.5250	1.5410
6	3.0	3.0	0.5	3.0	2.2502	1.0365

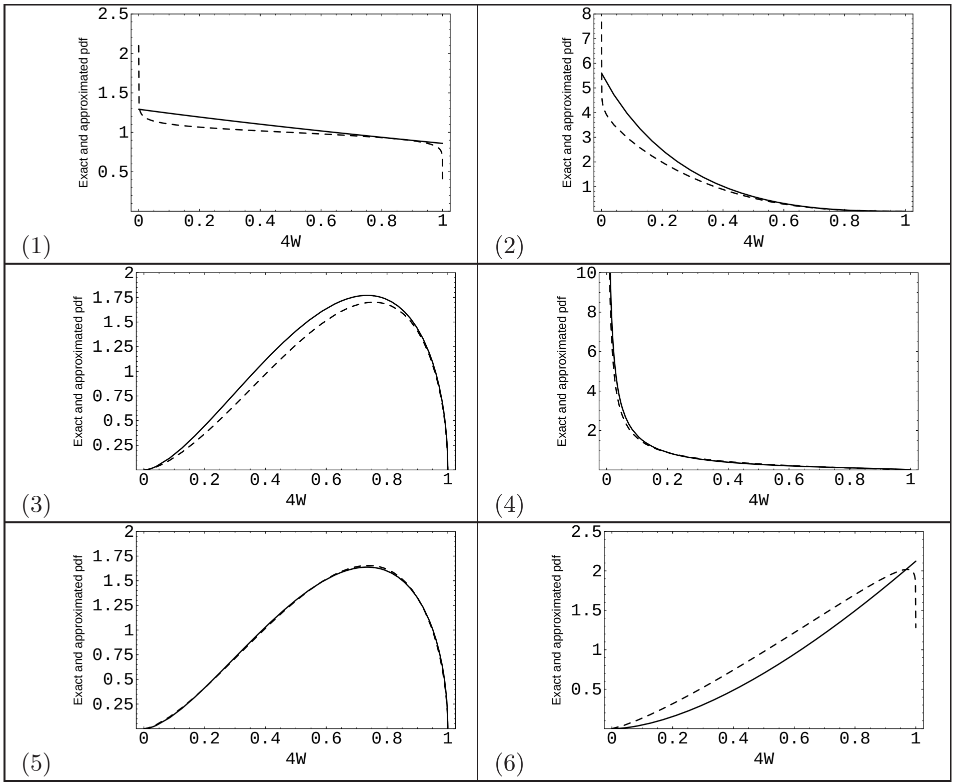

Figure 2: Graphics of the exact density function (solid curve) and the approximate (broken curve).

5. Application

In this section, we consider the data of fog and mist collect from five Colombian airports and present an application of the model given by (3).

Fog or mist is a collection of water droplets or ice crystals suspended in the air at or near the Earth's surface. The only difference between mist and fog is visibility. The phenomenon is called fog if the visibility is one kilometer or less; otherwise it is known as mist.

We consider data available at the website of IDEAM (Institute Hydrology, Meteorology and Environmental Studies, Colombia) collected from the following 5 major Colombian airports regarding the fog and mist:

- Ernesto Cortissoz Airport (Barranquilla)
- El Dorado Airport (Bogota)
- Alfonso Bonilla Aragón Airport (Cali)
- Rafael Núñez Airport (Cartagena)
- José María Córdova Airport (Medellin)

The data comprises average number of days each month in which mist or fog appeared during the period from 1975 to 1991. We consider the following variables:

X: the proportion of days with mist (the phenomenon weather provides a visibility of more than 1 km)

Y: proportion of days with fog (the phenomenon weather provides a visibility of 1 km or less)

In addition the following variables are of interest:

$X + Y$: proportion of days with the weather phenomenon (mist or fog)

$X/(X + Y)$: proportion of days with visibility greater than 1 km with respect to the total proportion of days exhibiting the phenomenon (mist or fog)

$Y/(X + Y)$: proportion of days with visibility less than 1 km with respect to the total proportion of days exhibiting the phenomenon (mist or fog)

Table 4, gives the estimates of a, b, c and λ, which were obtained using the maximum likelihood method, and by implementing Fisher scoring method (Kotz et al. (2000), p. 504). Table 5, gives estimated values of the moments $\mathrm{E}[X/(X+Y)]$, $\mathrm{E}[Y/(X+Y)]$ and $\mathrm{E}(X+Y)$ for five airports.

Table 4: Estimated values of a, b, c and λ.

Airport	a	b	c	λ
Barranquilla	0.620	0.266	153.00	−176.0
Bogota	8.290	3.370	3.82	12.3
Cali	0.303	0.088	70.80	−94.4
Cartagena	0.206	0.091	396.00	−407.0
Medellin	12.300	6.580	3.41	18.5

6. Conclusions of the Application

As conclusions, we can say that the proportion of days with visibility less than 1 km with respect to the total number of days presenting the phenomenon is similar for Barranquilla, Bogota and Cartagena airports. This ratio is a little lower for

Table 5: Estimated values of the moments.

Airport	$\mathrm{E}[X/(X+Y)]$	$\mathrm{E}[Y/(X+Y)]$	$\mathrm{E}(X+Y)$
Barranquilla	0.700	0.300	0.129
Bogotá	0.711	0.289	0.572
Cali	0.775	0.225	0.221
Cartagena	0.695	0.305	0.023
Medellín	0.651	0.349	0.675

the Cali and Medellin airports, the value of this ratio is higher. For example, we can say that the airport at Barranquilla has 30% of total days (with phenomenon) with fog. For Medellin, this percentage corresponds to 34.9% and for Cali to 22.5%. The proportion of days with phenomenon (mist or fog) is higher for the Medellin airport followed by the Bogota airport. Cartagena airport presents the lower proportion.

[Recibido: agosto de 2010 — Aceptado: agosto de 2011]

References

Balakrishnan, N. & Lai, C. D. (2009), *Continuous Bivariate Distributions*, second edn, Springer.

Barry, C. A., Castillo, E. & Sarabia, J. M. (1999), *Conditional Specification of Statistical Models*, Springer Series in Statistics, Springer-Verlag, New York.

Gordy, M. (1998), 'Computationally convenient distributional assumptions for common-value auctions', *Computational Economics* **12**, 61–78.

Gupta, A. K., Cardeño, L. & Nagar, D. K. (2001), 'Matrix variate Kummer-Dirichlet distributions', *Journal of Applied Mathematics* **1**(3), 117–139.

Gupta, A. K. & Nagar, D. K. (2000), *Matrix Variate Distributions*, Vol. 104 of *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics*, Chapman & Hall/CRC, Boca Raton, FL.

Gupta, A. K. & Song, D. (1996), 'Generalized Liouville distribution', *Computers & Mathematics with Applications* **32**(2), 103–109.

Gupta, A. K. & Wong, C. F. (1985), 'On three and five parameter bivariate Beta distributions', *International Journal for Theoretical and Applied Statistics* **32**(2), 85–91.

Gupta, R. D. & Richards, D. S. P. (2001), 'The history of the Dirichlet and Liouville distributions', *International Statistical Review* **69**(3), 433–446.

Hahn, G. J. & Shapiro, S. S. (1967), *Statistical Models in Engineering*, John Wiley and Sons, New York.

Hutchinson, T. P. & Lai, C. D. (1991), *The Engineering Statistician's Guide To Continuous Bivariate Distributions*, Rumsby Scientific Publishing, Adelaide.

Kotz, S., Balakrishnan, N. & Johnson, N. L. (2000), *Continuous Multivariate Distributions. Vol. 1. Models and applications*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, second edn, Wiley-Interscience, New York.

Lehmann, E. L. (1966), 'Some concepts of dependence', *Annals of Mathematical Statistics* **37**, 1137–1153.

Luke, Y. L. (1969), *The Special Functions and their Approximations*, Vol. 53 of *Mathematics in Science and Engineering*, Academic Press, New York.

Mardia, K. V. (1970), *Families of Bivariate Distributions*, Hafner Publishing Co., Darien, Conn. Griffin's Statistical Monographs and Courses, No. 27.

Marshall, A. W. & Olkin, I. (1979), *Inequalities: Theory of Majorization and its Applications*, Vol. 143 of *Mathematics in Science and Engineering*, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York.

Nadarajah, S. & Kotz, S. (2005), 'Some bivariate Beta distributions', *A Journal of Theoretical and Applied Statistics* **39**(5), 457–466.

Nadarajah, S. & Zografos, K. (2005), 'Expressions for Rényi and Shannon entropies for bivariate distributions', *Information Sciences* **170**(2-4), 173–189. *http://dx.doi.org/10.1016/j.ins.2004.02.020

Nagar, D. K. & Gupta, A. K. (2002), 'Matrix-variate Kummer-Beta distribution', *Journal of the Australian Mathematical Society* **73**(1), 11–25.

Nagar, D. K. & Zarrazola, E. (2005), 'Distributions of the product and the quotient of independent Kummer-Beta variables', *Scientiae Mathematicae Japonicae* **61**(1), 109–117.

Ng, K. W. & Kotz, S. (1995), Kummer-Gamma and Kummer-Beta univariate and multivariate distributions, Technical Report 84, Department of Statistics, The University of Hong Kong, Hong Kong.

Rényi, A. (1961), On measures of entropy and information, *in* 'Procedings 4th Berkeley Symposium Mathematical Statistics and Probability', University of California Press, Berkeley, California, pp. 547–561.

Shannon, C. E. (1948), 'A mathematical theory of communication', *The Bell System Technical Journal* **27**, 379–423, 623–656.

Sivazlian, B. D. (1981), 'On a multivariate extension of the Gamma and Beta distributions', *SIAM Journal on Applied Mathematics* **41**(2), 205–209.

Song, D. & Gupta, A. K. (1997), 'Properties of generalized Liouville distributions', *Random Operators and Stochastic Equations* **5**(4), 337–348.

Zografos, K. (1999), 'On maximum entropy characterization of Pearson's type II and VII multivariate distributions', *Journal of Multivariate Analysis* **71**(1), 67–75.
*http://dx.doi.org/10.1006/jmva.1999.1824

Zografos, K. & Nadarajah, S. (2005), 'Expressions for Rényi and Shannon entropies for multivariate distributions', *Statistics & Probability Letters* **71**(1), 71–84.
*http://dx.doi.org/10.1016/j.spl.2004.10.023

Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 513 a 543

Estimating the Discounted Warranty Cost of a Minimally Repaired Coherent System

Estimación del costo de garantía descontado para un sistema coherente bajo reparo mínimo

Nelfi Gertrudis González[1,a], Vanderlei Bueno[2,b]

[1] Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia

[2] Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brasil

Abstract

A martingale estimator for the expected discounted warranty cost process of a minimally repaired coherent system under its component level observation is proposed. Its asymptotic properties are also presented using the Martingale Central Limit Theorem.

***Key words*:** Expected cost, martingale central limit theorem, reliability, repairable system, semimartingale, stochastic point process.

Resumen

En este trabajo modelamos los costos de garantía descontados para un sistema coherente reparado mínimamente a nivel de sus componentes y proponemos un estimador martingalas para el costo esperado para un período de garantía fijo, también probamos sus propiedades asintóticas mediante el Teorema del Limite Central para Martingalas.

***Palabras clave*:** confiabilidad, costo esperado, proceso puntual estocástico, semi-martingalas, sistema reparable, teorema de límite central para martingalas.

1. Introduction

Warranties for durable consumer products are common in the market place. Its primary role is to offer a post sale remedy for consumers when a product fails

[a] Associate professor. E-mail: ngonzale@unal.edu.co
[b] Associate professor. E-mail: bueno@ime.usp.br

to fulfill its intended performance during the warranty period and generally, they also limit the manufacturer's liability for out-of-warranty product failure.

Manufacturers offer many types of warranties which have become an important promotional tool for their products. A discussion about various issues related to warranty policies can be found in Murthy (1990), Blischke & Murthy (1992*a*), Blischke & Murthy (1992*b*), Blischke & Murthy (1992*c*), Mitra & Patankar (1993), Blischke & Murthy (1994), Blischke & Murthy (1996).

Although warranties are used by manufacturers as a competitive strategy to boost their market share, profitability and image, they may cost a substantial amount of money and, from a manufacturer's perspective, the cost of a warranty program should be analyzed and estimated accurately.

A discounted warranty cost policy incorporates the time and provides an adequate measure for warranties because, in general, warranty costs can be treated as random cash flows in the future. Warranty issuers do not have to spend all the money at the stage of warranty planning, instead, they can allocate it along the life cycle of warranted products. Another reason why one should consider the time value is for the purpose of determining the warranty reserve, a fund set up specifically to meet future warranty claims. It is well known that the present value of warranty liabilities or rebates to be paid in the future are less than the face value and it is desirable to determine the warranty reserve according to the present value of the total warranty liability. Related issues to discounted warranty costs and warranty reserves have been studied by Mamer (1969), Mamer (1987), Patankar & Mitra (1995) and Thomas (1989), from both manufacturer and customer's perspectives for single-component products, either repairable or nonrepairable.

More recently, Jain & Maheshwari (2006) proposed a hybrid warranty model for renewing pro-rata warranties assuming constant failure rate and constant products maintenance and replacement costs. They derive the expected total discounted warranty costs for different lifetime distributions and determine the optimal number and optimal period for preventive maintenance after the expiry of the warranty; Jack & Murthy (2007) consider the costs for extended warranties offered after a base warranty and investigate optimal pricing strategies and optimal maintenance/replacement strategies; Hong-Zhong, Zhie-Jie, Yanfeng, Yu & Liping (2008) consider the cash flows of warranty reserve costs during the product lifecycle and estimate the expected warranty cost for reparable and non reparable products with both, the free replacement warranty and the pro-rata warranty policy. They also consider the case where the item has a heterogeneous usage intensity over the lifecicle and its usage is intermitent; Chattopadhyay & Rahman (2008) study lifetime warranties where the warranty coverage period depends on the lifetime of the product, they develop lifetime warranty policies and models for predicting failures and estimating costs; Jung, Park & Park (2010) consider optimal system maintenance policies during the post warranty period under the renewing warranty policy with maintenance costs dependent on life cycle.

In practice, most products are composed of several components. If warranties are offered for each component separately, then warranty models for single-component products can be applied directly. However, sometimes warranty terms are defined upon an entire system. For such warranties, it is necessary to consider the system structure as well as the component level warranty service cost (Thomas 1989). Warranty analysis for multi-component systems based on the system structure has been addressed in a few papers: Ritchken (1986) provides an example of a two-component parallel system under a two-dimensional warranty; Chukova & Dimitrov (1996) derive the expected warranty cost for two-component series system and parallel system under a free-replacement warranty; Hussain & Murthy (1998) also discuss warranty cost estimation for parallel systems under the setting that uncertain quality of new products may be a concern for the design of warranty programs; Bai & Pham (2006) obtained the first two centered moments of the warranty cost of renewable full-services warranties for complex systems with series-parallel and parallel-series structures. A Markovian approach to the analysis of warranty cost for a three-component system can be found in Balachandran, Maschmeyer & Livingstone (1981); Ja, Kulkarni, Mitra & Patankar (2002) study the properties of the discounted warranty cost and total warranty program costs for non renewable warranty policy with non stationary processes.

There are many ways to model the impact of repair actions on system failure times. For complex systems, repair is often assumed to be minimal, which restores its failure rate. For a review about modeling failure and maintenance data from repairable systems, see Li & Shaked (2003) and Lindqvist (2006). For a generalization of minimal repair to heterogeneous populations, i.e., when the lifetime distribution is a mixture of distributions, see Finkelstein (2004). Nguyen & Murthy (1984) present a general warranty cost model for single-component repairable products considering as-good-as-new-repair, minimal repair and imperfect repair, but the value of time is not addressed. In Ja et al. (2002), several warranty reserve models for single-component products are derived for non stationary sale processes. Ja, Kulkarni, Mitra & Patankar (2001) analyze a warranty cost model on minimally repaired single-component systems with time dependent costs. Bai & Pham (2004) consider the free-repair warranty and the pro-rata warranty policies to derive some properties of a discounted warranty cost for a series system of repairable and independent components using a non homogeneous Poisson process. Recently, Duchesne & Marri (2009) consider, the same problem by analyzing some distributional properties (mean, variance, characteristic function) of the corresponding discounted warranty cost and using a general competing risk model to approach system reliability; Sheu & Yu (2005) propose a repair-replacement warranty policy which splits the warranty period into two intervals where only minimal repair can be undertaken and a middle interval in which no more than one replacement is allowed. Their model applies to products with bathtub failure rate considering random minimal repair costs. Other work about repair strategies, including imperfect and minimal repair, which consider their effects on warranty costs, can be found in Yun, Murthy & Jack (2008), Chien (2008), Yeo & Yuan (2009) and Samatliy-Pac & Taner (2009).

For a series system with components which do not have common failures, system failures coincide with component failures and warranty models for single-component products can be applied directly. In this paper, we consider a discounted warranty cost policy of a repairable coherent system under a minimal repair process on its component level. In this case, the system is set up as a series system with its components that survive to their critical levels, that is, the time from which a failure of a component would lead to system failure and, therefore, it is seen as a series system. We use the Martingale Central Limit Theorem to approximate the warranty cost distribution for a fixed warranty period of length w, and to estimate the warranty cost through the component failure/repair point processes.

In the Introduction of this paper we survey the recent developments in warranty models. In Section 2, we consider the dependent components lifetimes, as they appear in time through a filtration and use the martingale theory, a natural tool to consider the stochastic dependence and the increasing information in time. In Section 3, we consider independent copies of a coherent system, and its components, as given in Section 2 and develop a statistical model for the discounted warranty cost. Also, in this Section, we give an example. The paper is self contained but a mathematical basis of stochastic processes applied to reliability theory can be found in Aven & Jensen (1999). The extended proofs are in the Appendix.

2. The Warranty Discounted Cost Model of a Coherent System on its Component Level

We consider the vector $\mathbf{S} = (S_1, S_2, \ldots, S_m)$ representing component lifetimes of a coherent system, with lifetime T, which are positive random variables in a complete probability space $(\Omega, \mathcal{F}, P)$. The components can be dependent but simultaneous failures are ruled out, that is, for all i, j with $i \neq j$, $P(S_i = S_j) = 0$. We observe the system on its component level throughout a filtration, a family of sub σ-algebras of $\mathcal{F}$, $(\mathcal{F}_t)_{t\geq 0}$

$$\mathcal{F}_t = \sigma\{1_{\{T>s\}}, 1_{\{S_i>s\}} : s \leq t, 1 \leq i \leq m\},$$

which is increasing, right-continuous and complete. Clearly, the $S_i, 1 \leq i \leq n$ are $(P, \mathcal{F}_t)$-stopping time.

An extended and positive random variable τ is an $(P, \mathcal{F}_t)$-stopping time if, and only if, $\{\tau \leq t\} \in \Im_t$, for all $t \geq 0$; an $(P, \mathcal{F}_t)$-stopping time τ is called predictable if an increasing sequence $(\tau_n)_{n\geq 0}$ of $(P, \mathcal{F}_t)$-stopping time, $\tau_n < \tau$, exists such that $\lim_{n\to\infty} \tau_n = \tau$; an $(P, \mathcal{F}_t)$-stopping time τ is totally inaccessible if $P(\tau = \sigma < \infty) = 0$ for all predictable $(P, \mathcal{F}_t)$-stopping time σ.

In what follows, to simplify the notation, we assume that relations such as $\subset, =, \leq, <, \neq$ between random variables and measurable sets, respectively, always hold "P-almost surely", i.e., with probability one, which means that the term P-a.s., is suppressed.

2.1. Component Minimal Repair

For each i, $1 \leq i \leq m$, we consider the simple counting process $N_t^i = 1_{\{S_i \leq t\}}$, i.e., the counting process corresponding to the simple point process $(S_{i,n})_{n \geq 1}$ with $S_i = S_{i,1}$ and $S_{i,n} = \infty$, for $n \geq 2$. We use the Doob-Meyer decomposition

$$N_t^i = A_t^i + M_t^i, \qquad M^i \in \mathcal{M}_0^2, \qquad i = 1, \ldots, m, \tag{1}$$

where $\mathcal{M}_0^2$ represents the class of mean zero and square integrable $(P, \mathcal{F}_t)$-martingales which are right-continuous with left-hand limits. A_t^i is a unique nondecreasing right continuous $(P, \mathcal{F}_t)$-predictable process with $A_0^i = 0$, called the $(P, \mathcal{F}_t)$-compensator of N_t^i.

We assume that the component lifetime $S_i,\ 1 \leq i \leq m$ is a totally inaccessible $(P, \mathcal{F}_t)$-stopping time, which is a sufficient condition for the absolutely continuity of A_t^i. It follows that

$$A_t^i = \int_0^t 1_{\{S_i > s\}} \lambda^i(s) ds < \infty, \qquad i = 1, \ldots, m, \tag{2}$$

where $\lambda^i(t)$ is the $(P, \mathcal{F}_t)$-failure rate of component i, a deterministic function of t.

Initially, consider the minimal repair process of component i. If we do a minimal repair at each failure of component i, the corresponding minimal repair counting process in $(0, t]$ is a non homogeneous Poisson process $\widetilde{N}_t^i = \sum\limits_{n=1}^{\infty} 1_{\{S_{i,n} \leq t\}}$, with Doob-Meyer decomposition given by,

$$\widetilde{N}_t^i = \int_0^t \lambda^i(s) ds + \widetilde{M}_t^i, \quad \widetilde{M}^i \in \mathcal{M}_0^2, \tag{3}$$

and therefore the expected number of minimal repairs of component i is $E[\widetilde{N}_t^i] = \int_0^t \lambda^i(s)\, ds$.

Let $H_i(t)$ be a deterministic, continuous (predictable) bounded and integrable function in $(0, t]$, corresponding to the minimal repair discounted cost of component i at time t, such that $\int_0^t H_i(s) \lambda^i(s) ds < \infty, 0 \leq t < \infty$.

The minimal repair cost process of component i is $\hat{B}_t^i = \sum\limits_{j=1}^{\widetilde{N}_t^i} H_i(S_{ij}) = \int_0^t H_i(s) d\widetilde{N}_s^i$, where S_{ij} is the j-th minimal repair time of component i and $S_{i1} = S_i$.

Since $H_i(s)$ is predictable, the process $\int_0^t H_i(s) d\widetilde{M}_s^i$ is a mean zero and square integrable $(P, \mathcal{F}_t)$-martingale and therefore, the $(P, \mathcal{F}_t)$-compensator of $\widehat{B}_t^i$ is B_t^i which is given by

$$B_t^i = \int_0^t H_i(s) \lambda^i(s) ds < \infty, \quad \forall \ \ 0 \leq t < \infty \tag{4}$$

Barlow and Proschan (1981) define the system lifetime T

$$T = \Phi(\mathbf{S}) = \min_{1\leq j\leq k} \max_{i\in K_j} S_i$$

where $K_j, 1 \leq j \leq k$ are minimal cut sets, that is, a minimal set of components whose joint failure causes the system to fail. Aven & Jensen (1999) define the critical level of component i as the $(P, \mathcal{F}_t)$-stopping time $Y_i,\ 1 \leq i \leq m$ which describes the time when component i becomes critical for the system, i.e., the time from which the failure of component i leads to system failure. If either the system or component i fail before the latter becomes critical ($T \leq Y_i$ or $S_i \leq Y_i$) we assume that $Y_i = \infty$. Therefore, as in Aven & Jensen (1999) we can write

$$T = \min_{i:Y_i<\infty} S_i \tag{5}$$

Therefore, concerning the system minimal repairs at the component level, it is sufficient to consider the component minimal repairs after its critical levels. In what follows we consider the set $\mathscr{C}^i = \{\omega \in \Omega : S_i(\omega) > Y_i(\omega)\}$, where Y_i is the critical level of component i, and the minimal repair point process restricted to $\mathscr{C}^i$, that is, the process N_t^{i*}, defined as

$$N_t^{i*} = 1_{\mathscr{C}^i} N_t^i \tag{6}$$

which counts the failures of component i when it is critical, implying system failure.

Theorem 1. *(González 2009) The $(P, \mathcal{F}_t)-$compensator process of the indicator process $N_t^i = 1_{\{S_i\leq t\}}$ in $\mathscr{C}^i$ is*

$$A_t^{i*} = \int_{Y_i}^t 1_{\{S_i>s\}}\lambda^i(s)ds = \int_0^t 1_{\{S_i>s\}}1_{\{Y_i<s\}}\lambda^i(s)ds < \infty, \forall\ \ 0 \leq t < \infty \tag{7}$$

***Note* 1.** Note that

$$E[N_t^i|S_i > Y_i] = E[A_t^{i*}|S_i > Y_i] = E\Big[\int_{Y_i}^t 1_{\{S_i>s\}}\lambda^i(s)ds\Big|S_i > Y_i\Big] \tag{8}$$

From Theorem 1 the next Corollary follows easily.

Corollary 1. *Let $\widetilde{N}_t^i$ be the minimal repair counting process for the component i. Let $H_i(t)$ be a deterministic, continuous (predictable), bounded and integrable function in $[0, t]$, corresponding to the discounted warranty cost of component i at time t, such that $\int_0^t H_i(s)\lambda^i(s)ds < \infty, 0 \leq t < \infty$. In $\mathscr{C}^i$ we have*

i. The $(P, \mathcal{F}_t)$-compensator of $\widetilde{N}_t^i$ is the process

$$\widetilde{A}_t^{i*} = \int_{Y_i}^t \lambda^i(s)ds = \int_0^t 1_{\{Y_i<s\}}\lambda^i(s)ds < \infty, \quad \forall\ \ 0 \leq t < \infty \tag{9}$$

ii. The $(P, \mathcal{F}_t)$-compensator of the minimal repair cost process of component i, $\widehat{B}_t^i = \sum_{j=1}^{\widetilde{N}_t^i} H_i(S_{ij})$ is the process

$$B_t^{i*} = \int_{Y_i}^t H_i(s)\lambda^i(s)ds = \int_0^t 1_{\{Y_i < s\}} H_i(s)\lambda^i(s)ds < \infty, \forall\ 0 \leq t < \infty \quad (10)$$

***Note* 2.** For each $i = 1, \dots, m$ and $\omega \in \mathscr{C}^i$, the process $B_t^i(\omega)$ given in (4), is equal to the process $B_t^{i*}(\omega)$.

2.2. Coherent System Minimal Repair

Now we are going to define the minimal repair counting process and its corresponding coherent system cost process.

Let $N_t = 1_{\{T \leq t\}}$ be the system failure simple counting process and its $(P, \mathcal{F}_t)$-compensator process A_t, with decomposition

$$N_t = A_t + M_t, \qquad M \in \mathcal{M}_0^2 \quad (11)$$

and

$$A_t = \int_0^t 1_{\{T > s\}} \lambda_s \, ds < \infty \quad (12)$$

where the process $(\lambda_t)_{t \geq 0}$ is the coherent system $(P, \mathcal{F}_t)$-failure rate process.

Since we do not have simultaneous failures the system will failure at time t when the first critical component for the system at t^- failures at t.

Under the above conditions Arjas (1981) proves that the $(P, \mathcal{F}_t)$-compensator of N_t is

$$A_t = \sum_{i=1}^m \left[A_{t \wedge T}^i - A_{Y_i}^i \right]^+ \quad (13)$$

and from (2) and (13) we get

$$A_t = \sum_{i=1}^m \int_0^t 1_{\{S_i > s\}} 1_{\{Y_i < s < T\}} \lambda^i(s) \, ds = \int_0^t 1_{\{T > s\}} \sum_{i=1}^m 1_{\{Y_i < s\}} \lambda^i(s) \, ds \quad (14)$$

From compensator unicity, it becomes clear that the $(P, \mathcal{F}_t)$-failure rate process of system is given by

$$\lambda_t = \sum_{i=1}^m 1_{\{Y_i < t\}} \lambda^i(t) \quad (15)$$

If the system is minimally repaired on its component level, its $(P, \mathcal{F}_t)$-failure rate process λ_t is restored at its condition immediately before failure and therefore the critical component that fails at the system failure time is minimally repaired. Therefore, the number of minimal repairs of the system on its component level, is $\widetilde{N}_t = \sum\limits_{n=1}^{\infty} 1_{\{T_n \leq t\}}$, with Doob-Meyer decomposition given by

$$\widetilde{N}_t = \int_0^t \lambda_s ds + \widetilde{M}_t = \sum_{i=1}^{m} \int_0^t 1_{\{Y_i < s\}} \lambda^i(s) ds + \widetilde{M}_t. \qquad \widetilde{M} \in \mathcal{M}_0^2$$

Definition 1. For a fixed $\omega \in \Omega$ let $\mathscr{C}^{\Phi}(\omega) = \{i \in \{1, \ldots, m\} : S_i(\omega) > Y_i(\omega)\}$ be the set of components surviving its corresponding critical levels. For each $i = 1, \ldots, m$, let C^i be the indicator variable

$$C^i(\omega) = \begin{cases} 1 & \text{if } i \in \mathscr{C}^{\Phi}(\omega) \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

Then, the minimal repair counting process of the coherent system is

$$\widetilde{N}_t(\omega) = \sum_{i \in \mathscr{C}^{\Phi}(\omega)} \widetilde{N}_t^i(\omega) = \sum_{i=1}^{m} C^i(\omega) \widetilde{N}_t^i(\omega) \tag{17}$$

with corresponding cost process

$$\widehat{B}_t(\omega) = \sum_{i \in \mathscr{C}^{\Phi}(\omega)} \widehat{B}_t^i(\omega) = \sum_{i=1}^{m} C^i(\omega) \widehat{B}_t^i(\omega) \tag{18}$$

***Note* 3.** Note that $C^i(\omega) = 1 \Leftrightarrow \omega \in \mathscr{C}^i$ and in each realization $\omega \in \Omega$, the indicator variables $C^i(\omega), i = 1, \ldots, m$, are constant in $[0, t]$. Therefore, if $C^i(\omega) = 0$, then $\widehat{B}_s^i = 0, \forall\, 0 \leq s \leq t$. It means that in each realization of the system repair/failure process, we only observe the repair/cost processes of components which fail after their corresponding critical levels. Therefore, in each realization, the repair/cost process for the system with structure Φ is equivalent to the repair/cost process for a series system of components which are critical for the initial system in such realization.

2.3. Martingale Estimator of the Warranty Cost

In the following results and definitions, for each realization w, the minimal repair costs of a coherent system is the sum of the minimal repair costs of its critical components in a given realization ω.

Suppose

$$\sum_{i=1}^{m} \int_0^t H_i(s) \lambda^i(s) ds < \infty, \ \forall\, 0 \leq t < \infty \tag{19}$$

For a fixed $\omega \in \Omega$, let $B_t(\omega)$ be the process

$$B_t(\omega) = \sum_{i \in \mathscr{C}^{\Phi}(\omega)} B^i_t(\omega) = \sum_{i=1}^{m} C^i(\omega) B^i_t(\omega) \tag{20}$$

Following Karr (1986), for each $i = 1, \dots, m$, $i \in \mathscr{C}^i$, the $(P, \mathcal{F}_t)$-martingale estimator for the process B^i_t, is the process

$$\widehat{B}^i_t(\omega) = \int_0^t H_i(s) d\widetilde{N}^i_s(\omega), \text{ in } \mathscr{C}^i \tag{21}$$

respectively.

Definition 2. For each $\omega \in \Omega$, the process $\widehat{B}_t(\omega)$ given in (18) is the $(P, \mathcal{F}_t)$-martingale estimator for the process B_t given in (20).

Proposition 1. *Let $H_i(t), i = 1, \dots, m$, be a bounded and continuous functions in $[0, t]$, such that*

$$\sum_{i=1}^{m} \int_0^t H_i^2(s) \lambda^i(s) ds < \infty, \ \forall \ 0 \leq t < \infty \tag{22}$$

Then, for each realization ω and each $i \in \mathscr{C}^{\Phi}(\omega)$, the processes $(\widehat{B}^i - B^i)_{t \geq 0}$, are orthogonal, mean zero, and square integrable $(P, \mathcal{F}_t)$martingales with predictable variation processes $(\langle \widehat{B}^i - B^i \rangle)_{t \geq 0}$ given by

$$\langle \widehat{B}^i - B^{i*} \rangle_t = \int_{Y_i}^t H_i^2(s) \lambda^i(s) ds = \int_0^t H_i^2(s) 1_{\{Y_i < s\}} \lambda^i(s) ds \tag{23}$$

respectively.

***Proof*.** Note that $\forall \ i \in \mathscr{C}^{\Phi}(\omega)$ we have $\omega \in \mathscr{C}^i$. Therefore, from Corollary 1, the $(P, \mathcal{F}_t)$-compensator of $\widehat{B}^i_t = \sum_{j=1}^{\widetilde{N}^i_t} H_i(S_{ij}) = \int_0^t H_i(s) d\widetilde{N}^i_s$ is the process $B^{i*}_t = \int_{Y_i}^t H_i(s) \lambda^i(s) ds = \int_0^t H_i(s) 1_{\{Y_i < s\}} \lambda^i(s) ds$ which represents B^i_t in $\mathscr{C}^i$ (see Note 2).

So, for all $i \in \mathscr{C}^{\Phi}(\omega)$, the predictable variation process of the martingale $(\widehat{B}^i_t - B^i_t)$ is the predictable variation process of the martingale $(\widehat{B}^i_t - B^{i*}_t)$,

$$\langle \widehat{B}^i - B^{i*} \rangle_t = \int_0^t H_i^2(s) d\langle \widetilde{M}^{i*} \rangle_s = \int_0^t H_i^2(s) 1_{\{Y_i < s\}} \lambda^i(s) \, ds$$

Otherwise, since $P(S_i = S_j) = 0$ P-a.s. for all i, j with $i \neq j$, the processes N^i_t and N^j_t do not have simultaneous jumps and so are $\widetilde{N}^i_t$ and $\widetilde{N}^j_t$. Then, for all $i, \in \mathscr{C}^{\Phi}(\omega)$, the $(P, \mathcal{F}_t)$-martingales $\widetilde{M}^{*i}_t$ and $\widetilde{M}^{*j}_t$ are orthogonal and square integrable, so that for $i \neq j$, the martingales $(\widehat{B}^i_t - B^{i*}_t)$ and $(\widehat{B}^j_t - B^{j*}_t)$ are also orthogonal and square integrable. It follows that, for all $i \in \mathscr{C}^{\Phi}(\omega)$, the predictable covariation process

$$\langle \widehat{B}^i - B^i, \widehat{B}^j - B^j \rangle_t = \langle \widehat{B}^i - B^{i*}, \widehat{B}^j - B^{j*} \rangle_t = 0$$

that is, for all $i \in \mathscr{C}^{\Phi}(\omega)$, $(\widehat{B}^i - B^i)(\widehat{B}^j - B^j)$ is a mean zero $(P, \mathcal{F}_t)$-martingale. □

Corollary 2. *Let $H_i(t)$, $i, 1 \leq i \leq m$ be bounded and continuous functions in $[0,t]$ satisfying the condition in (22), and the processes $(\widehat{B}_t)_{t\geq 0}$, $(B_t)_{t\geq 0}$ as were given in (18) and (20), respectively. Then, for a realization $\omega \in \Omega$ and the corresponding set $\mathscr{C}^{\Phi}(\omega)$, the process $(\widehat{B} - B)_{t\geq 0}$ is a mean zero and square integrable $(P, \mathcal{F}_t)$-martingale with predictable variation process $(\langle \widehat{B} - B \rangle)_{t\geq 0}$ given by*

$$\langle \widehat{B} - B \rangle_t = \sum_{i \in \mathscr{C}^{\Phi}(\omega)} \int_{Y_i}^{t} H_i^2(s) \lambda^i(s) ds = \sum_{i=1}^{m} C^i(\omega) \int_0^t H_i^2(s) 1_{\{Y_i < s\}} \lambda^i(s)\, ds \tag{24}$$

***Proof*.** For all $i \in \mathscr{C}^{\Phi}(\omega)$ and from Proposition 1, the processes $(\widehat{B}^i - B^i)_{t\geq 0} = (\widehat{B}^i - B^{i*})_{t\geq 0}, 1 \leq i \leq m$, are orthogonal, mean zero, and square integrable $(P, \mathcal{F}_t)-$martingales with predictable variation processes given by $\langle \widehat{B}^i - B^{i*} \rangle_t = \int_0^t H_i^2(s) 1_{\{Y_i<s\}} \lambda^i(s) ds$, respectively. Therefore,

$$\begin{aligned} \widehat{B}_t(\omega) - B_t(\omega) &= \sum_{i \in \mathscr{C}^{\Phi}(\omega)} (\widehat{B}_t^i(\omega) - B_t^i(\omega)) \\ &= \sum_{i \in \mathscr{C}^{\Phi}(\omega)} \int_0^t H_i(s) \widetilde{M}_s^{i*}(\omega) \in \mathcal{M}_0^2 \end{aligned} \tag{25}$$

and $\langle \widehat{B}^i - B^i, \widehat{B}^j - B^j \rangle = 0, \forall\ i \neq j$. From (23) we have

$$\langle \widehat{B} - B \rangle_t = \sum_{i \in \mathscr{C}^{\Phi}(\omega)} \langle \widehat{B}^i - B^{i*} \rangle_t = \sum_{i=1}^{m} C^i(\omega) \int_0^t H_i^2(s) 1_{\{Y_i<s\}} \lambda^i(s)\, ds$$

□

***Note* 4.** From (24) we have that the expected value of the predictable variation process of the system warranty cost process is

$$E[\langle \widehat{B} - B \rangle_t] = \sum_{i=1}^{m} P(S_i > Y_i) E\Big[\int_{Y_i}^{t} H_i^2(s) \lambda^i(s) ds \Big| S_i > Y_i \Big] \tag{26}$$

3. Statistical Model

3.1. Preliminary

We intend to estimate the expected minimal repair cost $E[\widehat{B}_t]$, over the interval $[0,t]$. First, we need asymptotic results for the estimator of each component expected warranty costs.

From Definitions 1, 2 and Corollary 2 we have

$$\begin{aligned} E[\widehat{B}_t] &= E\Big[\sum_{i\in\mathscr{C}^{\Phi}(\omega)} B_t^i\Big] = \sum_{i=1}^{m} P(S_i > Y_i)E\Big[\int_{Y_i}^{t} H_i(s)\lambda^i(s)ds\Big| S_i > Y_i\Big] \\ &= E[B_t] \end{aligned} \tag{27}$$

where $P(S_i > Y_i)E[\int_{Y_i}^t H_i(s)\lambda^i(s)|S_i > Y_i]$ corresponds to the system minimal repairs expected cost related to the component i.

We consider the sequences $(\widehat{B}_t^{i(j)}, C^{i(j)},\ i = 1,\dots,m)_{t\geq 0}$, $1 \leq j \leq n$, of n independent and identically distributed copies of the $m-$variate process $(\widehat{B}_t^i, C^i,\ i = 1,\dots,m)_{t\geq 0}$.

For $j = 1,\dots,n$ let $\mathscr{C}^{\Phi(j)} = \{i \in \{1,\dots,m\} : S_i^{(j)} > Y_i^{(j)}\}$ be the set of critical components for the $j-$th observed system, where $S_i^{(j)}$ is the first failure time of component i and $Y_i^{(j)}$ its critical level. Then, the minimal repairs expected cost for the $j-$th system is

$$\widehat{B}_t^{(j)} = \sum_{i\in\mathscr{C}^{\Phi(j)}} \widehat{B}_t^{i(j)} = \sum_{i=1}^{m} C^{i(j)}\widehat{B}_t^{i(j)} \tag{28}$$

and its compensator process is (from Corollary 2)

$$B_t^{(j)} = \sum_{i\in\mathscr{C}^{\Phi(j)}} B_t^{i(j)} = \sum_{i=1}^{m} C^{i(j)} \int_{Y_i^{(j)}}^{t} H_i(s)\lambda^i(s)\,ds \tag{29}$$

For n copies we consider the mean processes

$$\overline{\widehat{B}}_t^{(n)} = \frac{1}{n}\sum_{j=1}^{n}\widehat{B}_t^{(j)} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} C^{i(j)}\int_0^t H_i(s)d\widetilde{N}_s^{i(j)} \tag{30}$$

$$\overline{B}_t^{(n)} = \frac{1}{n}\sum_{j=1}^{n} B_t^{(j)} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} C^{i(j)}\int_{Y_i^{(j)}}^t H_i(s)\lambda^i(s)\,ds \tag{31}$$

Let

$$\overline{\widehat{B}}_t^{i(n)} = \frac{1}{n}\sum_{j=1}^{n} C^{i(j)}\widehat{B}_t^{i(j)} \qquad \text{and} \qquad \overline{B}_t^{i(n)} = \frac{1}{n}\sum_{j=1}^{n} C^{i(j)}B_t^{i(j)} \tag{32}$$

Then, from (30) and (31), we also have

$$\overline{\widehat{B}}_t^{(n)} = \sum_{i=1}^{m}\overline{\widehat{B}}_t^{i(n)} \qquad \text{and} \qquad \overline{B}_t^{(n)} = \sum_{i=1}^{m}\overline{B}_t^{i(n)} \tag{33}$$

For each $i = 1,\dots,m$ we propose $\overline{\widehat{B}}_t^{i(n)}$ as the estimator for the system minimal repairs expected cost related to the component i.

Theorem 2. *For each $i = 1, \dots, m$ let $B^{i*}(t)$ be*

$$B^{i*}(t) = P(S_i > Y_i)E\Big[\int_{Y_i}^{t} H_i(s)\lambda^i(s)ds\Big|S_i > Y_i\Big] \tag{34}$$

Then, under conditions in Proposition 1, $\overline{\widehat{B}}_t^{i(n)}$ is a consistent and unbiased estimator for the minimal repairs expected cost related to component i, $B^{i*}(t)$.

***Proof*.** See Appendix A. □

3.2. The Central Limit Theorem

In what follows we prove that the m-variate error process of the proposed estimators, $(\overline{\widehat{B}}_t^{i(n)} - B^{i*}(t), i = 1, \dots, m)$, conveniently standardized, satisfies the Martingale Central Limit Theorem, as in Karr (1986).

Theorem 3. *(Karr 1986, Theorem 5.11). For fixed m and for each n, $n \geq 1$, let $(M_t^{i(n)}, i = 1, \dots, m)$ be a sequence of orthogonal, mean zero and square integrable martingales with jumps,at time t, $\Delta M_t^{i(n)} = M_t^{i(n)} - M_t^{i(n)-}$, where $M_t^{i(n)-} = \lim\limits_{h\downarrow 0} M_{t-h}^{i(n)}$. For each $i, i = 1, \dots, m$ let $V_i(t)$ be a continuous and non decreasing function with $V_i(0) = 0$. If*

(a) $\forall\ t \geq 0$ and $i = 1, \dots, m$

$$\langle M^{i(n)}\rangle_t \xrightarrow[n\to\infty]{\mathcal{D}} V_i(t) \tag{35}$$

(b) There is a sequence $(c_n)_{n\geq 1}$, such that $c_n \xrightarrow[n\to\infty]{} 0$ and

$$P(\sup_{s\leq t} \mid \triangle M_s^{i(n)} \mid \leq c_n) \xrightarrow[n\to\infty]{} 1 \tag{36}$$

Then exist an m-variate Gaussian continuous process, $\mathbf{M} = (M^i, i = 1, \dots, m)$ where M^i is a martingale with

$$\langle M^i, M^k\rangle_t = 1_{\{i=k\}} V_i(t) \tag{37}$$

such that $\mathbf{M}^{(n)} = (M^{1(n)}, \dots, M^{m(n)}) \xrightarrow[n\to\infty]{\mathcal{D}} \mathbf{M} = (M^1, \dots, M^m)$ in $D[0, t]^m$

***Note* 5.** In the above theorem the conditions (a) and (b) are sufficient to prove the convergence of the finite-dimensional distributions and tightness of the sequence $\mathbf{M}^{(n)}$ in the m-dimensional space $D[0, t]^m$ of the right-continuous functions with left limits, in $[0, t]$ (Karr 1986).

Corollary 3. *Suppose that for each* $i, i = 1, \dots, m,$ $\int_0^t H_i^2(s)\lambda^i(s)ds < \infty$ *and let* $V_i^*(t)$ *be the function*

$$V_i^*(t) = P(S_i > Y_i)E\Big[\int_{Y_i}^t H_i^2(s)\lambda^i(s)ds\Big|S_i > Y_i\Big] \tag{38}$$

Let $\overline{\widehat{\mathbf{B}}}_t^{(n)} = \left(\overline{\widehat{B}}_t^{1(n)}, \dots, \overline{\widehat{B}}_t^{m(n)}\right)$ *and* $\overline{\mathbf{B}}_t^{(n)} = \left(\overline{B}_t^{1(n)}, \dots, \overline{B}_t^{m(n)}\right)$ *be m-variate processes. Then, the process* $\mathbf{M}^{(n)} = \sqrt{n}(\overline{\widehat{\mathbf{B}}}^{(n)} - \overline{\mathbf{B}}^{(n)}) \xrightarrow[n\to\infty]{\mathcal{D}} \mathbf{M}$ *in* $D[0,t]^m$*, where* $\mathbf{M}$ *is an m-variate Gaussian continuous process with martingales components.*

***Proof*.** We establish the conditions (a) and (b) of Theorem 3. Denote $M_t^{i(n)} = \sqrt{n}\left(\overline{\widehat{B}}_t^{i(n)} - \overline{B}_t^{i(n)}\right)$, $i = 1, \dots, m$. As $P(S_i = S_j) = 0$ for all i, j with $i \neq j$, from Proposition 1

$$\left(\overline{\widehat{B}}_t^{i(n)} - \overline{B}_t^{i(n)}\right) = \frac{1}{n}\sum_{j=1}^n C^{i(j)}\left(\widehat{B}_t^{i(j)} - B_t^{i(j)}\right), \ 1 \leq i \leq m,$$

are orthogonal, mean zero and square integrable $(P, \mathcal{F}_t)$-martingales, for each $n \geq 1$.

Therefore, for all $n \geq 1$ and $i \neq j$, $\langle M^{i(n)}, M^{j(n)}\rangle_t = 0$, from the Strong Law of Large Numbers and (60), for all i, $1 \leq i \leq m$

$$\langle M^{i(n)}\rangle_t = n\sum_{j=1}^n C^{i(j)}\Big[\frac{1}{n^2}\int_{Y_i^{(j)}}^t H_i^2(s)\lambda^i(s)ds\Big] = \frac{1}{n}\sum_{j=1}^n C^{i(j)}\Big[\int_{Y_i^{(j)}}^t H_i^2(s)\lambda^i(s)ds\Big]$$
$$\xrightarrow[n\to\infty]{} P(S_i > Y_i)E\Big[\int_{Y_i}^t H_i^2(s)\lambda^i(s)\,ds\Big|S_i > Y_i\Big] = V_i^*(t) < \infty \tag{39}$$

and we have $\langle M^{i(n)}\rangle_t \xrightarrow[n\to\infty]{\mathcal{D}} V_i^*(t)$, for all $t \geq 0$.

Furthermore, the jumps of $M^{i(n)}$ arise only from $\sqrt{n}\overline{\widehat{B}}_t^{i(n)}$ and they are of size $\triangle M_t^{i(n)} = \dfrac{H_i(t)}{\sqrt{n}}$. By hypothesis, $H_i(t)$ is continuous and bounded in $[0,t]$, say by a constant $\Gamma < \infty$. Taking $c_n = \Gamma n^{-\frac{1}{4}}$, the condition (b) of Theorem 3 is satisfied. Therefore, $\mathbf{M}^{(n)} \xrightarrow[n\to\infty]{\mathcal{D}} \mathbf{M}$, where $\mathbf{M}$ is an m-variate Gaussian continuous process, $\mathbf{M} = (M^i, i = 1, \dots, m)$, with martingale components M^i, $i = 1, \dots, m$ such that $\langle M^i, M^k\rangle_t = 1_{\{i=k\}}V_i^*(t)$. □

Proposition 2. *Let* $\mathbf{Z}_t^{(n)}$ *be the m-variate process* $\mathbf{Z}_t^{(n)} = \left(Z_t^{1(n)}, \dots, Z_t^{m(n)}\right)$ *where* $Z_t^{i(n)} = \sqrt{n}(\overline{B}_t^{i(n)} - B^{i*}(t))$*,* $i = 1, \dots, m$ *and suppose that for all* i, $1 \leq i \leq m$ *and* $t \geq 0$*,*

$$\sigma^{2i*}(t) = Var[C^iB_t^i] = E\Big[C^i\Big(\int_{Y_i}^t H_i(s)\lambda^i(s)ds\Big)^2\Big] - (B^{i*}(t))^2 < \infty. \tag{40}$$

Then $\mathbf{Z}_t^{(n)} \xrightarrow[n\to\infty]{\mathcal{D}} \mathbf{Z}_t$*, where* $\mathbf{Z}_t$ *is an* m*-variate Normal random vector with mean zero and covariance matrix* $\boldsymbol{\Sigma}(t)$ *such that* $\Sigma_{ij}(t) = 1_{\{i=j\}}\sigma^{2i*}(t)$*.*

***Proof*.** See Appendix B. □

Theorem 4. *Let* $\boldsymbol{\mu}(t) = (B^{1*}(t), \ldots, B^{m*}(t))$*,* $\delta^{2i*}(t) = Var[C^i \widehat{B}_t^i] < \infty$*,* $i = 1, \ldots, m$ *and suppose that the conditions of Corollary 3 and Proposition 2 holds. Then, the process* $\mathbf{E}_t^{(n)} = \sqrt{n}(\overline{\widehat{\mathbf{B}}}_t^{(n)} - \boldsymbol{\mu}(t)) \xrightarrow[n\to\infty]{\mathcal{D}} \mathbf{W}_t$ *in* $D[0,t]^m$*, where* $\mathbf{W}_t = (W_t^1, \ldots, W_t^m)$ *is an* m*-variate Gaussian process with mean zero and covariance matrix* $\mathbf{U}(t)$ *with* $U_{ij}(t) = 1_{\{i=j\}}\delta^{2i*}(t)$*.*

***Proof*.** See Appendix C. □

***Note* 6.** In order to apply Theorem 4 we must estimate for fixed t, the variances $\delta^{2i*}(t),\ i = 1, \ldots, m$, which can be done through the sample estimator of the variance.

For $t \geq 0$ we consider n independent and identically distributed copies of the $m-$variate process $((\widehat{B}_t^i, C^i),\ i = 1, \ldots, m)$, with covariance matrix given by

$$\mathbf{U}(t) = \begin{bmatrix} \delta^{21*}(t) & 0 & 0 & \cdots & 0 \\ 0 & \delta^{22*}(t) & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & \delta^{2m*}(t) \end{bmatrix} \tag{41}$$

We propose as estimator of $\mathbf{U}(t)$ to the sample covariance matrix, $\mathbf{S}^{(n)}(t)$, where $S_{ij}^{(n)}(t) = 1_{\{i=j\}} S_t^{2i(n)}$, that is

$$\mathbf{S}^{(n)}(t) = \begin{bmatrix} S_t^{21(n)} & 0 & 0 & \cdots & 0 \\ 0 & S_t^{22(n)} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & S_t^{2m(n)} \end{bmatrix} \tag{42}$$

with

$$S_t^{2i(n)} = \Big(\frac{n}{n-1}\Big)\Big[\frac{1}{n}\sum_{j=1}^{n}\big(C^{i(j)}\widehat{B}_t^{i(j)} - B^{i*}(t)\big)^2 - \big(\overline{\widehat{B}}_t^{i(n)} - B^{i*}(t)\big)^2\Big] \tag{43}$$

Therefore, for each $i, 1 \leq i \leq m$ and fixed $t \geq 0$, we calculate the corresponding sample estimator of variance, $S_t^{2i(n)}$, which satisfies the properties enunciated in the following proposition.

Proposition 3. *For each* $i, 1 \leq i \leq m$*,* $S_t^{2i(n)}$ *is an unbiased and uniformly consistent estimator for* $\delta^{2i*}(t)$ *and therefore,* $\mathbf{S}^{(n)}(t)$ *and* $\sum\limits_{i=1}^{m} S_t^{2i(n)}$ *are unbiased and uniformly consistent estimator for* $\mathbf{U}(t)$ *and* $\sum\limits_{i=1}^{m} \delta^{2i*}(t)$*, respectively.*

Proof. For each $i, 1 \leq i \leq m$ and $t \geq 0$ we have $E[C^i \widehat{B}_t^i] = E[\overline{\widehat{B}}_t^{i(n)}] = B^{i*}(t)$ and $\delta^{2i*}(t) = E[(C^i \widehat{B}_t^i - B^{i*}(t))^2]$. As the copies are independent and identically distributed from (43) we get

$$E[S_t^{2i(n)}] = \left(\frac{n}{n-1}\right)\left[\delta^{2i*}(t) - \frac{1}{n}\delta^{2i*}(t)\right] = \delta^{2i*}(t), \text{ and therefore,}$$

$$E[\mathbf{S}^{(n)}(t)] = \mathbf{U}(t) \quad \text{and} \quad E\Big[\sum_{i=1}^{m} S_t^{2i(n)}\Big] = \sum_{i=1}^{m} \delta^{2i*}(t) \tag{44}$$

Also, we apply the Strong Law of Large Number to obtain, for all $t \geq 0$,

$$\frac{1}{n}\sum_{j=1}^{n}\big(C^{i(j)}\widehat{B}_t^{i(j)} - B^{i*}(t)\big)^2 \xrightarrow[n\uparrow\infty]{} \delta^{2i*}(t)$$

From the Strong Law of Large Number and the Continuous Mapping Theorem (See, Billingsley 1968), we have,

$$\big(\overline{\widehat{B}}_t^{i(n)} - B^{i*}(t)\big)^2 \xrightarrow[n\uparrow\infty]{} 0$$

and $\left(\frac{n}{n-1}\right) \xrightarrow[n\uparrow\infty]{} 1$. From the above results and (43), for all i, $1 \leq i \leq m$ we conclude

$$S_t^{2i(n)} \xrightarrow[n\uparrow\infty]{} \delta^{2i*}(t), \quad \forall\ t \geq 0$$

Then

$$S_s^{2i(n)} \xrightarrow[n\uparrow\infty]{} \delta^{2i*}(s), \quad \forall\ s \leq t, \quad \sup_{s\leq t}|S_s^{2i(n)} - \delta^{2i*}(s)| \xrightarrow[n\uparrow\infty]{} 0$$

and therefore,

$$\sup_{s\leq t}\big(S_s^{2i(n)} - \delta^{2i*}(s)\big)^2 \xrightarrow[n\uparrow\infty]{} 0$$

It follows from the above results that

$$E\Big[\sup_{s\leq t}\big(S_s^{2i(n)} - \delta^{2i*}(s)\big)^2\Big] \xrightarrow[n\uparrow\infty]{} 0 \tag{45}$$

This result gives the uniformly consistence of the estimators $S_t^{2i(n)}$ and $\sum_{i=1}^{m} S_t^{2i(n)}$, which also warranties the consistence of the estimator $\mathbf{S}^{(n)}(t)$ given in (42). □

3.3. Estimation of the Expected Warranty Cost for a Fixed Warranty Period of Length *w*

From (27) and (34), the expected warranty cost for a fixed period of length w is $B^*(w) = E[\widehat{B}_{\mathrm{w}}] = \sum_{i=1}^{m} B^{i*}(w) = E[B_{\mathrm{w}}]$. In this section we obtain a $(1-\alpha)100\%$ confidence interval from results in Section 2.3 to Section 3.2.

Let $\mathbf{1}_m = (1, 1, \ldots, 1)$ be the m-dimensional unit vector and $(\mathbf{A})^t$ the transpose of corresponding vector or matrix $\mathbf{A}$. From (33) and Theorem 2 an estimator of $B^*(w)$ is $\widehat{B}^*(w) = \overline{\widehat{B}}_{\mathrm{w}}^{(n)}$, which can be write as

$$\overline{\widehat{B}}_{\mathrm{w}}^{(n)} = \sum_{i=1}^{m} \overline{\widehat{B}}_{\mathrm{w}}^{i(n)} = \mathbf{1}_m \big(\overline{\widehat{\mathbf{B}}}_{\mathrm{w}}^{(n)}\big)^t = \overline{\widehat{\mathbf{B}}}_{\mathrm{w}}^{(n)} \big(\mathbf{1}_m\big)^t \tag{46}$$

where $\overline{\widehat{\mathbf{B}}}_{\mathrm{w}}^{(n)} = (\overline{\widehat{B}}_{\mathrm{w}}^{1(n)}, \ldots, \overline{\widehat{B}}_{\mathrm{w}}^{m(n)})$ Also, we can write the expected warranty cost $B^*(w)$ as

$$B^*(\mathrm{w}) = \sum_{i=1}^{m} B^{i*}(\mathrm{w}) = \mathbf{1}_m \big(\boldsymbol{\mu}(\mathrm{w})\big)^t = \boldsymbol{\mu}(\mathrm{w})\big(\mathbf{1}_m\big)^t \tag{47}$$

with $\boldsymbol{\mu}(w) = (B^{1*}(w), \ldots, B^{m*}(w))$ as defined in Theorem 4.

Theorem 5.

i. $\overline{\widehat{B}}_{\mathrm{w}}^{(n)}$ *is a consistent and unbiased estimator for* $B^*(w)$.

ii. A consistent and unbiased estimator for $Var[\overline{\widehat{B}}_{\mathrm{w}}^{(n)}]$ *is* $\widehat{Var}[\overline{\widehat{B}}_{\mathrm{w}}^{(n)}] = \frac{1}{n}\sum_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}$.

iii. An approximate $(1-\alpha)100\%$ *confidence interval for* $B^*(\mathrm{w})$, *is*

$$\overline{\widehat{B}}_{\mathrm{w}}^{(n)} \pm Z_{1-\alpha/2}\sqrt{\frac{1}{n}\sum_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}} \tag{48}$$

where Z_γ *is the* γ*-quantile of the standard normal distribution.*

Proof.

i. For each i, $1 \leq i \leq m$, from Theorem 2, $\overline{\widehat{B}}_{\mathrm{w}}^{i(n)}$ is a consistent and unbiased estimator for $B^{i*}(w)$. Then, $\overline{\widehat{B}}_{\mathrm{w}}^{(n)} = \sum_{i=1}^{m} \overline{\widehat{B}}_{\mathrm{w}}^{i(n)}$ is a consistent and unbiased estimator for $B^*(w)$.

ii. Since for $i \neq j$ the processes $\widehat{B}_{\mathrm{w}}^{i}$ and $\widehat{B}_{\mathrm{w}}^{j}$ do not have simultaneous jumps, we have

$$\mathrm{Var}[\overline{\widehat{B}}_{\mathrm{w}}^{(n)}] = \frac{1}{n}\sum_{i=1}^{m} \delta^{2i*}(\mathrm{w}) = \frac{1}{n}\mathbf{1}_m \mathbf{U}(\mathrm{w})\big(\mathbf{1}_m\big)^t = \frac{1}{n}\mathrm{Var}[\big(\widehat{B}_{\mathrm{w}}^{1}, \ldots, \widehat{B}_{\mathrm{w}}^{m}\big)\big(\mathbf{1}_m\big)^t]$$

Therefore, from Proposition 3, (42) and (44), an unbiased and consistent estimator for $\mathrm{Var}[\overline{\widehat{B}}_{\mathrm{w}}^{(n)}]$ is

$$\widehat{\mathrm{Var}}[\overline{\widehat{B}}_{\mathrm{w}}^{(n)}] = \frac{1}{n}\sum_{i=1}^{m} S_{\mathrm{w}}^{2i(n)} = \frac{1}{n}\mathbf{1}_m \mathbf{S}^{(n)}(\mathrm{w})\big(\mathbf{1}_m\big)^t \tag{49}$$

iii. As a consequence of Theorem 4, and the Crámer-Wold procedure (See, Fleming & Harrington 1991, Lemma 5.2.1) we have

$$\mathbf{1}_m\big(\mathbf{E}_{\mathrm{w}}^{(n)}\big)^t = \mathbf{E}_{\mathrm{w}}^{(n)}\big(\mathbf{1}_m\big)^t = \sum_{i=1}^{m} E_{\mathrm{w}}^{i(n)} \xrightarrow[n\to\infty]{\mathcal{D}}$$

$$\mathbf{1}_m\big(\mathbf{W}_{\mathrm{w}}\big)^t = \mathbf{W}_{\mathrm{w}}\big(\mathbf{1}_m\big)^t \sim N(0, \mathbf{1}_m\mathbf{U}(\mathrm{w})\big(\mathbf{1}_m\big)^t) = N\left(0,\ \sum_{i=1}^{m}\delta^{2i*}(\mathrm{w})\right)$$

then

$$\frac{\sum\limits_{i=1}^{m} E_{\mathrm{w}}^{i(n)}}{\sqrt{\sum\limits_{i=1}^{m}\delta^{2i*}(\mathrm{w})}} \xrightarrow[n\to\infty]{\mathcal{D}} N(0,1) \tag{50}$$

From Proposition 3 and the Slutzky Theorem,

$$\begin{aligned}\frac{\sum\limits_{i=1}^{m} E_{\mathrm{w}}^{i(n)}}{\sqrt{\sum\limits_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}}} &= \frac{\sqrt{n}\sum\limits_{i=1}^{m}(\overline{\widehat{B}}_{\mathrm{w}}^{i(n)} - B^{i*}(\mathrm{w}))}{\sqrt{\sum\limits_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}}} \\ &= \frac{\sqrt{n}(\overline{\widehat{B}}_{\mathrm{w}}^{(n)} - B^{*}(\mathrm{w}))}{\sqrt{\sum\limits_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}}} \xrightarrow[n\to\infty]{\mathcal{D}} N(0,1)\end{aligned} \tag{51}$$

From the last equation we get

$$\lim_{n\to\infty} P\left\{\frac{\sqrt{n}|\overline{\widehat{B}}_{\mathrm{w}}^{(n)} - B^{*}(\mathrm{w})|}{\sqrt{\sum\limits_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}}} \leq Z_{1-\alpha/2}\right\} \geq P\left\{|Z| \leq Z_{1-\alpha/2}\right\} = 1-\alpha$$

and a $(1-\alpha)100\%$ approximate pointwise confidence interval for $B^*(w)$, is

$$\overline{\widehat{B}}_{\mathrm{w}}^{(n)} \pm Z_{1-\alpha/2}\sqrt{\frac{1}{n}\sum_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}}$$

□

The confidence interval for $B^*(w)$ given in (48) can have negative values and it is not acceptable. We propose to build a confidence interval through a convenient bijective transformation $g(x)$ such that $\frac{d}{dx}g(x)\big|_{x=B^*(w)} \neq 0$, which does not contain negative values. Conveniently, we consider $g(x) = \log x$, $x > 0$ with $\frac{d}{dx}g(x) = 1/x, x > 0$.

Corollary 4. *Suppose that for a fixed* $\mathrm{w} > 0$, $B^*(\mathrm{w}) > 0$ *and* $\overline{\widehat{B}}_{\mathrm{w}}^{(n)} > 0$. *Then*

$$\overline{\widehat{B}}_{\mathrm{w}}^{(n)} \times \exp\left\{\pm Z_{1-\alpha/2}\sqrt{\sum_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}\Big/ n\big[\overline{\widehat{B}}_{\mathrm{w}}^{(n)}\big]^2}\right\} \tag{52}$$

is an approximate $(1-\alpha)100\%$ *confidence interval for* $B^*(\mathrm{w})$.

***Proof*.** Using the Delta Method (See, Lehmann 1999, Section 2.5) and formula (50) we get

$$\sqrt{n}[\log(\overline{\widehat{B}}_{\mathrm{w}}^{(n)}) - \log(B^*(\mathrm{w}))] \xrightarrow[n\to\infty]{\mathcal{D}} N\Big(0, [B^*(\mathrm{w})]^{-2}\sum_{i=1}^{m}\delta^{2i*}(\mathrm{w})\Big) \tag{53}$$

From literal i in Theorem 5, Proposition 3, and the Continuous Mapping Theorem (Billingsley 1968),

$$\frac{\overline{\widehat{B}}_{\mathrm{w}}^{(n)}}{\sqrt{\sum_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}}} \xrightarrow[n\to\infty]{} \frac{B^*(\mathrm{w})}{\sqrt{\sum_{i=1}^{m}\delta^{2i*}(\mathrm{w})}}, \tag{54}$$

Therefore, for fixed w, from (53), (54) and using Slutsky Theorem, we have

$$\frac{\sqrt{n}\overline{\widehat{B}}_{\mathrm{w}}^{(n)}}{\sqrt{\sum_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}}}[\log(\overline{\widehat{B}}_{\mathrm{w}}^{(n)}) - \log(B^*(\mathrm{w}))] \xrightarrow[n\to\infty]{\mathcal{D}} N(0,1) \tag{55}$$

From the last equation, an approximate $(1-\alpha)100\%$ confidence interval for $\log(B^*(\mathrm{w}))$ is

$$\log(\overline{\widehat{B}}_{\mathrm{w}}^{(n)}) \pm Z_{1-\alpha/2}\sqrt{\sum_{i=1}^{m} S_{\mathrm{w}}^{2i(n)}\Big/ n\big[\overline{\widehat{B}}_{\mathrm{w}}^{(n)}\big]^2} \tag{56}$$

from which, applying the inverse transformation, that is, $\exp(x)$, we obtain (52). □

3.4. Example

To illustrate the results we simulated the minimal repair warranty cost process for a parallel system of three independent components with lifetimes $S_i \sim \text{Weibull}(\theta_i, \beta_i)$, $i = 1, 2, 3$, respectively, where θ_i is the scale parameter and β_i is the shape parameter, that is, with survival function $\overline{F}^i(t) = \exp[-(t/\theta_i)^{\beta_i}]$ and hazard rate function $\lambda^i(t) = (\beta_i/\theta_i^{\beta_i})t^{\beta_i - 1}$, $t > 0$.

We use two possible cost functions: the first one is $H_i(t) = C_i e^{-\delta t}$ and the second one is $H_i(t) = C_i\left(1 - \frac{t}{w}\right)e^{-\delta t}$, $0 \leq t \leq w$, with $\delta = 1$ in both cases. Clearly they are bounded and continuous functions in $[0, t]$. The parameter values are indicated in Table 1, $w = 5$ is the fixed warranty period and the sample sizes are $n = 30, 50, 100, 500, 1000, 2000, 5000, 10000$.

The critical levels of the components for the system under minimal repair are

$$Y_i = \begin{cases} \max\limits_{j \neq i} S_j & \text{if } \max\limits_{j \neq i} S_j < S_i, \\ \infty & \text{if } \max\limits_{j \neq i} S_j \geq S_i, \end{cases} \qquad i = 1, 2, 3 \tag{57}$$

Therefore, if the component failure times are observed in order S_2, S_3, S_1, then $T = \max\{S_1, S_2, S_3\} = \min\limits_{\{Y_i < \infty\}} S_i = S_1$, and, in this case, component 1, is the only one critical for the system. Therefore, after the second component failure time, S_3, the system is reduced to component 1, which is minimally repaired in each observed failure over the warranty period.

Table 1: Parameter values.

i	θ_i	β_i	C_i
1	1	1.5	3
2	1	1.5	3
3	2	2.0	5

The simulation results considering the cost function as $H_i(t) = C_i e^{-\delta t}$ are:

In Table 2, the limits correspond to the confidence interval defined in (52), with confidence level of $\alpha = 0.05$. In Figure 1, we show the 95% approximate pointwise confidence intervals for sample size of $n = 100$ and $w \in (0, 5]$.

Table 2: Estimations for some sample sizes, $H_i(t) = C_i e^{-\delta t}$, $w = 5$, $\alpha = 0.05$.

n	$\widehat{B}^*(w)$	$\sum\limits_{i=1}^{3} S_w^{2i(n)}$	$\sum\limits_{i=1}^{3} S_w^{2i(n)}/n$	Lower limit	Upper limit
30	1.90	2.30	0.07654	1.430	2.529
50	1.89	2.96	0.05921	1.471	2.435
100	1.85	2.95	0.02954	1.543	2.221
500	1.83	2.84	0.00568	1.685	1.980
1000	1.81	2.76	0.00276	1.712	1.918
2000	1.85	2.84	0.00142	1.780	1.928
5000	1.84	2.83	0.00057	1.794	1.887
10000	1.86	2.90	0.00029	1.825	1.891

Table 3, presents the theoretical values for the expected cost for a warranty period of length $w = 5$, where $E[B_w^i] = \int_0^w H_i(s)\lambda^i(s)\,ds$ (that is, when the component i is minimally repaired at each observed failure) and $E[B_w^i \mid S_i > Y_i] = E\left[\int_{Y_i}^{w} H_i(s)\lambda^i(s)ds \mid S_i > Y_i\right]$. Based on these results, we can conclude that for the considered system, the estimated values are closer to the expected values for sample sizes greater than $n = 50$.

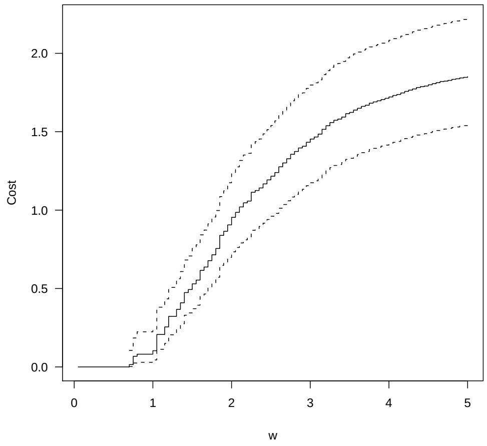

Figure 1: 95% Approximate pointwise confidence intervals using limits in (52) with simulated samples and $H_i(t) = C_i e^{-\delta t}$.

Table 3: Expected costs, $H_i(t) = C_i e^{-\delta t}$, $\mathrm{w} = 5$.

i	$E[B^i_{\mathrm{w}}]$	$E[B^i_{\mathrm{w}} \mid S_i > Y_i]$	$P(S_i > Y_i)$	$P(S_i > Y_i)E[B^i_{\mathrm{w}} \mid S_i > Y_i]$
1	3.9138	2.14648	0.1620753	0.348
2	3.9138	2.14648	0.1620753	0.348
3	2.3989	1.70345	0.6758494	1.151
			System cost	1.847

The following results correspond to the Monte Carlo simulations in which we got the mean cost for $w = 5$ and 1000 samples of size $n = 100$ and $n = 200$, respectively. Table 4, presents several statistics and the Shapiro Wilk normality test. In Figure 2, we show the histograms of mean costs.

Table 4: Statistics of Monte Carlo simulation, $H_i(t) = C_i e^{-\delta t}$, $\mathrm{w} = 5$.

n	$\overline{X}_n$	S^2_n	$\widetilde{X}_n$	$P_{2.5}$	P_{25}	P_{75}	$P_{97.5}$	S.Wilk	P-value
100	1.848	0.01574	1.848	1.612	1.761	1.936	2.091	0.9990	0.8576
200	1.843	0.00851	1.841	1.666	1.782	1.905	2.036	0.9987	0.7200

From results in Tables 2 to 4, and Figures 1 and 2, we observe that the mean cost is approximately 1.85 for a warranty period of length $\mathrm{w} = 5$. Also, the sample variance and the 2.5th and 97.5th sample percentiles for the mean costs from samples of size $n = 100$ showed in Table 4. They are close to the corresponding values in Table 2 for $\sum_{i=1}^{3} S^{2i(n)}_{\mathrm{w}}/n$ and the confidence limits, respectively, and it becomes clear that, in this case, the normal approximation is already achieved with samples of size 100.

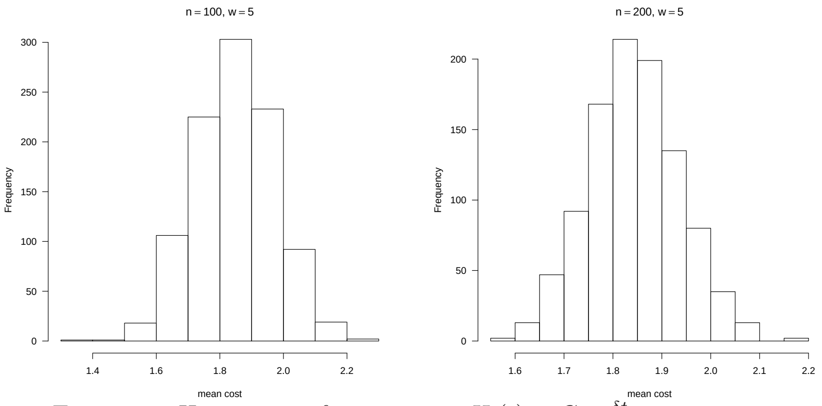

FIGURE 2: Histograms of mean costs, $H_i(t) = C_i e^{-\delta t}$, $n = 100, 200$.

The results related to the minimal repair costs with functions given by $H_i(t) = C_i\left(1 - \frac{t}{w}\right) e^{-\delta t}$ are showed in Tables 5, 6 and 7 and Figures 3 and 4. The conclusions are similar to the previous case.

TABLE 5: Estimations for different sample sizes, $H_i(t) = C_i\left(1 - \frac{t}{w}\right) e^{-\delta t}$, $w = 5$, $\alpha = 0.05$.

n	$\widehat{B}^*(w)$	$\sum_{i=1}^{3} S_w^{2i(n)}$	$\sum_{i=1}^{3} S_w^{2i(n)}/n$	Lower limit	Upper limit
30	1.16	1.35	0.04516	0.811	1.662
50	1.10	1.27	0.02544	0.827	1.460
100	1.04	1.23	0.01232	0.847	1.286
500	1.08	1.30	0.00259	0.985	1.185
1000	1.02	1.22	0.00122	0.953	1.090
2000	1.03	1.25	0.00063	0.978	1.076
5000	1.04	1.27	0.00025	1.007	1.069
10000	1.04	1.26	0.00013	1.014	1.058

TABLE 6: Expected costs, $H_i(t) = C_i\left(1 - \frac{t}{w}\right) e^{-\delta t}$, $w = 5$.

i	$E[B_w^i]$	$E[B_w^i \mid S_i > Y_i]$	$P(S_i > Y_i)$	$P(S_i > Y_i)E[B_w^i \mid S_i > Y_i]$
1	2.8076	1.26053	0.1620753	0.204
2	2.8076	1.26053	0.1620753	0.204
3	1.5236	0.93862	0.6758494	0.634
			System cost	1.043

TABLE 7: Statistics of Monte Carlo simulation, $H_i(t) = C_i\left(1 - \frac{t}{w}\right) e^{-\delta t}$, $w = 5$.

n	$\widehat{X}_n$	S_n^2	$\widetilde{X}_n$	$P_{2.5}$	P_{25}	P_{75}	$P_{97.5}$	S.Wilk	P-value
100	1.038	0.00925	1.037	0.846	0.973	1.100	1.227	0.9987	0.6595
200	1.042	0.00418	1.040	0.917	0.996	1.085	1.171	0.9985	0.5533

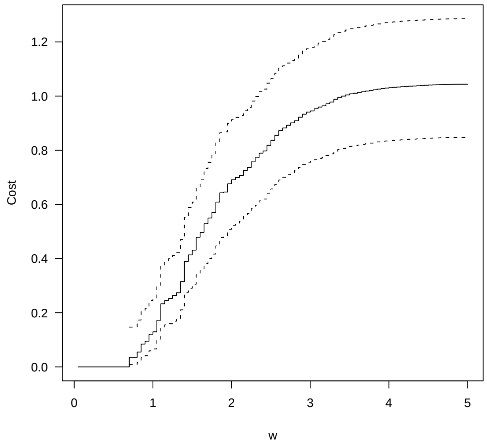

Figure 3: 95% Approximate pointwise confidence intervals using limits in (52) with simulated samples and $H_i(t) = C_i\left(1 - \frac{t}{w}\right)e^{-\delta t}$.

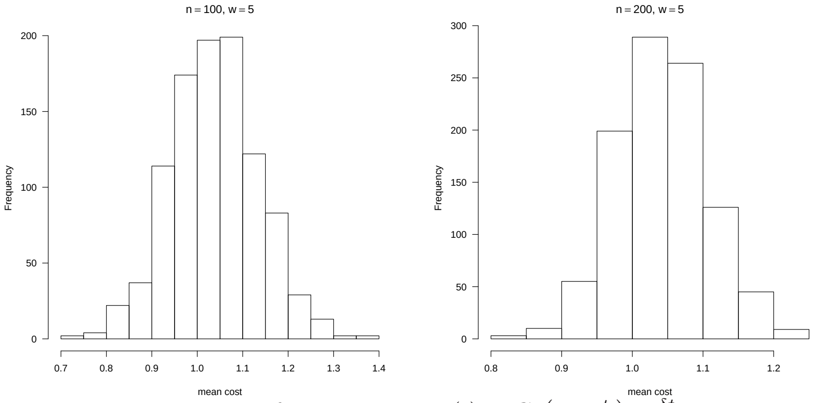

Figure 4: Histograms of mean costs, $H_i(t) = C_i\left(1 - \frac{t}{w}\right)e^{-\delta t}$, $n = 100, 200$.

4. Conclusions

A martingale estimator for the expected discounted warranty cost process of a minimally repaired coherent system under its component level observation was proposed. Its asymptotic properties were also presented using the Martingale Central Limit Theorem.

Acknowledgements

Dr. Bueno was partially supported by FAPESP process No. 2010/52227-0, Brazil and Mrs. González was partially supported by CNPq process No. 140903/2007-0, Brazil.

[Recibido: octubre de 2010 — Aceptado: agosto de 2011]

References

Arjas, E. (1981), 'The failure and hazard processes in multivariate reliability systems', *Mathematics of Operations Research* **6**(4), 551–562.

Aven, T. & Jensen, U. (1999), *Stochastic Models in Reliability*, Springer-Verlag, Inc., New York.

Bai, J. & Pham, H. (2004), 'Discounted warranty cost of minimally repaired series systems', *IEEE Transactions on Reliability* **53**, 37–42.

Bai, J. & Pham, H. (2006), 'Cost analysis on renewable full-service warranties for multi-component systems', *European Journal of Operational Research* **168**, 492–508.

Balachandran, K. R., Maschmeyer, R. & Livingstone, J. (1981), 'Product warranty period: A Markovian approach to estimation and analysis of repair and replacement costs', *The Accounting Review* **1**, 115–124.

Billingsley, P. (1968), *Convergence of Probability Measures*, John Wiley and Sons, Inc., New York.

Blischke, W. R. & Murthy, D. N. P. (1992*a*), 'Product warranty management-1: A taxonomy for warranty policies', *European Journal of Operational Research* **62**, 127–148.

Blischke, W. R. & Murthy, D. N. P. (1992*b*), 'Product warranty management-2: An integrated framework for study', *European Journal of Operational Research* **62**, 261–281.

Blischke, W. R. & Murthy, D. N. P. (1992*c*), 'Product warranty management-3: A review of mathematical models', *European Journal of Operational Research* **63**, 1–34.

Blischke, W. R. & Murthy, D. N. P. (1994), *Warranty Cost Analysis*, Marcel Dekker, Inc., New York.

Blischke, W. R. & Murthy, D. N. P. (1996), *Product Warranty Handbook*, Marcel Dekker, Inc., New York.

Chattopadhyay, G. & Rahman, A. (2008), 'Development of lifetime warranty policies and models for estimating costs', *Reliability Engineering and System Safety* **93**, 522–529.

Chien, Y. H. (2008), 'A general age-replacement model with minimal repair under renewing free-replacement warranty', *European Journal of Operational Research* **186**, 1046–1058.

Chukova, S. & Dimitrov, B. (1996), Warranty analysis for complex systems, *in* W. R. Blischke & D. N. P. Murthy, eds, 'Product Warranty Handbook', Chukova-Dimitrov, New York.

Duchesne, T. & Marri, F. (2009), 'General distributional properties of discounted warranty costs with risk adjustment under minimal repair', *IEEE Transactions on Reliability* **58**, 143–151.

Finkelstein, M. S. (2004), 'Minimal repair in heterogeneous populations', *Journal of Applied Probability* **41**, 281–286.

Fleming, T. R. & Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, John Wiley and Sons, Inc., New York.

González, N. G. (2009), Processos de burn-in e de garantia em sistemas coerentes sob o modelo de tempo de vida geral, Tese De Doutorado, Universidade de Sao Paulo, Instituto de Matemática e Estatística.

Hong-Zhong, H., Zhie-Jie, L., Yanfeng, L., Yu, L. & Liping, H. (2008), 'A warranty cost model with intermittent and heterogeneous usage', *Eksploatacja I Niezawodnosc-Maintenance and Reliability* **40**, 9–15.

Hussain, A. Z. M. O. & Murthy, D. N. P. (1998), 'Warranty and redundancy design with uncertain quality', *IIE Transactions* **30**, 1191–1199.

Ja, S. S., Kulkarni, V., Mitra, A. & Patankar, G. (2001), 'A non renewable minimal-repair warranty policy with time-dependent costs', *IEEE Transactions on Reliability* **50**, 346–352.

Ja, S. S., Kulkarni, V., Mitra, A. & Patankar, G. (2002), 'Warranty reserves for nonstationary sales processes', *Naval Research Logistics* **49**, 499–513.

Jack, N. & Murthy, D. N. P. (2007), 'A flexible extended warranty an related optimal strategies', *Journal of the Operational Research Society* **58**, 1612–1620.

Jain, M. & Maheshwari, S. (2006), 'Discounted costs for repairable units under hybrid warranty', *Applied Mathematics and Computation* **173**, 887–901.

Jung, K. M., Park, M. & Park, D. H. (2010), 'System maintenance cost depend on life cycle under renewing warranty policy', *Reliability Engineering and System Safety* **95**, 816–821.

Karr, A. F. (1986), *Point Processes and their Statistical Inference*, Marcel Dekker, Inc., New York.

Lehmann, E. L. (1999), *Elements of Large-Sample Theory*, Springer-Verlag, Inc., New York.

Li, H. & Shaked, M. (2003), 'Imperfect repair models with preventive maintenance', *Journal of Applied Probability* **40**, 1043–1059.

Lindqvist, B. H. (2006), 'On the statistical modeling and analysis of repairable systems', *Statistical Science* **21**, 532–551.

Lipster, R. S. & Shiryaev, A. N. (2001), *Statistics of Random Processes I. General Theory*, 2 edn, Springer-Verlag, Inc., New York.

Mamer, J. W. (1987), 'Discounted and per unit costs of product warranty', *Management Science* **33**, 916–930.

Mamer, W. (1969), 'Determination of warranty reserves', *Management Science* **15**, 542–549.

Mitra, A. & Patankar, J. G. (1993), 'Market share and warranty costs for renewable warranty programs', *International Journal of Production Economics* **20**, 111–123.

Murthy, D. N. P. (1990), 'Optimal reliability choice in product design', *Engineering Optimization* **15**, 280–294.

Nguyen, D. G. & Murthy, D. N. P. (1984), 'Cost analysis of warranty policies', *Naval Research Logistics Quarterly* **31**, 525–541.

Patankar, J. & Mitra, A. (1995), 'Effect of warranty execution on warranty reserve costs', *Management Science* **41**, 395–400.

Ritchken, P. H. (1986), 'Optimal replacement policies for irreparable warranty item', *IEEE Transactions on Reliability* **35**, 621–624.

Samatliy-Pac, G. & Taner, M. R. (2009), 'The role of repair strategy in warranty cost minimization: An investigation via quasi-renewal processes', *European Journal of Operational Research* **197**, 632–641.

Sheu, S. H. & Yu, S. L. (2005), 'Warranty strategy accounts for bathtub failure rate and random minimal repair cost', *An International Journal Computers & Mathematics with Applications* **49**, 1233–1242.

Thomas, M. U. (1989), 'A prediction model of manufacturer warranty reserves', *Management Science* **35**, 1515–1519.

Yeo, W. M. & Yuan, X. M. (2009), 'Optimal warranty policies for systems with imperfect repair', *European Journal of Operational Research* **199**, 187–197.

Yun, W. Y., Murthy, D. N. P. & Jack, N. (2008), 'Warranty servicing with imperfect repair', *International Journal of Production Economics* **111**, 159–169.

Appendix A. Proof of Theorem 2

If $i \in \mathscr{C}^{\Phi}(\omega)$, from Proposition 1 and the martingale property we have

$$E[\widehat{B}_t^i | S_i > Y_i] = E\Big[\int_0^t H_i(s) d\widetilde{N}_s^i \Big| S_i > Y_i\Big] = E\Big[\int_{Y_i}^t H_i(s)\lambda^i(s) ds \Big| S_i > Y_i\Big]$$

Since the sequences $(\widehat{B}_t^{i(j)}, C^{i(j)}, 1 \leq i \leq m)$, $1 \leq j \leq n$, are independent and identically distributed copies of the m-variate process $(\widehat{B}_t^i, C^i, 1 \leq i \leq m)$, from (32) we have

$$E[\overline{\widehat{B}}_t^{i(n)}] = \frac{1}{n}\sum_{j=1}^{n} P(S_i > Y_i) E\Big[\int_{Y_i}^t H_i(s)\lambda^i(s) ds \Big| S_i > Y_i\Big]$$

and therefore, $E[\overline{\widehat{B}}_t^{i(n)}] = \frac{1}{n}\sum\limits_{j=1}^{n} B^{i*}(t) = B^{i*}(t) = E[\widehat{B}_t^{i(n)}]$.

To set the consistency of proposed estimator we have to prove that

$$E[\sup_{s \leq t}(\overline{\widehat{B}}_s^{i(n)} - B^{i*}(s))^2] \xrightarrow[n\uparrow\infty]{} 0 \tag{58}$$

First, from (32) and Proposition 1 and for fixed n we have

$$\overline{\widehat{B}}_t^{i(n)} - \overline{B}_t^{i(n)} = \frac{1}{n}\sum_{j=1}^{n} C^{i(j)}(\widehat{B}_t^{i(j)} - B_t^{i(j)}) = \frac{1}{n}\sum_{j=1}^{n} C^{i(j)}(\widehat{B}_t^{i(j)} - B_t^{i*(j)}) \tag{59}$$

is a mean zero and square integrable $(P, \mathcal{F}_t)$-martingale. Furthermore, from the independence conditions and (23) we have

$$\langle \overline{\widehat{B}}^{i(n)} - \widehat{B}^{i(n)} \rangle_t = \frac{1}{n} \times \Big[\frac{1}{n}\sum_{j=1}^{n} C^{i(j)} \int_{Y_i^{(j)}}^t H_i^2(s)\lambda^i(s) ds\Big] \tag{60}$$

By hypothesis, for each $i = 1, \ldots, m$

$$E\Big[C^i \int_{Y_i}^t H_i^2(s)\lambda^i(s) ds\Big] = P(S_i > Y_i) E\Big[\int_{Y_i}^t H_i^2(s)\lambda^i(s) ds \Big| S_i > Y_i\Big] < \infty \tag{61}$$

and, therefore, using the Strong Law of Large Numbers we have

$$\frac{1}{n}\sum_{j=1}^{n} C^{i(j)} \int_{Y_i^{(j)}}^t H_i^2 \lambda^i(s) ds \xrightarrow[n\uparrow\infty]{} P(S_i > Y_i) E\Big[\int_{Y_i}^t H_i^2(s)\lambda^i(s) ds \Big| S_i > Y_i\Big] \tag{62}$$

Using (60) and (62) we conclude that

$$\langle \overline{\widehat{B}}^{i(n)} - \overline{B}^{i(n)} \rangle_t \xrightarrow[n\to\infty]{} 0 \times P(S_i > Y_i) E\Big[\int_{Y_i}^t H_i^2(s)\lambda^i(s) ds \Big| S_i > Y_i\Big] = 0 \tag{63}$$

Furthermore, we have (Lipster & Shiryaev 2001, Theorem 2.4)

$$E[\sup_{s\leq t}(\overline{\widehat{B}}_s^{i(n)} - \widehat{B}_s^{i(n)})^2] \leq 4E[(\overline{\widehat{B}}_t^{i(n)} - \overline{B}_t^{i(n)})^2] = 4E[\langle \overline{\widehat{B}}^{i(n)} - \overline{B}^{i(n)} \rangle_t] \tag{64}$$

where the last equality is because $(\overline{\widehat{B}}_t^{i(n)} - \overline{B}_t^{i(n)})$ is a mean zero and square integrable $(P, \mathcal{F}_t)$-martingale. From (63) and (64), we have

$$E[\sup_{s\leq t}(\overline{\widehat{B}}_s^{i(n)} - \overline{B}_s^{i(n)})^2] \xrightarrow[n\uparrow\infty]{} 0 \tag{65}$$

Also, from the Strong Law of Large Numbers and continuity in t, we get

$$(\overline{B}_s^{i(n)} - B^{i*}(s)) \xrightarrow[n\to\infty]{} 0, \quad \forall s \leq t \text{ and therefore, } \sup_{s\leq t} |\overline{B}_s^{i(n)} - B^{i*}(s)| \xrightarrow[n\to\infty]{} 0$$

then, we conclude

$$\sup_{s\leq t}(\overline{B}_s^{i(n)} - B^{i*}(s))^2 \xrightarrow[n\to\infty]{} 0$$

and

$$E[\sup_{s\leq t}(\overline{B}_s^{i(n)} - B^{i*}(s))^2] \xrightarrow[n\to\infty]{} 0 \tag{66}$$

Furthermore, we have

$$E[\sup_{s\leq t}(\overline{\widehat{B}}_s^{i(n)} - B^{i*}(s))^2] \leq E[\sup_{s\leq t}(\overline{\widehat{B}}_s^{i(n)} - \overline{B}_s^{i(n)})^2] + E[\sup_{s\leq t}(\overline{B}_s^{i(n)} - B^{i*}(s))^2]$$

and taking limits in the above inequality, from (65) and (66) we get

$$\lim_{n\to\infty} E[\sup_{s\leq t}(\overline{\widehat{B}}_s^{i(n)} - B^{i*}(s))^2] = 0 \tag{67}$$

and (58) is proved.

Appendix B. Proof of Proposition 2

First, as the sequences $(\widehat{B}_t^{i(j)}, C^{i(j)}, 1 \leq j \leq n)$ are independent and identically distributed copies of $(\widehat{B}_t^i, C^i)$, we have that, for all $t \geq 0$ and $i = 1, \dots, m$,

$$E[\overline{B}_t^{i(n)}] = \frac{1}{n}\sum_{j=1}^{n} P(S_i > Y_i) E\left[\int_{Y_i}^{t} H_i(s)\lambda^i(s)ds \middle| S_i > Y_i\right] = B^{i*}(t)$$

and $\overline{B}_t^{i(n)}$ is an unbiased estimator for $B^{i*}(t)$.

Furthermore, from the Strong Law of Large Numbers, $\overline{B}_t^{i(n)}$ converges almost surely to $B^{i*}(t)$.

Otherwise, as $(\mathbf{Z}_t^{(n)})_{n\geq 1}$ is a sequence of independent and identically distributed random vectors and the components do not have simultaneous failures, the processes $\overline{B}_t^{i(n)}$ and $\overline{B}_t^{j(n)}$ are uncorrelated in $[0,t]$, for all i, j, $i \neq j$, all n and

$$COV[Z_t^{i(n)}, Z_t^{j(n)}] = COV[\sqrt{n}\ \overline{B}_t^{i(n)}, \sqrt{n}\ \overline{B}_t^{j(n)}] = COV[C^i B_t^i, C^j B_t^j] = 0 \quad (68)$$

Consequently

$$\mathrm{Var}[Z_t^{i(n)}] = \mathrm{Var}[\sqrt{n}\ \overline{B}_t^{i(n)}] = \mathrm{Var}[C^i B_t^i] = \sigma^{2i*}(t) \quad (69)$$

Therefore, applying the Central Limit Theorem for a sequence of independent and identically distributed random vectors with mean $\boldsymbol{\mu}(t) = (B^{1*}(t), \dots, B^{m*}(t))$ and finite covariance matrix $\boldsymbol{\Sigma}(t)$, where $\Sigma_{ij}(t) = 1_{\{i=j\}}\sigma^{2i*}(t)$, we obtain that $\mathbf{Z}_t^{(n)} \xrightarrow[n\to\infty]{\mathcal{D}} \mathbf{Z}_t$, where $\mathbf{Z}_t$ is an m-variate Normal random vector with mean zero and covariance matrix $\boldsymbol{\Sigma}(t)$. In what follows we prove the convergence of the finite-dimensional distributions of the process $\mathbf{Z}^{(n)}$. For that we consider:

(a) Since $\forall\ t \geq 0, n \geq 1, i \neq j$, $COV[Z_t^{i(n)}, Z_t^{j(n)}] = COV[C^i B_t^i, C^j B_t^j] = 0$, we have $\forall\ t_k \leq t_l$, $t_k, t_l \in [0,t]$, $COV[Z_{t_k}^{i(n)}, Z_{t_l}^{j(n)}] = COV[C^i B_{t_k}^i, C^j B_{t_l}^j] = 0$;

(b) From the above, we can prove the convergence of the finite-dimensional distributions of the process $\mathbf{Z}^{(n)}$ using the Crámer-Wold procedure: proving the convergence for each component $Z^{i(n)}$, for all $0 \leq t_1 \leq t_2 \leq \cdots \leq t_k \leq t$, we prove that $\forall\ a_{il}$ arbitrary constants,

$$\sum_{i=1}^{m}\sum_{l=1}^{k-1} a_{il}(Z_{t_{l+1}}^{i(n)} - Z_{t_l}^{i(n)}) \xrightarrow[n\to\infty]{\mathcal{D}} \sum_{i=1}^{m}\sum_{l=1}^{k-1} a_{il}(Z_{t_{l+1}}^{i} - Z_{t_l}^{i}) \quad (70)$$

which, using the Cramer-Wold procedure, is equivalent to

$$(\mathbf{Z}_{t_1}^{(n)}, \mathbf{Z}_{t_2}^{(n)}, \dots, \mathbf{Z}_{t_k}^{(n)}) \xrightarrow[n\to\infty]{\mathcal{D}} (\mathbf{Z}_{t_1}, \mathbf{Z}_{t_2}, \dots, \mathbf{Z}_{t_k})$$

Now, for each i, $1 \leq i \leq m$ and $t_1 \leq t_2 \in [0,t]$, consider n independent and identically distributed copies of $(C^i B_{t_1}^i, C^i B_{t_2}^i)$. Then, for each n and $i = 1, \dots, m$ we get the random vector $(Z_{t_1}^{i(n)}, Z_{t_2}^{i(n)})$. Therefore

$$E(Z_{t_1}^{i(n)}, Z_{t_2}^{i(n)}) = (0,0),\ \forall\ n \geq 1, i = 1, \dots, m.$$

Furthermore, since the copies $C^{i(j)} B_{t_1}^{i(j)} B_{t_2}^{i(j)}$, $j = 1, \dots, n$ are independent and identically distributed and as, for independent copies j and k, the random variables $C^{i(j)} B_{t_1}^{i(j)}$ and $C^{i(k)} B_{t_2}^{i(k)}$ are also independent, we have

$$COV[Z_{t_1}^{i(n)}, Z_{t_2}^{i(n)}] = E[C^i B_{t_1}^i B_{t_2}^i] - B^{i*}(t_1)B^{i*}(t_2) = \sigma^{i*}(t_1, t_2) < \infty. \quad (71)$$

From (69), $\mathrm{Var}[Z_{t_1}^{i(n)}] = \sigma^{2i*}(t_1)$ and $\mathrm{Var}[Z_{t_2}^{i(n)}] = \sigma^{2i*}(t_2)$. Then, from the Central Limit Theorem for a sequence of independent and identically distributed random vectors, with finite mean vector and finite covariance matrix, we have

$$(Z_{t_1}^{i(n)}, Z_{t_2}^{i(n)}) \xrightarrow[n\to\infty]{\mathcal{D}} (Z_{t_1}^{i}, Z_{t_2}^{i}),\ \forall\ t_1 \le t_2 \in [0,t] \tag{72}$$

where $(Z_{t_1}^{i}, Z_{t_2}^{i})$ is a bivariate normal vector with mean zero and covariance matrix $\mathbf{\Sigma}^i(t_1,t_2)$,

$$\mathbf{\Sigma}^i(t_1,t_2) = \begin{bmatrix} \sigma^{2i*}(t_1) & \sigma^{i*}(t_1,t_2) \\ \sigma^{i*}(t_1,t_2) & \sigma^{2i*}(t_2) \end{bmatrix} \tag{73}$$

Using an induction argument we can generalize the above result for all partition $0 \le t_1 \le t_2 \le \cdots \le t_k \le t$, of the interval $[0,t]$ and we get for all $i,\ \ 1 \le i \le m$,

$$(Z_{t_1}^{i(n)}, Z_{t_2}^{i(n)}, \dots, Z_{t_k}^{i(n)}) \xrightarrow[n\to\infty]{\mathcal{D}} (Z_{t_1}^{i}, Z_{t_2}^{i}, \dots, Z_{t_k}^{i})$$

where $(Z_{t_1}^{i}, Z_{t_2}^{i}, \dots, Z_{t_k}^{i})$ is a k-variate Normal vector with mean zero and finite covariance matrix.

Finally, we analyze Stone's tightness condition in $D[0,t]^m$ (Fleming & Harrington 1991), that is: If for each $i, 1 \le i \le m$ and for all $\epsilon > 0$,

$$\lim_{\delta\downarrow 0} \limsup_{n\uparrow 0} P\Big\{ \sup_{\substack{|s-u|<\delta \\ 0\le s,u\le t}} \Big| Z_s^{i(n)} - Z_u^{i(n)} \Big| > \epsilon \Big\} = 0 \tag{74}$$

Since $Z_s^{i(n)}$ is continuous and monotone in $[0,t]$, we have

$$\begin{aligned} &P\Big\{ \sup_{\substack{|s-u|<\delta \\ 0\le s,u\le t}} \Big| Z_s^{i(n)} - Z_u^{i(n)} \Big| \le \epsilon \Big\} \\ &\quad \le P\Big\{ \Big| Z_s^{i(n)} - Z_u^{i(n)} \Big| \le \epsilon, \text{ for } s \text{ and } u \text{ fixed: } 0 \le s,u \le t, |s-u| < \delta \ \Big\} \end{aligned} \tag{75}$$

From (72) and (73), for all $0 \le s \le u$

$$(Z_s^{i(n)} - Z_u^{i(n)}) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, \gamma^2(s,u)),\ \gamma^2(s,u) = \sigma^{2i*}(s) + \sigma^{2i*}(u) - 2\sigma^{i*}(s,u). \tag{76}$$

Finally, from (69) and (71) it is clear that $\lim_{\delta\downarrow 0} \gamma^2(s,u) = 0,\ |s-u| < \delta,\ 0 \le s,u \le t$. Then, from (75) we have

$$\begin{aligned} \lim_{\delta\downarrow 0} \lim_{n\to\infty} P\Big\{ \sup_{\substack{|s-u|<\delta \\ 0\le s,u\le t}} \Big| Z_s^{i(n)} - Z_u^{i(n)} \Big| \le \epsilon \Big\} &\le \lim_{\delta\downarrow 0} 2\Phi\left(\frac{\epsilon}{\sqrt{\gamma^2(s,u)}} \right) - 1 \\ &= 2\Phi(\infty) - 1 = 1 \qquad \square \end{aligned}$$

Appendix C. Proof of Theorem 4

From Theorem 2 we have $E[\mathbf{E}_t^{(n)}]=0$ for all $n\geq 1$ and $t\geq 0$.

Let $\mathbf{M}_t^{(n)}=\sqrt{n}(\overline{\widehat{\mathbf{B}}}_t^{(n)}-\overline{\mathbf{B}}_t^{(n)})$ and $\mathbf{Z}_t^{(n)}=\sqrt{n}(\overline{\mathbf{B}}_t^{(n)}-\boldsymbol{\mu}(t))$. Note that

$$\mathbf{E}_t^{(n)}=\sqrt{n}(\overline{\widehat{\mathbf{B}}}_t^{(n)}-\boldsymbol{\mu}(t))=\mathbf{M}_t^{(n)}+\mathbf{Z}_t^{(n)} \tag{77}$$

Now, for all $t\geq 0$ and $1\leq i\leq m$ we are going to calculate the asymptotic variance for the processes $E_t^{i(n)}=\sqrt{n}(\overline{\widehat{B}}_t^{i(n)}-B^{i*}(t))=M_t^{i(n)}+Z_t^{i(n)}$. For fixed t,

$$\text{Var}[E_t^{i(n)}]=\text{Var}[M_t^{i(n)}]+\text{Var}[Z_t^{i(n)}]+2\;COV[M_t^{i(n)},Z_t^{i(n)}] \tag{78}$$

Since the copies are independent and identically distributed, from Corollary 3 and for all $t\geq 0$, we have that $\text{Var}[M_t^{i(n)}]$ corresponds to

$$E[\langle M^{i(n)}\rangle_t]=E\Big[\frac{1}{n}\sum_{j=1}^{n}C^{i(j)}\int_{Y_i^{(j)}}^{t}H_i^2(s)\lambda^i(s)ds\Big]=E[C^i(\widehat{B}_t^i-B_t^i)^2]=V_i^*(t); \tag{79}$$

and $\text{Var}[Z_t^{i(n)}]$ is given by (69).

In order to calculate $COV[M_t^{i(n)},Z_t^{i(n)}]$, we use the covariance definition, the martingale property, the fact that for independent copies j and l, $C^{i(j)}\widehat{B}_t^{i(j)}$ and $C^{i(l)}B_t^{i(l)}$ are also independent, concluding

$$COV[M_t^{i(n)},Z_t^{i(n)}]=E[C^i\widehat{B}_t^iB_t^i]-E[C^i(B_t^i)^2] \tag{80}$$

Therefore, from (69), (79) and (80), we obtain in (78) that, for all $n\geq 1$ and $t\geq 0$

$$\text{Var}[E_t^{i(n)}]=E[C^i(\widehat{B}_t^i)^2]-(B^{i*}(t))^2$$

In addition,we have $E[C^i\widehat{B}_t^i]=B^{i*}(t)$ and then,

$$\text{Var}[E_t^{i(n)}]=\text{Var}[C^i\widehat{B}_t^i]=\delta^{2i*}(t) \tag{81}$$

We also calculate $COV[E_t^{i(n)},E_t^{j(n)}]$ for $n\geq 1$ and $i\neq j$, and since processes $\widehat{B}_t^i$ and $\widehat{B}_t^j$ do not have simultaneous jumps, we obtain,

$$COV[E_t^{i(n)},E_t^{j(n)}]=COV[C^i\widehat{B}_t^i,C^j\widehat{B}_t^j]=0 \tag{82}$$

From results (81) and (82) we conclude that the asymptotic covariance for the process $\mathbf{E}_t^{(n)}$ is $\mathbf{U}(t)$ where $U_{ij}(t)=1_{\{i=j\}}\delta^{2i*}(t)$. Next, we set the asymptotic normality of $\mathbf{E}_t^{(n)}$ by considering the results from its asymptotic covariance structure and the convergence in distribution of the processes $\mathbf{M}_t^{(n)}$ (Corollary 3) and $\mathbf{Z}_t^{(n)}$ (Proposition 2):

As the processes $\mathbf{M}^{(n)}$ and $\mathbf{Z}^{(n)}$ satisfy the tightness condition in $D[0,t]^m$ and their finite-dimensional distributions converge to Gaussian continuous processes, such that $\forall\ t_k, t_l \in [0,t]$, $COV[E^{i(n)}_{t_k}, E^{j(n)}_{t_l}] = 0$, the process $\mathbf{E}^{(n)} = \mathbf{M}^{(n)} + \mathbf{Z}^{(n)}$ also satisfies the tightness condition.

Also, its finite-dimensional distributions converge to Gaussian continuous processes and for all partition $0 \le t_1 \le t_2 \le \cdots \le t_k \le t$, we can prove that,

$\forall\ a_{il}$ arbitrary constants,

$$\sum_{i=1}^{m}\sum_{l=1}^{k-1} a_{il}(E^{i(n)}_{t_{l+1}} - E^{i(n)}_{t_l}) = \sum_{i=1}^{m}\sum_{l=1}^{k-1} a_{il}(M^{i(n)}_{t_{l+1}} - M^{i(n)}_{t_l}) + \sum_{i=1}^{m}\sum_{l=1}^{k-1} a_{il}(Z^{i(n)}_{t_{l+1}} - Z^{i(n)}_{t_l})$$

$$\xrightarrow[n\to\infty]{\mathcal{D}}$$

$$\sum_{i=1}^{m}\sum_{l=1}^{k-1} a_{il}(M^{i}_{t_{l+1}} - M^{i}_{t_l}) + \sum_{i=1}^{m}\sum_{l=1}^{k-1} a_{il}(Z^{i}_{t_{l+1}} - Z^{i}_{t_l}) = \sum_{i=1}^{m}\sum_{l=1}^{k-1} a_{il}(W^{i}_{t_{l+1}} - W^{i}_{t_l}) \ \square$$

which, using the Cramer-Wold procedure, is equivalent to

$$(\mathbf{E}^{(n)}_{t_1}, \mathbf{E}^{(n)}_{t_2}, \ldots, \mathbf{E}^{(n)}_{t_k}) = (\mathbf{M}^{(n)}_{t_1}, \mathbf{M}^{(n)}_{t_2}, \ldots, \mathbf{M}^{(n)}_{t_k}) + (\mathbf{Z}^{(n)}_{t_1}, \mathbf{Z}^{(n)}_{t_2}, \ldots, \mathbf{Z}^{(n)}_{t_k}) \xrightarrow[n\to\infty]{\mathcal{D}}$$
$$(\mathbf{M}_{t_1}, \mathbf{M}_{t_2}, \ldots, \mathbf{M}_{t_k}) + (\mathbf{Z}_{t_1}, \mathbf{Z}_{t_2}, \ldots, \mathbf{Z}_{t_k}) = (\mathbf{W}_{t_1}, \mathbf{W}_{t_2}, \ldots, \mathbf{W}_{t_k})$$

Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 545 a 566

On Certain Properties of A Class of Bivariate Compound Poisson Distributions and an Application to Earthquake Data

Ciertas propiedades de una clase de distribuciones Poisson compuesta bivariadas y una aplicación a datos de terremotos

Gamze Özel[a]

Department of Statistics, Hacettepe University, Ankara, Turkey

Abstract

The univariate compound Poisson distribution has many applications in various areas such as biology, seismology, risk theory, forestry, health science, etc. In this paper, a bivariate compound Poisson distribution is proposed and the joint probability function of this model is derived. Expressions for the product moments, cumulants, covariance and correlation coefficient are also obtained. Then, an algorithm is prepared in Maple to obtain the probabilities quickly and an empirical comparison of the proposed probability function is given. Bivariate versions of the Neyman type A, Neyman type B, geometric-Poisson, Thomas distributions are introduced and the usefulness of these distributions is illustrated in the analysis of earthquake data.

Key words**:** Bivariate distribution, Coefficient of correlation, Compound Poisson distribution, Cumulant, Moment.

Resumen

La distribución compuesta de Poisson univariada tiene muchas aplicaciones en diversas áreas tales como biología, ciencias de la salud, ingeniería forestal, sismología y teoría del riesgo, entre otras. En este artículo, una distribución compuesta de Poisson bivariada es propuesta y la función de probabilidad conjunta de este modelo es derivada. Expresiones para los momentos producto, acumuladas, covarianza y el coeficiente de correlación respectivos son obtenidas. Finalmente, un algoritmo preparado en lenguaje Maple es descrito con el fin de calcular probabilidades asociadas rápidamente y con el fin de hacer una comparación de la función de probabilidad propuesta. Se introducen además versiones bivariadas de las distribuciones tipo A y tipo B de Neyman, geométrica-Poisson y de Thomas y se ilustra la utilidad de estas distribuciones aplicadas al análisis de datos de terremoto.

Palabras clave**:** coeficiente de correlación, conjuntas, distribución bivariada, distribución compuesta de Poisson, momento.

[a] Doctor. E-mail: gamzeozl@hacettepe.edu.tr

1. Introduction

Bivariate discrete random variables taking integer non-negative values, have received considerable attention in the literature, in an effort to explain phenomena in various areas of application. For an extensive account of bivariate discrete distributions one can refer to the books by Kocherlakota & Kocherlakota (1992), Johnson, Kotz & Balakrishnan (1997) and the review articles by Papageorgiou (1997) and Kocherlakota & Kocherlakota (1997). There is however, a variety of applications, e.g. in an accident or family studies (see Cacoullos & Papageorgiou 1980, Sastry 1997). The bivariate Poisson distribution (BPD) is probably the best known bivariate discrete distribution (Holgate 1964). It is appropriate for modeling paired count data exhibiting correlation. Paired count data arise in a wide context including marketing (number of purchases of different products), epidemiology (incidents of different diseases in a series of districts), accident analysis (number of accidents in a site before and after infrastructure changes), medical research (the number of seizures before and after treatment), sports (the number of goals scored by each one of the two opponent teams in soccer), econometrics (number of voluntary and involuntary job changes).

Bivariate compound distributions can be especially used in actuarial science to model a business book containing bivariate claim count distributions and bivariate claims severities (Ambagaspitiya 1998). In most actuarial studies, the assumption of independence between classes of business in an insurance business book containing is made. However this assumption is not verified in practice. For example, in the case of a catastrophe such as an earthquake, the damages covered by homeowners and private passenger automobile insurance can not be considered independent (Cossette, Gaillardetz, Marceau & Rioux 2002). In this situation, bivariate compound Poisson distribution (BCPD) is useful when the claim count distribution is bivariate Poisson and the claim size distribution is bivariate.

Although the case of BPD has attracted some attention in the literature, BCPD has not been systematically studied. The studies on such a distribution are sparse due to computational problems involved in its implementation. Hesselager (1996) studied the BCPD but mainly from the recursive evaluation of its joint probability function. On the other hand, non-existence of explicit probabilities and algorithm of the BCPD hinders its use in probability theory itself and its applications in seismology, actuarial science, survival analysis, etc. (see Ozel & Inal 2008, Wienke, Ripatti, Palmgren & Yashin 2010). Consequently, since relative results are sparse and case oriented, the aim of this study is to obtain a general technique for deriving the probabilistic characteristics and obtain an algorithm for the computation of probabilities.

The rest of the paper is organised as follows. In Section 2, some preliminary results are given. In Section 3, the probabilistic characteristics of the BCPD are proposed based on the derivation of the joint probability generating function (pgf). This pgf enables us to obtain the joint probability function of the BCPD. In addition, explicit expressions for the product moments, cumulants, covariance and correlation coefficient are obtained. Then numerical examples and an application to earthquakes in Turkey are presented in Section 4, by means of the proposed algorithm in Maple. The conclusion is given in Section 5.

2. Some Preliminary Results

Let N be a Poisson random variable with parameter $\lambda > 0$ and let X_i, $i = 1, 2, \dots$ be i.i.d. non-negative, integer-valued random variables, independent of N. S has a compound Poisson distribution (CPD), when defined as

$$S = \sum_{i=1}^{N} X_i \tag{1}$$

If $E(X)$ and $V(X)$ are the common mean and variance of the random variables X_1, $i = 1, 2, \dots$, then, the moments of S are given by

$$E(S) = \lambda E(X), \quad V(S) = \lambda[V(X) + [E(X)]^2] \tag{2}$$

The probability function of S is given by

$$p_S(s) = P(S = s) = \sum_{n=0}^{\infty} P(X_1 + X_2 + \cdots + X_n = s \mid N = n) P(N = n), \quad s = 0, 1, 2 \dots \tag{3}$$

However, it is not easy to yield an explicit formula for the probability function of S from (3), and this obstructs use of the CPD completely (see, for example Bruno, Camerini, Manna & Tomassetti 2006, Rolski, Schmidli, Schmidt & Teugels 1999). Panjer (1981) described a procedure for recursive evaluation of the CPD when N is Poisson distributed.

Let N be a Poisson distributed random variable with parameter λ and let S be a compound Poisson distributed random variable. Panjer (1981) showed that when N satisfies a recursion in the form $p_N(n) = \frac{\lambda}{n} p_N(n-1)$, $n = 1, 2, 3 \dots$ than S satisfies

$$\begin{aligned} p_S(0) &= e^{-\lambda[1-p_X(0)]} \\ p_S(s) &= \lambda \sum_{i=1}^{s} \frac{i}{s} p_X(i) p_S(s-i), \quad s = 1, 2, 3 \dots \end{aligned} \tag{4}$$

where $p_X(x)$ is the common probability function of X_i, $i = 1, 2, 3 \dots$ Since (4) is based on a recursive scheme, it causes difficulties in computation time and computer memory for the large values of s (Rolski et al. 1999). The explicit probabilities of S are obtained by Ozel & Inal (2010) as in (6) by using (5).

Let X_i, $i = 1, 2, 3 \dots$, be i.i.d. discrete random variables with the probabilities $P(X_i = j) = p_j$, $j = 0, 1, 2 \dots$ and let define the parameters $\lambda_j = \lambda p_j$. The common probability generating function (pgf) of X_i, $i = 1, 2, 3 \ldots$, is given by $g_X(s) = \sum_{j=0}^{\infty} p_j s^j = p_0 + p_1 s + p_2 s^2 + \cdots$ and the pgf of S is given by

$$\begin{aligned} g_S(z) &= \sum_{n=0}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} [g_X(z)]^n = e^{-\lambda} \left[1 + \frac{\lambda g_X(z)}{1!} + \frac{(\lambda g_X(z))^2}{2!} + \cdots \right] \\ &= e^{\lambda[g_X(z)-1]} = e^{\lambda[(p_0+p_1 z+\cdots+p_m z^m)-1]} \\ &= e^{-\lambda(1-p_0)} e^{\lambda_1 z+\lambda_2 z^2+\cdots+\lambda_m z^m} \end{aligned} \tag{5}$$

Let N be a Poisson distributed random variable with parameter $\lambda > 0$ and $\lambda_j = \lambda p_j$, $j = 1, 2, \ldots, m$. Then, the explicit formula for the probability function of S is determined by using (5) as follows:

$$\begin{aligned} P(S=0) &= e^{-\lambda(1-p_0)} \\ P(S=1) &= e^{-\lambda(1-p_0)} \frac{\lambda_1}{1!} \\ P(S=2) &= e^{-\lambda(1-p_0)} \left[\frac{\lambda_1^2}{2!} + \frac{\lambda_2}{1!} \right] \\ P(S=3) &= e^{-\lambda(1-p_0)} \left[\frac{\lambda_1^3}{3!} + \frac{\lambda_1 \lambda_2}{1!1!} + \frac{\lambda_3}{1!} \right] \\ P(S=4) &= e^{-\lambda(1-p_0)} \left[\frac{\lambda_1^4}{4!} + \frac{\lambda_1^2 \lambda_2}{2!1!} + \frac{\lambda_1 \lambda_3}{1!1!} + \frac{\lambda_2^2}{2!} + \frac{\lambda_4}{1!} \right] \\ P(S=5) &= e^{-\lambda(1-p_0)} \left[\frac{\lambda_1^5}{5!} + \frac{\lambda_1^3 \lambda_2}{3!1!} + \frac{\lambda_1^2 \lambda_3}{2!1!} + \frac{\lambda_1 \lambda_2^2}{1!2!} + \frac{\lambda_1 \lambda_4}{1!1!} + \frac{\lambda_2 \lambda_3}{1!1!} + \frac{\lambda_5}{1!} \right] \\ &\vdots \end{aligned} \tag{6}$$

According to the above probabilities for $s = 1, 2, \ldots$, the on the right terms depend on how s can be partitioned into different forms using integers $1, 2, \ldots$,m. For example, if $s = 5$, it is partitioned in seven ways and all the partitions of five are $\{1, 1, 1, 1, 1\}$, $\{1, 1, 1, 2\}$, $\{1, 2, 2\}$, $\{1, 1, 3\}$, $\{2, 3\}$, $\{1, 4\}$, $\{5\}$. Note that S has a Neyman type A distribution if X_i, $i = 1, 2, \ldots$ are Poisson distributed in (1). Similarly, if X_i, $i = 1, 2, \ldots$ are truncated Poisson distributed, S has a Thomas distribution. S has a Neyman type B distribution if X_i, $i = 1, 2, \ldots$, are binomial distributed. If X_i, $i = 1, 2, \ldots$ are geometric distributed, S has a geometric-Poisson (Pólya-Aeppli) distribution. Let us point out that (6) is also extended by Ozel & Inal (2011) for these special cases of the CPD and by Ozel & Inal (2008) for the compound Poisson process with an application for earthquakes in Turkey. There has also been an increasing interest in bivariate discrete probability distributions and many forms of these distributions have been studied (see, for example, Kocherlakota & Kocherlakota 1992, Johnson et al. 1997). The BPD has been constructed by Holgate (1964) as in (7) using the trivariate reduction method.

Let M_0, M_1, M_2 be independent Poisson variables with parameters $\lambda_0, \lambda_1, \lambda_2$, respectively. Then, $N_1 = M_0 + M_1$ and $N_2 = M_0 + M_1$ follow a BPD and the joint probability function is given by

$$p_{N_1,N_2}(n_1,n_2) = P(N_1 = n_1, N_2 = n_2) = \\ e^{-(\lambda_0+\lambda_1+\lambda_2)} \sum_{i=0}^{\min(n_1,n_2)} \frac{\lambda_1^{n_1-i}\lambda_2^{n_2-i}\lambda_0^{i}}{(n_1-i)!(n_2-i)!i!}, \quad n_1, n_1 = 0, 1, 2, \dots \tag{7}$$

The formula in (7), allows positive dependence between N_1 and N_2. Marginally, each random variable follows a Poisson distribution with $E(N_1) = V(N_1) = \lambda_0+\lambda_1$ and $E(N_2) = V(N_2) = \lambda_0 + \lambda_2$. Moreover, $Cov(N_1, N_2) = \lambda_0$, and hence λ_0 is a measure of dependence between the two random variables. Then, the correlation coefficient of N_1 and N_2 is given by

$$\rho = \frac{\lambda_0}{\sqrt{(\lambda_0 + \lambda_1)(\lambda_0 + \lambda_2)}}$$

This implies that $\lambda_0 = 0$ is a necessary and sufficient condition for N_1 and N_2 to be independent. Also, $\lambda_0 = 1$, if and only if, N_1 and N_2 are linearly dependent.

In Section 3, the concept of the CPD is extended to the bivariate case.

3. Main Results

3.1. The Joint Probability Function

Let M_0, M_1, M_2 be independent Poisson variables with parameters $\lambda_0, \lambda_1, \lambda_2$, respectively, and let $N_1 = M_0+M_1$, $N_2 = M_0+M_2$ be bivariate Poisson distributed random variables with parameters $\lambda_0+\lambda_1$ and $\lambda_0+\lambda_2$. Then, (S_1, S_2) has a BCPD when defined as

$$\left(S_1 = \sum_{i=1}^{N_1} X_i, S_2 \sum_{i=1}^{N_2} Y_i\right) \tag{8}$$

where X_i and Y_i, $i = 1, 2, \dots$ i.i.d. integer-valued random variables and independent of N_1 and N_2.

In particular, if X_i and Y_i, $i = 1, 2, \dots$ are Poisson distributed with parameters μ_1 and μ_2 in (8), S_1 and S_2 have a bivariate Neyman type A distribution. If X_i and Y_i, $i = 1, 2, \dots$ are binomial distributed with parameters (m_1, p_1) and (m_2, p_2), S_1 and S_2 have a bivariate Neyman type B distribution. Let X_i and Y_i, $i = 1, 2, \dots$ are truncated Poisson distributed with the probability functions $p_j = P(X_i = j) = e^{-\alpha_1}\frac{\alpha_1^{j-1}}{(j-1)!}$, $j = 1, 2, 3, \dots$ and $q_k = P(Y_i = k) = e^{-\alpha_2}\frac{\alpha_2^{j-1}}{(j-1)!}$, $k = 1, 2, 3, \dots$ for $\alpha_1, \alpha_2 > 0$, respectively. Then, the pair of (S_1, S_2) has a bivariate Thomas distribution. If X_i and Y_i, $i = 1, 2, \dots$ are geometric distributed with parameters θ_1 and θ_2, S_1 and S_2 have a bivariate geometric-Poisson distribution.

The joint probability function of S_1 and S_2 takes the following form

$$\begin{aligned} p_{S_1,S_2}(s_1,s_2) = \\ \sum_{n_1}^{\infty}\sum_{n_2}^{\infty} p(n_1,n_2)P(X_1+\cdots+X_{n_1}=s_1 \mid N_1=n_1) \\ P(Y_1+\cdots+Y_{n_2}=s_2 \mid N_2=n_2), \quad s_1,s_2=0,1,\dots \end{aligned} \tag{9}$$

where $p_{S_1,S_2}(s_1,s_2)=P(S_1=s_1,S_2=s_2)$. Since the probability function given in (9) contains a summation over i from 0 to ∞, it is not suitable to obtain probabilities quickly (Ambagaspitiya 1998). More generally, for large n_1 and n_2, it is difficult to use (9) because of the high order of convolutions involved.

Hesselager (1996), in his pioneering work on recursive computation of the bivariate compound distributions, considered three classes of Poisson distributions and related compound distributions. A brief description of related recursive relations is given as follows:

Let M_0, M_1, M_2 be independent Poisson variables with parameters $\lambda_0, \lambda_1, \lambda_2$. Let $p_X(x)$ and $p_Y(y)$ be the common probability function of X_i, Y_i, $i=1,2,\dots$, respectively. Then, the joint probability function of S_1 and S_2 satisfies the recursive relations

$$\begin{aligned} p_{S_1,S_2}(s_1,s_2) &= \frac{\lambda_1}{s_1}\sum_{x=1}^{s_1} x p_X(x) p_{S_1,S_2}(s_1-x,s_2) + \\ &\frac{\lambda_0}{s_1}\sum_{x=1}^{s_1}\sum_{y=0}^{s_2} x p_X(x) p_Y(y) p_{S_1,S_2}(s_1-x,s_2-y) \\ p_{S_1,S_2}(s_1,s_2) &= \frac{\lambda_2}{s_1}\sum_{x=1}^{s_1} y p_Y(y) p_{S_1,S_2}(s_1,s_2-y) + \\ &\frac{\lambda_0}{s_2}\sum_{x=0}^{s_1}\sum_{y=1}^{s_2} y p_X(x) p_Y(y) p_{S_1,S_2}(s_1-x,s_2-y) \\ &\qquad s_1,s_2=1,2,\dots \end{aligned} \tag{10}$$

Although the use of these recursions considerably reduces the number of computations to obtain probabilities $P(S_1=s_1,S_2=s_2)$, $s_1,s_2=0,1,2,\dots$ compared with the traditional method based on convolutions in (9), these computations are still time consuming since each probability depends on all the preceding ones. It occurs in underflow problems which are not always easy to overcome and therefore restrict its applicability further (Sundt 1992). Thus, it can be applied only in some practical circumtances or in an approximate manner.

Finally to establish the probabilistic characteristics of the BCPD. We first compute the joint pgf of S_1 and S_2 as follows:

Let X_i, Y_i, $i=1,2,\dots$ be i.i.d. discrete random variables with the probabilities $P(X_i=j)=p_j$, $j=0,1,2,\dots,m$ and $P(Y_i=k)=q_k$, $k=0,1,2,\dots,r$. Then, the joint pgf of S_1 and S_2 is found to be

$$
\begin{aligned}
g_{S_1,S_2}(z_1,z_2) &= \sum_{s_1}^{\infty}\sum_{s_2}^{\infty} P\left(\sum_{i=1}^{N_1} X_i = s_1, \sum_{i=1}^{N_2} Y_i = s_2\right) z_1^{s_1} z_2^{s_2} \\
&= \sum_{s_1}^{\infty}\sum_{s_2}^{\infty}\sum_{n_1}^{\infty}\sum_{n_2}^{\infty} \\
& P\left(\sum_{i=1}^{n_1} X_i = s_1, \sum_{i=1}^{n_2} Y_i = s_2 \mid N_1 = n_1, N_2 = n_2\right) \\
& p_{N_1,N_2}(n_1,n_2) z_1^{s_1} z_2^{s_2} \\
&= \sum_{n_1}^{\infty}\sum_{n_2}^{\infty} p_{N_1,N_2}(n_1,n_2) \sum_{s_1}^{\infty}\sum_{s_2}^{\infty} \\
& P\left(\sum_{i=1}^{n_1} X_i = s_1, \sum_{i=1}^{n_2} Y_i = s_2 \mid N_1 = n_1, N_2 = n_2\right) z_1^{s_1} z_2^{s_2}
\end{aligned}
$$

Since X_i, Y_i, $i = 1, 2, \ldots$ are i.i.d. random variables, we have

$$
\begin{aligned}
g_{S_1,S_2}(z_1,z_2) &= \sum_{n_1}^{\infty}\sum_{n_2}^{\infty} p_{N_1,N_2}(n_1,n_2) \sum_{s_1}^{\infty} P(X_1 + \cdots + X_{n_1} = s_1) z_1^{s_1} \\
& \sum_{s_2}^{\infty} P(Y_1 + \cdots + Y_{n_2} = s_1) z_2^{s_2} \\
&= \sum_{n_1}^{\infty}\sum_{n_2}^{\infty} p_{N_1,N_2}(n_1,n_2) g_{X_1+\cdots+X_{n_1}}(z_1) g_{Y_1+\cdots+Y_{n_2}}(z_2) \\
&= \sum_{n_1}^{\infty}\sum_{n_2}^{\infty} p_{N_1,N_2}(n_1,n_2) [g_X(z_1)]^{n_1} [g_Y(z_2)]^{n_2} \\
&= g_{N_1,N_2}[g_X(z_1), g_Y(z_2)]
\end{aligned}
\tag{11}
$$

where $g_X(z_1)$, $g_Y(z_2)$ are the common pgfs of X_i, Y_i, $i = 1, 2, \ldots$, respectively.

Let $N_1 = M_0 + M_1$, $N_2 = M_0 + M_2$ be a BPD with parameters $\lambda_0 + \lambda_1$ and $\lambda_0 + \lambda_2$, then the joint pgf of N_1 and N_2 is given by

$$
\begin{aligned}
g_{N_1,N_2}(z_1,z_2) &= g_{M_0+M_1,M_0+M_2}(z_1,z_2) \\
&= E(z_1^{M_0+M_1} z_2^{M_0+M_2}) \\
&= E(z_1^{M_1}) E(z_2^{M_2}) E(z_1 z_2)^{M_0} \\
&= \exp[\lambda_1(z_1 - 1) + \lambda_2(z_2 - 1) + \lambda_0(z_1 z_2 - 1)]
\end{aligned}
\tag{12}
$$

From (11) and (12), the joint pgf of S_1 and S_2 is obtained by the following expression

$$
\begin{aligned}
g_{S_1,S_2}(z_1,z_2) &= \exp\bigl(\lambda_1[g_X(z_1)-1]+\lambda_2[g_Y(z_2)-1] \\
&\quad+\lambda_0[g_X(z_1)g_Y(z_2)-1]\bigr) \\
&= \exp\bigl(\lambda_1[p_0+p_1z_1+p_2z_1^2+\cdots+p_mz_1^m-1] \\
&\quad+\lambda_2[q_0+q_1z_1+q_2z_2^2+\cdots+q_rz_2^r-1] \\
&\quad+\lambda_0\bigl[(p_0+p_1z_2+p_2z_1^2+\cdots+p_mz_1^m) \\
&\qquad(q_0+q_1z_2+q_2z_2^2+\cdots+q_rz_2^r)-1\bigr]\bigr) \\
&= \exp\bigl(-(\lambda_0+\lambda_1+\lambda_2)\bigr) \\
&\quad\exp\bigl(\lambda_1(p_0+p_1z_1+\cdots+p_mz_1^m) \\
&\quad+\lambda_2(q_0+q_1z_2+\cdots+q_rz_2^r) \\
&\quad+\lambda_0[(p_0+p_1z_1+\cdots+p_mz_1^m)(q_0+q_1z_2+\cdots+q_rz_2^r)]\bigr)
\end{aligned}
\tag{13}
$$

Now we are interested in studying the joint probability function of the pair S_1 and S_2. The joint pgf in (13) can be differentiated any number of times with respect to s_1 and s_2 and evaluated at $(0,0)$ yielding

$$
\begin{aligned}
P(S_1=0,S_2=0) &= g_{S_1,S_2}(0,0) \\
P(S_1=s_1,S_2=s_2) &= \frac{\left.\frac{\partial^{S_1+S_2}g_{S_1,S_2}(z_1,z_2)}{\partial z_1^{s_1}z_2^{s_2}}\right|_{z_1=z_2=0}}{s_1!s_2!}, \quad s_1s_2=0,1,2,\dots
\end{aligned}
\tag{14}
$$

Differentiating the joint pgf given by (13) and substituting in (14) and after some algebraic manipulations, the probabilities $p_{S_1,S_2}(s_1,s_2)=P(S_1=s_1,S_2=s_2)$, $s_1s_2=0,1,2,\dots$ are obtained as

$$
\begin{aligned}
p_{S_1,S_2}(0,0) &= e^{-(\lambda_0+\lambda_1+\lambda_2)}e^{(\lambda_1p_0+\lambda_2q_0+\lambda_0p_0q_0)} \\
p_{S_1,S_2}(1,0) &= p_{S_1,S_2}(0,0)\left[p_1\frac{\Lambda_x}{1!}\right] \\
p_{S_1,S_2}(2,0) &= p_{S_1,S_2}(0,0)\left[p_1^2\frac{\Lambda_x^2}{2!}+p_2\frac{\Lambda_x}{1!}\right] \\
p_{S_1,S_2}(3,0) &= p_{S_1,S_2}(0,0)\left[p_1^3\frac{\Lambda_x^3}{3!}+p_1p_2\frac{\Lambda_x^2}{2!}+p_3\frac{\Lambda_x}{1!}\right] \\
p_{S_1,S_2}(0,1) &= p_{S_1,S_2}(0,0)\left[q_1\frac{\Lambda_y}{1!}\right]
\end{aligned}
$$

$$
\begin{aligned}
p_{S_1,S_2}(0,2) &= p_{S_1,S_2}(0,0)\left[q_1^2\frac{\Lambda_y^2}{2!}+q_2\frac{\Lambda_y}{1!}\right]\\
p_{S_1,S_2}(0,3) &= p_{S_1,S_2}(0,0)\left[q_1^3\frac{\Lambda_y^3}{3!}+q_1q_2\frac{\Lambda_y^2}{2!}+q_3\frac{\Lambda_y}{1!}\right]\\
p_{S_1,S_2}(1,1) &= p_{S_1,S_2}(0,0)\left[p_1q_1\left(\frac{\Lambda_x\Lambda_y}{1!1!}+\lambda_0\right)\right]\\
p_{S_1,S_2}(1,2) &= p_{S_1,S_2}(0,0)\left[p_1q_1^2\left(\frac{\Lambda_x\Lambda_y^2}{1!2!}+\frac{\Lambda_y}{1!}\right)+p_1q_2\left(\frac{\Lambda_x\Lambda_y}{1!1!}+\lambda_0\right)\right]\\
p_{S_1,S_2}(1,3) &= p_{S_1,S_2}(0,0)\Bigg[p_1q_1^3\left(\frac{\Lambda_x\Lambda_y^3}{1!3!}+\frac{\Lambda_y^2}{2!}\right)+p_1q_1q_2\left(\frac{\Lambda_x\Lambda_y^2}{1!2!}+\frac{\Lambda_y}{1!}\right)\\
&\qquad +p_1q_3\left(\frac{\Lambda_x\Lambda_y}{1!1!}+\lambda_0\right)\Bigg]\\
p_{S_1,S_2}(2,1) &= p_{S_1,S_2}(0,0)\left[p_1^2q_1\left(\frac{\Lambda_x^2\Lambda_y}{1!2!}+\frac{\Lambda_x}{1!}\right)+p_2q_1\left(\frac{\Lambda_x\Lambda_y}{1!1!}+\lambda_0\right)\right]\\
p_{S_1,S_2}(2,2) &= p_{S_1,S_2}(0,0)\Bigg[p_1^2q_1^2\left(\frac{\Lambda_x^2\Lambda_y^2}{2!2!}+\frac{\Lambda_x\Lambda_y}{1!1!}+\lambda_0^2\right)\\
&\qquad +p_1^2q_2\left(\frac{\Lambda_x^2\Lambda_y}{2!1!}+\frac{\Lambda_x}{2!1!}\right)+p_2q_1^2\left(\frac{\Lambda_x\Lambda_y^2}{1!2!}+\frac{\Lambda_y}{1!}\right)\\
&\qquad +p_2q_2\left(\frac{\Lambda_x\Lambda_y}{1!1!}+\lambda_0\right)\Bigg]\\
p_{S_1,S_2}(2,3) &= p_{S_1,S_2}(0,0)\Bigg[p_1^2q_1^3\left(\frac{\Lambda_x^2\Lambda_y^3}{2!3!}+\frac{\Lambda_x\Lambda_y^2}{1!2!}+\frac{\Lambda_y}{1!}\right)\\
&\qquad +p_1^2q_1q_2\left(\frac{\Lambda_x^2\Lambda_y^2}{2!2!}+\frac{\Lambda_x\Lambda_y}{1!1!}+\lambda_0^2\right)\\
&\qquad +p_2q_1^3\left(\frac{\Lambda_x\Lambda_y^3}{3!1!}+\frac{\Lambda_y^2}{2!}\right)+p_2q_1q_2\left(\frac{\Lambda_x\Lambda_y^2}{1!2!}+\frac{\Lambda_y^2}{2!}+\frac{\Lambda_y}{1!}\right)\\
&\qquad +p_1^2q_3\left(\frac{\Lambda_x^2\Lambda_y}{2!1!}+\frac{\Lambda_x}{1!}\right)+p_2q_3\left(\frac{\Lambda_x\Lambda_y}{1!1!}+\lambda_0\right)\Bigg]
\end{aligned}
\tag{15}
$$

where $\Lambda_x = (\lambda_1 + \lambda_0 q_0)$ and $\Lambda_y = (\lambda_2 + \lambda_0 p_0)$. According to above probabilities $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 1, 2, 3, \dots$ the on the right side terms p_j, $j = 1, 2, \dots, m$ and q_k, $k = 1, 2, \dots, r$ depend on how s_1 and s_2 can be partitioned into different forms using integers $1, 2, \dots$ Similarly, the terms Λ_x and Λ_y also have an order related with the powers of p_j, $j = 1, 2, \dots, m$ and q_k, $k = 1, 2, \dots, r$ based on the integer partitions. Furthermore, the denominators of Λ_x and Λ_y suitable to these partitions. For example, if $(s_1 = 1, s_2 = 3)$, the partitions of p_j for $j = 1$ and q_k, $k = 1, 2, 3$ are $(p_1, q_1^3), (p_1, q_1q_2), (p_1, q_3)$ and the partitions of Λ_x and Λ_y

are $\left[\left(\frac{\Lambda_x}{1!},\frac{\Lambda_y^3}{3!},\frac{\Lambda_y^2}{2!}\right)\right]$ for p_1, q_1^3, $\left[\left(\frac{\Lambda_x}{1!},\frac{\Lambda_y^2}{2!},\frac{\Lambda_y}{1!}\right)\right]$ for p_1, q_1q_2, $\left[\left(\frac{\Lambda_x}{1!},\frac{\Lambda_y^1}{1!}\right)\right]$ for p_1, q_3. Using these properties, an algorithm is prepared in Maple for the joint probability function of the BCPD.

A general formula is given in (15) for the joint probability function of the BCPD. $P(X_i = j) = p_j$, $j = 0, 1, 2, \dots .m$ and $P(Y_i = k) = q_k$, $k = 0, 1, 2, \dots, r$ are defined in (15) to obtain joint probabilities of bivariate Neyman type A and B, Thomas and geometric-Poisson distribution respectively,

$$
\begin{aligned}
p_j &= e^{-\mu_1}\mu_1^j/j!, && j = 0, 1, 2, \dots\\
q_k &= e^{-\mu_2}\mu_2^k/k!, && k = 0, 1, 2, \dots\\
p_j &= \binom{m_1}{j}p_1^j(1-p_1)^{m_1-j}, && j = 0, 1, 2, \dots, m_1\\
q_k &= \binom{m_2}{k}p_2^k(1-p_2)^{m_2-k}, && k = 0, 1, 2, \dots, m_2\\
p_j &= e^{-\alpha_1}\alpha_1^{(j-1)}/(j-1)!, && j = 1, 2, \dots\\
q_k &= e^{-\alpha_2}\alpha_2^{(k-1)}/(k-1)!, && k = 1, 2, \dots\\
p_j &= \theta_1(1-\theta_1)^j, && j = 0, 1, 2, \dots\\
q_k &= \theta_2(1-\theta_2)^k, && k = 0, 1, 2, \dots
\end{aligned}
$$

3.2. Joint Moment Characteristics

We turn now to the consideration of moments and coefficient of correlation for the BCPD. As far as we know, product moments, cumulants, coefficient of correlation and covariance of the BCPD have never been investigated before (Homer 2006). We start with finding (a, b)-th product moment $\mu'(a, b) = E(S_1^a S_2^b)$. We derive the product moments of S_1 and S_2 by calculating the joint moment generating function

$$
\begin{aligned}
M(z_1, z_2) = \exp(-(\lambda_0+\lambda_1+\lambda_2))\exp\big(&\lambda_1[p_0 + p_1\exp(z_1) + \cdots + p_m\exp(z_1^m)]\\
&+ \lambda_2[q_0 + q_1\exp(z_2) + \cdots + q_r\exp(z_2^r)]\\
&+ \lambda_0[(p_0 + p_1\exp(z_1) + \cdots + p_m\exp(z_1^m))\\
&\quad (q_0 + q_1\exp(z_2) + \cdots + q_r\exp(z_2^r))]\big)
\end{aligned}
$$

Differentiating $M(z_1, z_2)$ at $z_1 = z_2 = 0$, the (a, b)-th product moments are given by

$$\begin{aligned}
\mu'(1,1) &= \mu_X^{[1]}\mu_Y^{[1]}(\Lambda_1 + \Lambda_2 + \Lambda_0)\\
\mu'(2,1) &= \left(\mu_X^{[1]}\right)^2 \mu_Y^{[1]}(\Lambda_1^2\Lambda_2 + \Lambda_1) + \mu_X^{[2]}\mu_Y^{[1]}(\Lambda_1\Lambda_2 + \Lambda_0)\\
\mu'(3,1) &= \left(\mu_X^{[1]}\right)^3 \mu_Y^{[1]}(\Lambda_1^3\Lambda_2 + \Lambda_1^2) + \mu_X^{[1]}\mu_X^{[2]}\mu_Y^{[1]}(\Lambda_1^2\Lambda_2 + \Lambda_1)\\
&\quad + \mu_X^{[3]}\mu_Y^{[1]}(\Lambda_1\Lambda_2 + \Lambda_0)\\
\mu'(2,2) &= \left(\mu_X^{[1]}\right)^2 \left(\mu_Y^{[1]}\right)^2 (\Lambda_1^2\Lambda_2^2 + \Lambda_1\Lambda_2 + \Lambda_0^2)\\
&\quad + \mu_X^{[2]} \left(\mu_Y^{[1]}\right)^2 (\Lambda_1 + \Lambda_2^2 + \Lambda_2) + \left(\mu_X^{[1]}\right)^2 \mu_Y^{[2]}(\Lambda_1^2\Lambda_2 + \Lambda_1)\\
&\quad + \mu_X^{[2]}\mu_Y^{[2]}(\Lambda_1\Lambda_2 + \Lambda_0)\\
\mu'(2,3) &= \left(\mu_X^{[2]}\right)^2 \left(\mu_Y^{[1]}\right)^3 (\Lambda_1^2\Lambda_2^3 + \Lambda_1\Lambda_2^2 + \Lambda_2)\\
&\quad + \mu_X^{[2]} \left(\mu_Y^{[1]}\right)^3 (\Lambda_1\Lambda_2^3 + \Lambda_2^2) + \left(\mu_X^{[1]}\right)^2 \mu_Y^{[1]}\mu_Y^{[2]}(\Lambda_1^2\Lambda_2^2 + \Lambda_1\Lambda_2 + \Lambda_0^2)\\
&\quad + \mu_X^{[2]}\mu_Y^{[1]}\mu_Y^{[2]}(\Lambda_1\Lambda_2^2 + \Lambda_2)\\
&\quad + \left(\mu_X^{[1]}\right)^2 \mu_Y^{[3]}(\Lambda_1^2\Lambda_2 + \Lambda_1)\mu_X^{[2]}\mu_Y^{[3]}(\Lambda_1\Lambda_2 + \Lambda_0)
\end{aligned} \tag{16}$$

3.3. Cumulants

The joint cumulant generating function of S_1 and S_2 is the logarithm of the joint moment generating function $M(z_1, z_2)$ and is given by

$$\begin{aligned}
\kappa_{S_1,S_2}(z_1, z_2) = -(\lambda_0 + \lambda_1 + \lambda_2)\lambda_1[p_0 + p_1\exp(z_1) + \cdots + p_m\exp(z_1^m)]\\
+ \lambda_2[q_0 + q_1\exp(z_2) + \cdots + q_r\exp(z_2^r)] + \lambda_0[(p_0 + p_1\exp(z_1) + \cdots + p_m\exp(z_1^m))\\
(q_0 + q_1\exp(z_2) + \cdots + p_r\exp(z_2^r))]
\end{aligned} \tag{17}$$

From (17) we have

$$\begin{aligned}
\kappa_{1,1} &= \lambda_1\mu_X + \lambda_2\mu_Y + \lambda_0\mu_X\mu_Y\\
\kappa_{1,2} &= \lambda_1\mu_X + \lambda_2\mu_Y^2 + \lambda_0\mu_X\mu_Y^2\\
\kappa_{2,2} &= \lambda_1\mu_X^2 + \lambda_2\mu_Y^2 + \lambda_0\mu_X^2\mu_Y^2\\
\kappa_{2,3} &= \lambda_1\mu_X^2 + \lambda_2\mu_Y^3 + \lambda_0\mu_X^2\mu_Y^3
\end{aligned}$$

where μ_X and μ_Y are the expected values of X_i and Y_i, $i = 1, 2, \ldots$, respectively.

3.4. Independence of S_1 and S_2

The covariance of S_1 and S_2 is obtained using (2) and (16)

$$\begin{aligned} Cov(S_1, S_2) &= E(S_1S_2) - E(S_1)E(S_2) \\ &= E(X)E(Y)[(\lambda_0+\lambda_1)(\lambda_0+\lambda_2)+\lambda_0] \\ &- [(\lambda_0+\lambda_1)E(X)][(\lambda_0+\lambda_2)E(Y)] \\ &= \lambda_0 E(X)E(Y) \end{aligned} \tag{18}$$

Let σ_{s_1} and σ_{s_2} be standard deviations of the random variables S_1 and S_2, then the coefficient of correlation of S_1 and S_2 is obtained from (2) and (18) as follows

$$\begin{aligned} \rho = Corr(S_1, S_2) &= \frac{Cov(S_1, S_2)}{\sigma_{s_1}\sigma_{s_2}} \\ &= \frac{\lambda_0 E(X)E(Y)}{\sqrt{(\lambda_0+\lambda_1)[V(X)+[E(X)]^2](\lambda_0+\lambda_2)[V(Y)+[E(Y)]^2]}} \end{aligned} \tag{19}$$

Note that the correlation of S_1 and S_2 assumes only positive values. This implies that $\rho = 0$ is a necessary condition for S_1 and S_2 to be independent. Also, $\rho = 1$ if and only if S_1 and S_2 are linearly dependent.

3.5. Asymptotics

If $(\lambda_0+\lambda_1) \to \infty$, $(\lambda_0+\lambda_2) \to \infty$, then

$$(Z_1, Z_2) = \left(\frac{S_1 - (\lambda_0+\lambda_1)E(X)}{\sqrt{(\lambda_0+\lambda_1)[V(X)+[E(X)]^2]}}, \frac{S_2 - (\lambda_0+\lambda_2)E(Y)}{\sqrt{(\lambda_0+\lambda_2)[V(Y)+[E(Y)]^2]}} \right) \tag{20}$$

follows a standardized normal bivariate distribution and asymptotically, $\frac{(Z_1^2 - 2\rho Z_1 Z_2 + Z_2^2)}{1-\rho^2}$ is a Chi-squared distribution with two degrees of freedom.

4. Some Numerical Examples

As an illustration of the BCPD and algorithm, a variety of special cases for the BCPD is considered. An algorithm is prepared in Maple for the joint probability function of the BCPD. This algorithm can also be used for the special cases of the BCPD. The probabilities $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \ldots$ are presented in Table 1, which are calculated from (15) for the bivariate Neyman type A distribution. In these calculations, X_i, $i = 1, 2, \ldots$ have a Poisson distribution with parameter $\mu_1 = 0.35$ and Y_i, $i = 1, 2, \ldots$ have a Poisson distribution with parameter $\mu_2 = 0.65$; M_0, M_1, M_2 are independent Poisson distributed random variables with parameters $\lambda_0 = 0.5, \lambda_1 = 0.7, \lambda_2 = 0.1$, respectively.

Table 2 presents $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \ldots$ for the bivariate Neyman type B distribution where X_i, $i = 1, 2, 3, \ldots$ are binomial distributed with

Table 1: The probabilities $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \ldots$, with the parameters $(\mu_1 = 0.35, \mu_2 = 0.65)$ and $(\lambda_0 = 0.5, \lambda_1 = 0.7, \lambda_2 = 0.1)$.

	s_1					
s_2	0	1	2	3	4	5
0	0.2836	0.1163	0.0674	0.0436	0.0212	0.0192
1	0.0985	0.0776	0.0167	0.0091	0.0149	0.0064
2	0.0867	0.0113	0.0095	0.0074	0.0097	0.0052
3	0.0065	0.0095	0.0074	0.0037	0.0087	0.0049
4	0.0042	0.0082	0.0062	0.0019	0.0063	0.0037
5	0.0038	0.0075	0.0057	0.0011	0.0041	0.0024

parameters $(m_1 = 5, p_1 = 0.02)$ and Y_i, $i = 1, 2, \ldots$ are binomial distributed with parameters $(m_2 = 15, p_2 = 0.3)$; M_0, M_1, M_2 are independent Poisson distributed random variables with parameters $\lambda_0 = 0.4, \lambda_1 = 0.6, \lambda_2 = 0.2$, respectively.

Table 2: The probabilities $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \ldots$, with the parameters $(m_1 = 5, p_1 = 0.02)$, $(m_2 = 15, p_2 = 0.3)$ and $(\lambda_0 = 0.4, \lambda_1 = 0.6, \lambda_2 = 0.2)$.

	s_1					
s_2	0	1	2	3	4	5
0	0.2836	0.1163	0.0674	0.0436	0.0212	0.0192
1	0.0985	0.0776	0.0167	0.0091	0.0149	0.0064
2	0.0867	0.0113	0.0095	0.0074	0.0097	0.0052
3	0.0065	0.0095	0.0074	0.0037	0.0087	0.0049
4	0.0042	0.0082	0.0062	0.0019	0.0063	0.0037
5	0.0038	0.0075	0.0057	0.0011	0.0041	0.0024

The probabilities $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \ldots$ are shown in Table 3, for the bivariate Thomas distribution. In these calculations X_i, $i = 1, 2, 3, \ldots$, have a truncated Poisson distribution with parameter $\alpha_1 = 0.75$ and Y_i, $i = 1, 2, 3, \ldots$ have a truncated Poisson distribution with parameter $\alpha_2 = 2$; M_0, M_1, M_2 are independent Poisson distributed random variables with parameters $\lambda_0 = 0.5, \lambda_1 = 0.4, \lambda_2 = 0.2$, respectively.

The probabilities $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \ldots$ are presented in Table 4, for the bivariate geometric-Poisson distribution. In these calculations, X_i, $i = 1, 2, 3, \ldots$ have a geometric distribution with parameter $\theta_1 = 0.25$ and Y_i, $i = 1, 2, 3, \ldots$, have a geometric distribution with parameter $\theta_2 = 0.5$; M_0, M_1, M_2 are independent Poisson distributed random variables with parameters $\lambda_0 = 0.9, \lambda_1 = 0.5, \lambda_2 = 0.2$, respectively.

The results are also illustrated with an analysis of the earthquake data in Turkey. The data is obtained from the database of the Kandilli Observatory, Turkey. Earthquakes are an unavoidable natural disasters for Turkey since a significant portion of Turkey is subject to frequent destructive mainshocks, their foreshock and aftershock sequences. In this study, mainshocks that occured in

Table 3: The probabilities $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \ldots$, with the parameters $(\alpha_1 = 0.75, \alpha_2 = 2)$ and $(\lambda_0 = 0.5, \lambda_1 = 0.4, \lambda_2 = 0.2)$.

	s_1						
s_2	0	1	2	3	4	5	6
0	0.4266	0.0540	0.0533	0.0306	0.0225	0.0094	0.0082
1	0.0707	0.0288	0.0131	0.0090	0.0061	0.0085	0.0069
2	0.0468	0.0114	0.0096	0.0074	0.0056	0.0067	0.0053
3	0.0421	0.0094	0.0089	0.0052	0.0042	0.0052	0.0047
4	0.0019	0.0061	0.0072	0.0043	0.0035	0.0048	0.0034
5	0.0003	0.0043	0.0064	0.0038	0.0027	0.0032	0.0028
6	0.0002	0.0036	0.0056	0.0029	0.0018	0.0025	0.0019

Table 4: The probabilities $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \ldots$, with the parameters $(\theta_1 = 0.25, \theta_2 = 0.5)$ and $(\lambda_0 = 0.9, \lambda_1 = 0.5, \lambda_2 = 0.2)$.

	s_1							
s_2	0	1	2	3	4	5	6	7
0	0.3122	0.0374	0.0430	0.0449	0.0387	0.0212	0.0145	0.0093
1	0.0285	0.0173	0.0323	0.0146	0.0214	0.0109	0.0098	0.0086
2	0.0097	0.0115	0.0237	0.0099	0.0138	0.0093	0.0083	0.0074
3	0.0149	0.0092	0.0116	0.0084	0.0097	0.0082	0.0045	0.0062
4	0.0099	0.0083	0.0092	0.0063	0.0085	0.0073	0.0037	0.0053
5	0.0076	0.0064	0.0092	0.0055	0.0073	0.0064	0.0021	0.0047
6	0.0068	0.0035	0.0086	0.0048	0.0062	0.0056	0.0001	0.0036
7	0.0052	0.0023	0.0062	0.0027	0.0053	0.0043	0.0001	0.0027

Turkey between 1900 and 2010, having surface wave magnitudes $M_s \geq 5.0$, their foreshocks within five days with $M_s \geq 3.0$ and aftershocks within one month with $M_s \geq 4.0$, are considered. In this area, 132 mainshocks with surface magnitude $M_s \geq 5.0$ have occured between 1900 and 2010.

(Kocyigit & Ozacar 2003)

A BCPD is constructed to explain the total number of foreshocks and aftershocks in Turkey. For this purpose, the neotectonic subdivision of Turkey is considered for the first time with the BCPD. To better understand the neotectonic features and active tectonics of Turkey, the simplied tectonic map of Turkey is given in Figure 1.

As seen in Figure 1, Turkey is divided into three main neotectonic domains: area of extensional neotectonic regime, area of strike-slip neotectonic regime with normal component and area of strike-slip neotectonic regime with thrust component. The mainshocks in Turkey are separated according to these neotectonic zones to obtain more reliable results. Let M_0 be the number of mainshocks in the area of extensional neotectonic regimes, M_1 be the number of mainshocks in the area of strike-slip neotectonic regime with normal component and M_2 be the area of strike-slip neotectonic regime with thrust component. Then X_i,

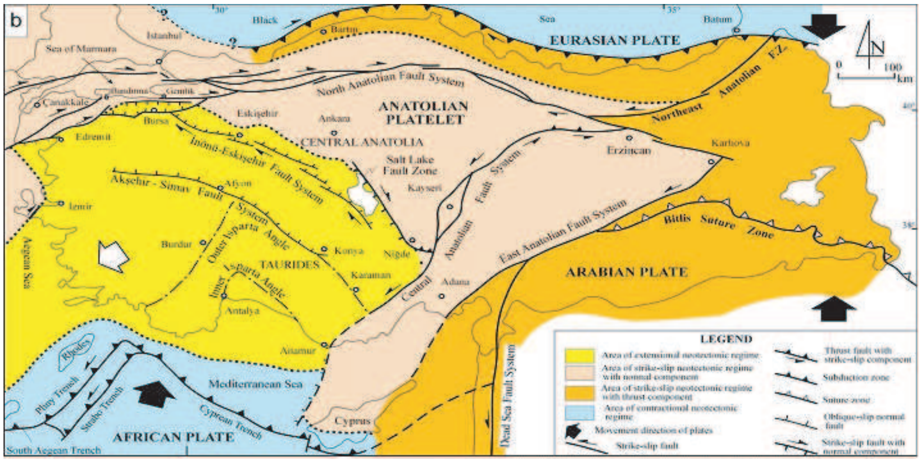

Figure 1: Neotectonic subdivision of Turkey and adjacent areas (Kocyigit & Ozacar 2003).

$i = 1, 2, 3, \dots$ are defined as the number of foreshocks of i^{th} mainshock and Y_i, $i = 1, 2, 3, \dots$ are defined as the number of aftershocks of i^{th} mainshock. Hence, $\left(S_1 = \sum_{i=1}^{N_1} X_i, S_2 = \sum_{i=1}^{N_2} Y_i\right)$ shows the total number of foreshocks and aftershocks for the mainshocks. If the following conditions hold, the pair of (S_1, S_2) has a BCPD:

Condition 1 Fit of the Poisson distribution to the mainshocks: Several studies have modelled earthquakes in Turkey as a Poisson distribution (Kalyoncuoglu 2007, Ozel & Inal 2008). The test for goodness of fit is performed to compare the observed frequency distributions of the mainshocks to the theoretical Poisson distribution. Chi-square values of M_0, M_1, M_2 are calculated as (0.082 with $df = 9$, p-value$= 0.248$), (0.068 with $df = 15$, p-value $= 0.563$), and (0.875 with $df = 10$, p-value $= 0.351$, respectively. These values indicate that M_0, M_1, M_2 fit the Poisson distribution with parameters $\lambda_0 = 2.83, \lambda_1 = 0.862, \lambda_2 = 0.145$ at the level of 0.05, respectively.

Condition 2 Independency tests of the random variables N_1, N_2, X_i and Y_i, $i = 1, 2, \dots$: Previous studies have indicated that there is no correlation between the number of mainshocks, foreshocks and aftershocks (Agnew & Jones 1991). Spearman's ρ test verifies the absence of correlation between N_1 and X_i, $i = 1, 2, \dots$ (Spearman's $\rho = 0.092$; p-value $= 0.759$). No correlation is also found between N_2 and Y_i, $i = 1, 2, \dots$ (Spearman's $\rho = 0.017$; p-value $= 0.473$). Similarly, it is shown that there is no statistically significant dependence between X_i and Y_i, $i = 1, 2, \dots$ (Spearman's $\rho = 0.098$; p-value $= 0.764$).

Condition 3 Fit of the binomial distribution to the foreshocks: As discussed in Jones (1985), if the occurrence of foreshock sequences is assumed as independent from the occurrence of mainshocks without foreshocks, then the distribution of foreshocks in the set of all earthquakes can be treated as a binomial distribution. The percentage, p, of foreshocks is an estimate of the probability that a future earthquake will be a foreshock. After obtaining the frequency distribution of foreshocks and the result of the test for goodness of fit ($\chi^2 = 1.437$, with $df = 36$, p-value $= 0.925$), it is seen that X_i, $i = 1, 2, \ldots$ have a binomial distribution with parameters $m = 35, p = 0.953$ at the level of 0.05.

Condition 4 Fit of the geometric distribution to the aftershocks: It is pointed that in the literature the number of aftershocks of a shock has a geometric distribution (Christophersen & Smith 2000). The test for goodness of fit is carried out to compare the theoretical geometric distribution to the experimental geometric distribution for the number of aftershocks. The test for goodness of fit ($\chi^2 = 1.184$, with $df = 30$, p-value $= 0.273$) shows that Y_i, $i = 1, 2, \ldots$ have a geometric distribution with parameter $\theta = 0.086$.

Because all conditions hold, it can be written $\left(S_1 = \sum_{i=1}^{N_1} X_i, S_2 = \sum_{i=1}^{N_2} Y_i\right)$ and suggested that (S_1, S_2) has a BCPD. Then, $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \ldots$ are computed using (15) for the parameters $\lambda_0 = 2.83, \lambda_1 = 0.862, \lambda_2 = 0.145$; $(m = 35, p = 0.953)$; $\theta = 0.086$ and presented in Table 5.

Table 5: The probabilities $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \ldots$, with the parameters $\theta = 0.086$ and $(m = 35, p = 0.953)$ and $(\lambda_0 = 2.83, \lambda_1 = 0.862, \lambda_2 = 0.145)$.

						s_1					
s_2	0	1	2	3	4	5	6	7	8	9	10
0	0.3630	0.0071	0.0001	0.0049	0.0041	0.0040	0.0037	0.0026	0.0013	0.0010	0.0009
1	0.0075	0.0063	0.0053	0.0045	0.0038	0.0036	0.0035	0.0034	0.0013	0.0009	0.0009
2	0.0001	0.0056	0.0048	0.0041	0.0034	0.0035	0.0034	0.0032	0.0012	0.0008	0.0007
3	0.0058	0.0050	0.0043	0.0037	0.0031	0.0035	0.0032	0.0032	0.0009	0.0008	0.0006
4	0.0051	0.0044	0.0037	0.0033	0.0022	0.0021	0.0021	0.0019	0.0009	0.0007	0.0005
5	0.0041	0.0040	0.0034	0.0031	0.0020	0.0019	0.0019	0.0017	0.0008	0.0006	0.0005
6	0.0035	0.0032	0.0031	0.0030	0.0019	0.0019	0.0016	0.0015	0.0006	0.0005	0.0003
7	0.0021	0.0020	0.0020	0.0029	0.0016	0.0015	0.0014	0.0014	0.0006	0.0004	0.0003
8	0.0018	0.0018	0.0013	0.0015	0.0015	0.0013	0.0009	0.0011	0.0004	0.0003	0.0001
9	0.0013	0.0013	0.0009	0.0013	0.0010	0.0009	0.0007	0.0009	0.0004	0.0003	0.0001
10	0.0010	0.0008	0.0008	0.0009	0.0008	0.0008	0.0007	0.0098	0.0002	0.0001	0.0001

It can be seen from Table 5 that the joint probability recurrence of zero foreshock and zero aftershock is approximately 0.363. The expected values, variances, joint moments, cumulants for S_1 and S_2 are given in Table 6.

Table 6: Expected values, variances and some joint moments and cumulants of S_1 and S_2.

$E(S_1)$	$E(S_2)$	$V(S_1)$	$V(S_2)$	$\mu'(1,1)$	$\mu'(2,1)$	$\kappa_{1,1}$	$\kappa_{1,2}$
123.14	34.59	4113.34	804.49	6384.32	22430.97	1128.05	12811.28

As shown in Table 6 that approximately to 123 foreshocks with $M_s \geq 3.0$ and 35 aftershocks with $M_s \geq 4.0$ are expected in Turkey. It can be concluded from Table 5 that the expected value of total number of foreshocks is less than the expected value of total number of aftershocks. The coefficient of correlation between S_1 and S_2 is found as 0.60 using (19). This result seemed to indicate that increases on the incidence of foreshocks might lead to a more occurences of aftershocks.

5. Conclusion

In this paper the joint probability function, moments, cumulants, covariance and coefficient of correlation of BCPD are obtained. It is concluded that $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \dots$ can be computed easily for the BCPD if p_j, $j = 1, 2, \dots, m$ and q_k, $k = 1, 2, \dots, r$ are known. As seen in Section 3, (9) and (10) need long and tedious computations but $P(S_1 = s_1, S_2 = s_2)$, $s_1, s_2 = 0, 1, 2, \dots$ can be computed accurately from (15) and its proposed algorithm in Maple. Then, some important probabilistic characteristics such as moments, cumulants, covariance, and correlation coefficient of the BCPD are provided. Some numerical examples and an application to the earthquake data have been also presented to illustrate the usage of the bivariate geometric-Poisson, Thomas, Neyman type A and B distributions. The results can be informative regarding BCPD and its applications

Acknowledgements

The author wishes to thank the editor Leonardo Trujillo and anonymous referees for their constructive comments on an earlier version of this manuscript which resulted in this improved version.

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

References

Agnew, D. C. & Jones, L. M. (1991), 'Prediction probabilities from foreshocks', *Journal of Geophysical Research* **96**(11), 959–971.

Ambagaspitiya, R. (1998), 'Compound bivariate Lagrangian Poisson distributions', *Insurance: Mathematics and Economics* **23**(1), 21–31.

Bruno, M. G., Camerini, E., Manna, A. & Tomassetti, A. (2006), 'A new method for evaluating the distribution of aggregate claims', *Applied Mathematics and Computation* **176**, 488–505.

Cacoullos, T. & Papageorgiou, H. (1980), 'On some bivariate probability models applicable to traffic accidents and fatalities', *International Statistical Review* **48**, 345–346.

Christophersen, A. & Smith, E. G. C. (2000), A global model for aftershock behaviour, Proceedings of the 12th World Conference on Earthquake Engineering, Auckland, New Zealand. Paper 0379.

Cossette, H., Gaillardetz, P., Marceau, E. & Rioux, J. (2002), 'On two dependent individual risk models', *Insurance: Mathematics and Economics* **30**, 153–166.

Hesselager, O. (1996), 'Recursions for certain bivariate counting distributions and their compound distributions', *ASTIN Bulletin* **26**, 35–52.

Holgate, P. (1964), 'Estimation for the bivariate Poisson distribution', *Biometrika* **51**, 241–245.

Homer, D. L. (2006), Aggregating bivariate claim severities with numerical fourier inversion, Report, CAS Research Working Party on Correlations and Dependencies among all Risk Sources. 205-230.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1997), *Discrete Multivariate Distributions*, Wiley, New York.

Jones, L. M. (1985), 'Foreshocks and time-dependent earthquake hazard assessment in Southern California', *Bulletin of the Seismological Society of America* **75**, 1669–1679.

Kalyoncuoglu, Y. (2007), 'Evaluation of seismicity and seismic hazard parameters in turkey and surrounding area using a new approach to the Gutenberg-Richter relation', *Journal of Seismology* **11**, 31–148.

Kocherlakota, S. & Kocherlakota, K. (1992), *Bivariate Discrete Distributions*, Marcel Decker, New York.

Kocherlakota, S. & Kocherlakota, K. (1997), Bivariate discrete distributions, *in* S. Kotz, C. B. Read & D. L. Banks, eds, 'Encyclopedia of Statistical Sciences-Update', Vol. 2, Wiley, New York, pp. 68–83.

Kocyigit, A. & Ozacar, A. (2003), 'Extensional neotectonic regime through the ne edge of the outher isparta angle, SW Turkey: New field and seismic data', *Turkish Journal of Earth Sciences* **12**, 67–90.

Ozel, G. & Inal, C. (2008), 'The probability function of the compound Poisson process and an application to aftershock sequences', *Environmetrics* **19**, 79–85.

Ozel, G. & Inal, C. (2010), 'The probability function of a geometric Poisson distribution', *Journal of Statistical Computation and Simulation* **80**, 479–487.

Ozel, G. & Inal, C. (2011), 'On the probability function of the first exit time for generalized Poisson processes', *Pakistan Journal of Statistics* **27**(4). In press.

Panjer, H. (1981), 'Recursive evaluation of a family of compound distributions', *ASTIN Bulletin* **12**, 22–26.

Papageorgiou, H. (1997), Multivariate discrete distributions, *in* C. B. Kotz, S. Read & D. L. Banks, eds, 'Encyclopedia of Statistical Sciences-Update', Vol. 1, Wiley, New York, pp. 408–419.

Rolski, T., Schmidli, H., Schmidt, V. & Teugels, J. (1999), *Stochastic Processes for Insurance and Finance*, John Wiley and Sons.

Sastry, N. (1997), 'A nested frailty model for survival data with an application to the study of child survival in Northeast Brazil', *Journal of the American Statistical Association* **92**(438), 426–435.

Sundt, B. (1992), 'On some extensions of Panjer's class of counting distributions', *ASTIN Bulletin* **22**, 61–80.

Wienke, A., Ripatti, S., Palmgren, J. & Yashin, A. (2010), 'A bivariate survival model with compound Poisson frailty', *Statistics in Medicine* **29**(2), 275–283.

Appendix A. Maple Code for the Joint Probability Function of the BCPD

```
    # $Source: /u/maple/research/lib/bcpd/jpf, v $

bcpd/jpf‘:=proc(L::{set,nvl},q::posint)
local p0, q0, i, lambda1,lambda2, f_final, R, F,
LambdaP_i, LambdaP_n, j, S1, S2, subscript, k, a, b, c,
        n, p, m, z, us, say, y_denom, denom v;

Partitionproduct := proc( n, f, g, statistic) local j, R, visit;
visit:= proc(L) local i, A, S, U, V, W;
   A:= add(pow (x, L[i]), i= 1.. nops (L));
   S:= [seq(coeff (A,x,i), i=1..n)];
   V:= mul (pow (f(i), S[i],i=1,..,n);
   W:= mul (pow (g(i), S[i])*S[i]!, i=1..n);
   U:= abs (n!*V/W);
    i statistic = "sum" then R := R+U
   elif statistic = "part" then R := [op (R), U]
   elif statistic = "len" then R [nops (L)] := R[nops (L)] + U
   elif statistic = "big" then R [L(1)] := R[ K[1]] + U;
   fi;
end;
if n = 0 then if statistic = "sum"
then RETURN (1) else RETURN ([1]) fi fi;
  if statistic = "sum" then R := 0
elif statistic = "part" then R : = []
else R := [ seq (0, j=1..n)] fi;
GeneratePartitions (n, visit);
R end:
 F0 := exp(-lambda1);
   if k =1 then
        F := lambda1;
    else
         i := k-2;
  n := 2;
  F := lambda1^k/k!;
        W := lambda2^k/k!;
    else
  F := F + (lambda1^i/i!)* (LambdaP_n);
        W := W + (lambda2^i/i!)* (LambdaP_n);
  i := i-1;
  n := n+1;
    fi;
    if i < 0
    fi;
  F := 0;
  k := k+1;
   if k > = subscript then k := 4 ;
        else
         i := k-2;
         m := 1;
         n := 2;
      nL := nops (L);
           F := LambdaP_i * LambdaP_n;
           nL := n;
           us := 1;
           if  i =n  then  us := us+1;
           else
           y_denom := seq(us!,1);
           if F >0 then
             do b = nvl(b,0) + F/y_denom while denom =k
           m := m+1;
           fi
        fi;
for z from 1 to n do
  p:=0;
   if subscript >0 then
            p:=p+1;
   fi
  z:=z-1;
   elif z=1 or p>0;
      od;
if p=0 then
subscript := F
                   F := p*LambdaP_n;
                   denom := subscript/(n+1);
                   y_denom:= denom*denom!/(denom-1)!
                     if F and w>0 then
                      do b = NVL(b,0) + F/y_denom while denom =k
                      y_denom :=1;
                      m:=m+1;
                     fi
       fi;
    i := i-1;
    n := n+1;
    p := 0;
    while i < k-trunc(k/2)
           k := k + 1;
                while k > :fNumber;
                            F := 0;
                            i := 1;
                      fi
         j := i-1;
         n := 1;
     for j from 1 to n do
       F := F + lambda1^n/n!;
       W := W + lambda1^n/n!;
      od;
      j := j-1;
      n := n+1;
      while j < 3;
     fi
     F := 0;
     W := 0;
     i := i+1;
     while i > :say
        set value=nvl(nvl (a,0)+ nvl(b,0)+nvl(c,0),0)*F0
     while denom> 0;
   end:
   return F
for i from 1 to n do
            f_final:= 1;
      f_final:= f_final*i;
           i:=i+1;
      while i>n
   fi
   return f_final
fi;
say :=0;
   for i from 1 to n do
      v_value(i):=substr (p_string, i, 1);
         if v_value(i):=p_string then
              say :=say+1;
               fi
               i:= i+1;
               while i-1> length(p);
             od;
   fi
end:

#savelib(’‘bcpd/jpf‘’):
```

Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 567 a 588

Comparison among High Dimensional Covariance Matrix Estimation Methods

Comparación entre métodos de estimación de matrices de covarianza de alta dimensionalidad

Karoll Gómez[1,3,a], Santiago Gallón[2,3,b]

[1] Departamento de Economía, Facultad de Ciencias Humanas y Económicas, Universidad Nacional de Colombia, Medellín, Colombia

[2] Departamento de Estadística y Matemáticas - Departamento de Economía, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia

[3] Grupo de Econometría Aplicada, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia

Abstract

Accurate measures of the volatility matrix and its inverse play a central role in risk and portfolio management problems. Due to the accumulation of errors in the estimation of expected returns and covariance matrix, the solution to these problems is very sensitive, particularly when the number of assets (p) exceeds the sample size (T). Recent research has focused on developing different methods to estimate high dimensional covariance matrixes under small sample size. The aim of this paper is to examine and compare the minimum variance optimal portfolio constructed using five different estimation methods for the covariance matrix: the sample covariance, RiskMetrics, factor model, shrinkage and mixed frequency factor model. Using the Monte Carlo simulation we provide evidence that the mixed frequency factor model and the factor model provide a high accuracy when there are portfolios with p closer or larger than T.

***Key words*:** Covariance matrix, High dimensional data, Penalized least squares, Portfolio optimization, Shrinkage.

Resumen

Medidas precisas para la matriz de volatilidad y su inversa son herramientas fundamentales en problemas de administración del riesgo y portafolio. Debido a la acumulación de errores en la estimación de los retornos esperados y la matriz de covarianza la solución de estos problemas son muy sensibles, en particular cuando el número de activos (p) excede el tamaño muestral (T).

[a] Assistant professor. E-mail: kgomezp@unal.edu.co
[b] Assistant professor. E-mail: santiagog@udea.edu.co

La investigación reciente se ha centrado en desarrollar diferentes métodos para estimar matrices de alta dimensión bajo tamaños muestrales pequeños. El objetivo de este artículo consiste en examinar y comparar el portafolio óptimo de mínima varianza construido usando cinco diferentes métodos de estimación para la matriz de covarianza: la covarianza muestral, el RiskMetrics, el modelo de factores, el shrinkage y el modelo de factores de frecuencia mixta. Usando simulación Monte Carlo hallamos evidencia de que el modelo de factores de frecuencia mixta y el modelo de factores tienen una alta precisión cuando existen portafolios con p cercano o mayor que T.

***Palabras clave*:** matrix de covarianza, datos de alta dimension, mínimos cuadrados penalizados, optimización de portafolio, shrinkage.

1. Introduction

It is well known that the volatility and correlation of financial asset returns are not directly observed and have to be calculated from return data. An accurate measure of the volatility matrix and its inverse is fundamental in empirical finance with important implications for risk and portfolio management. In fact, the optimal portfolio allocation requires solving the Markowitz's mean-variance quadratic optimization problem, which is based on two inputs: the expected (excess) return for each stock and the associated covariance matrix. In the case of portfolio risk assessment, the smallest and highest eigenvalues of the covariance matrix are referred to as the minimum and maximum risk of the portfolio, respectively. Additionally, the volatility itself has also become an underlying asset of the derivatives that are actively traded in the financial market of futures and options.

Consequently, many applied problems in finance require a covariance matrix estimator that is not only invertible, but also well-conditioned. A symmetric matrix is well-conditioned if the ratio of its maximum and minimum eigenvalues is not too large. Then it has full-rank and can be inverted. An ill-conditioned matrix has a very large ratio and is close to being numerically non-invertible. This can be an issue especially in the case of large-dimensional portfolios. The larger number of assets p with respect to the sample size T, the more spread out the eigenvalues obtained from a sample covariance matrix due to the imprecise estimation of this input (Bickel & Levina 2008).

Therefore, the optimal portfolio problem is very sensitive to errors in the estimates of inputs. This is especially true when the number of stocks under consideration is large compared to the return history in the sample. Traditionally the literature, the inversion matrix maximizes the effects of errors in the input assumptions and, as a result, practical implementation is problematic. In fact, those can produce the allocation vector that we get based on the empirical data can be very different from the allocation vector we want based on the theoretical inputs, due to the accumulation of estimation errors (Fan, Zhang & Yu 2009). Also, Chopra & Ziemba (1993) showed that small changes in the inputs can produce large changes in the optimal portfolio allocation. These simple arguments suggest that severe problems might arise in the high-dimensional Markowitz problem.

Covariance estimation for high dimensional vectors is a classical difficult problem, sometimes referred as the “curse of dimensionality”. In recent years, different parametric and nonparametric methods have been proposed to estimate a high dimensional covariance matrix under small sample size. The most usual candidate is the empirical sample covariance matrix. Unfortunately, this matrix contains severe estimation errors. In particular, when solving the high-dimensional Markowitz problem, one can be underestimating the variance of certain portfolios, that is the optimal vectors of weights (Chopra & Ziemba 1993).

Other nonparametric methods such as 250-day moving average, RiskMetrics exponential smoother and exponentially weighted moving average with different weighting schemes have long been used and are widely adopted particularly for market practitioners. More recently, with the availability of high frequency databases, the technique of realized covariance proposed by Barndorff-Nielsen & Shephard (2004) has gained popularity, given that high frequency data provides opportunities for better inference of market behavior.

Parametric methods have been also proposed. Multivariate GARCH models –MGARCH– were introduced by Bollerslev, R. & Wooldridge (1988) with their early work on time-varying covariance in large dimensions, developing the diagonal vech model and later the constant correlation model (Bollerslev 1990). In general, this family model captures the temporal dependence in the second-order moments of asset returns. However, they are heavily parameterized and the problem becomes computationally unfeasible in a high dimension system, usually for $p \geq 100$ (Engle, Shephard & Sheppard 2008).

A useful approach to simplifying the dynamic structure of the multivariate volatility process is to use a factor model. Fan, Fan & Lv (2008) showed that the factor model is one of the most frequently used effective ways to achieve dimension-reduction. Given that financial volatilities move together over time across assets and markets is reasonable to impose a factor structure (Anderson, Issler & Vahid 2006). The three factor model of Fama & French (1992) is the most widely used in financial literature. Another approach that has been used to reduce the noise inherent in covariance matrix estimators is the shrinkage technique by Stein (1956). Ledoit & Wolf (2003) used this approach to decrease the sensitivity of the high-dimensional Markowitz-optimal portfolios to input uncertainty.

In this paper we examine and compare the minimum variance optimal portfolios constructed using five methods of estimating high dimensional covariance matrix: the sample covariance, RiskMetrics, shrinkage estimator, factor model and mixed frequency factor model. These approaches are widely used both by practitioners and academics. We use the global portfolio variance minimization problem with the gross exposure constraint proposed by Fan et al. (2009) for two reasons: *i*) to avoid the effect of estimation error in the mean on portfolio weights and *ii*) the error accumulation effect from estimation of vast covariance matrices.

The goal of this study is to evaluate the performance of the different methods in terms of their precision to estimate a covariance matrix in the high dimensional minimum variance optimal portfolios allocation context.[1] The simulated Fama-French three factor model was used to generate the returns of $p = 200$ and $p = 500$ stocks over a period of 1 and 3 years of daily and intraday data. Using the Monte Carlo simulation we provide evidence than the mixed frequency factor model and the factor model using daily data show a high accuracy when there are portfolios with p closer or larger than T.

The paper is organized as follows. In Section 2, we present a general review of different methods to estimate high dimensional covariance matrices. In Section 3, we describe the global portfolio variance minimization problem with the gross exposure constraint proposed by Fan et al. (2009), and the optimization methodology used to solve it. In Section 4, we compare the minimum variance optimal portfolio obtained using simulated stocks returns and five different estimation methods for the covariance matrix. Also in this section we include an empirical study using the data of 100 industrial portfolios by Kenneth French web site. Finally, in Section 5 we conclude.

2. General Review of High Dimensional Covariance Matrix Estimators

In this Section, we introduce different methods to estimate the high dimensional covariance matrix which is the input for the portfolio variance minimization problem. Let us first introduce some notation used throughout the paper. Consider a p-dimensional vector of returns, $\boldsymbol{r}_t = (r_{1t}, \ldots, r_{pt})'$, on a set of p stocks with the associated $p \times p$ covariance matrix, $\boldsymbol{\Sigma}_t$, $t = 1, \ldots, T$.

2.1. Sample Covariance Matrix

The most usual candidate for estimating $\boldsymbol{\Sigma}$ is the empirical sample covariance matrix. Let $\boldsymbol{R}$ be a $p \times T$ matrix of p returns on T observations. The sample covariance matrix is defined by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T-1}\boldsymbol{R}\left(\boldsymbol{I} - \frac{1}{T}\boldsymbol{\imath}\boldsymbol{\imath}'\right)\boldsymbol{R}' \tag{1}$$

where $\boldsymbol{\imath}$ denotes a $T \times 1$ vector of ones and $\boldsymbol{I}$ is the identity matrix of order T.[2] The (i, j)th element of $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma}^{ij} = (T-1)^{-1}\sum_{t=1}^{T}\left(r_t^i - \bar{r}^i\right)\left(r_t^j - \bar{r}^i\right)$ where r_t^i and r_t^j are the ith and jth returns of the assets i and j on $t = 1, \ldots, T$, respectively; and $\bar{r}^i$ is the mean of the ith return.

[1] Other authors have compared a set of models which are suitable to handle large dimensional covariance matrices. Voev (2008) compares the forecasting performance and also proposes a new methodology which improves the sample covariance matrix. Lam, Fung & Yu (2009) also compare the predictive power of different methods.

[2] When $p \geq T$ the rank of $\hat{\boldsymbol{\Sigma}}$ is $T-1$ which is the rank of the matrix $\boldsymbol{I} - \frac{1}{T}\boldsymbol{\imath}\boldsymbol{\imath}'$, thus it is not invertible. Then, when p exceeds $T-1$ the sample covariance matrix is rank deficient, (Ledoit & Wolf (2003)).

Although the sample covariance matrix is always unbiased estimator is well known that the sample covariance matrix is an extremely noisy estimator of the population covariance matrix when p is large (Dempster 1979).[3] Indeed, estimation of covariance matrix for samples of size T from a p-variate Gaussian distribution, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_p)$, has unexpected features if both p and T are large such as extreme eigenvalues of $\boldsymbol{\Sigma}_p$ and associated eigenvectors (Bickel & Levina 2008).[4]

2.2. Exponentially Weighted Moving Average Methods

Morgan's RiskMetrics covariance matrix, which is very popular among market practitioners, is just a modification of the sample covariance matrix which is based on an exponentially weighted moving average method. This method attaches greater importance on the more recent observations while further observations on the past have smaller exponential weights. Let us denote $\boldsymbol{\Sigma}_{RM}$ the RiskMetrics covariance matrix, the (i, j)th element is given by

$$\boldsymbol{\Sigma}_{RM}^{ij} = (1-\omega)\sum_{t=1}^{T}\omega^{t-1}\left(r_t^i - \bar{r}^i\right)\left(r_t^j - \bar{r}^j\right) \tag{2}$$

where $0 < \omega < 1$ is the decay factor. Morgan (1996) suggest to use a value of 0.94 for this factor. It can be write also as follows:

$$\boldsymbol{\Sigma}_{RM,t} = \omega \boldsymbol{r}_{t-1}\boldsymbol{r}'_{t-1} + (1-\omega)\boldsymbol{\Sigma}_{RM,t-1}$$

which correspond a BEKK scalar integrated model by Engle & Kroner (1995).

Other straightforward methods such as rolling averages and exponentially weighted moving average using different weighting schemes have long been used and are widely adopted specially among practitioners.

2.3. Shrinkage Method

Regularizing large covariance matrices using the Stein (1956) shrinkage method have been used to reduce the noise inherent in covariance estimators. In his seminal paper Stein found that the optimal trade-off between bias and estimation error can be handled simply taking properly a weighted average of the biased and unbiased estimators. This is called shrinking the unbiased estimator full of estimation error towards a fixed target represented by the biased estimator.

This procedure improved covariance estimation in terms of efficiency and accuracy. The shrinkage pulls the most extreme coefficients towards more central values, systematically reducing estimation error where it matters most. In summary, such method produces a result to exhibit the following characteristics: i) the

[3] There is a fair amount of theoretical work on eigenvalues of sample covariance matrices of Gaussian data. See Johnstone (2001) for a review.

[4] For example, the larger p/T the more spread out the eigenvalues of the sample covariance matrix, even asymptotically.

estimate should always be positive definite, that is, all eigenvalues should be distinct from zero and ii) the estimated covariance matrix should be well-conditioned.

Ledoit & Wolf (2003) used this approach to decrease the sensitivity of the high-dimensional Markowitz-optimal portfolios to input uncertainty. Let us denote $\mathbf{\Sigma}_S$ the shrinkage estimators of the covariance matrix, which generally have the form

$$\mathbf{\Sigma}_S = \alpha \boldsymbol{F} + (1-\alpha)\widehat{\mathbf{\Sigma}} \tag{3}$$

where $\alpha \in [0,1]$ is the shrinkage intensity optimally chosen, $\boldsymbol{F}$ corresponds to a positive definite matrix which is the target matrix and $\widehat{\mathbf{\Sigma}}$ represents the sample covariance matrix.

The shrinkage intensity is chosen as the optimal α with respect to a loss function (risk), $L(\alpha)$, defined as a quadratic measure of distance between the true and the estimated covariance matrices based on the Frobenius norm. That is

$$\alpha^* = \arg\min \mathbb{E}\left[\left\|\alpha \boldsymbol{F} + (1-\alpha)\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|^2\right]$$

Given that α^* is non observable, Ledoit & Wolf (2004) proposed a consistent estimator of α for the case when the shrinkage target is a matrix in which all pairwise correlations are equal to the same constant. This constant is the average value of all pairwise correlations from the sample covariance matrix. The covariance matrix resulting from combining this correlation matrix with the sample variances, known as equicorrelated matrix, is the shrinkage target.

Ledoit & Wolf (2003) also proposed to estimate the covariance matrix of stock returns by an optimally weighted average of two existing estimators: the sample covariance matrix with the single-index covariance matrix or the identity matrix.[5]

An alternative method frequently used proposes banding the sample covariance matrix or estimating a banded version of the inverse population covariance matrix. A relevant assumption, in particular for time series data, is that the covariance matrix is banded, meaning that the entries decay based on their distance from the diagonal. Thus, Furrer & Bengtsson (2006) proposed to shrink the covariance entries based on this distance from the diagonal. In other words, this method keeps only the elements in a band along its diagonal and gradually shrinking the off-diagonal elements toward zero.[6] Wu & Pourahmadi (2003) and Huang, Liu, Pourahmadi & Liu (2006) estimate the banded inverse covariance matrix by using thresholding and L_1penalty, respectively.[7]

2.4. Factor Models

The factor model is one of the most frequently used effective ways for dimension reduction, and a is widely accepted statistical tool for modeling multivariate

[5] The single-index covariance matrix corresponds to a estimation using one factor model given the strong consensus about the use of the market index as a natural factor.

[6] This method is also known how "tapering" the sample covariance matrix.

[7] Thresholding a matrix is to retain only the elements whose absolute values exceed a given value and replace others by zero.

volatility in finance. If few factors can completely capture the cross sectional variability of data then the number of parameters in the covariance matrix estimation can be significatively reduced (Fan et al. 2008). Let us consider the $p \times 1$ vector $\boldsymbol{r}_t$. Then the K-factor model is written as

$$\boldsymbol{r}_t = \boldsymbol{\Lambda f}_t + \boldsymbol{\nu}_t = \sum_{k=1}^{K} \boldsymbol{\lambda}_k \cdot f_{kt} + \boldsymbol{\nu}_t \tag{4}$$

where $\boldsymbol{f}_t = (f_{1t}, \dots, f_{Kt})'$ is the K-dimensional factor vector, $\boldsymbol{\Lambda}$ is a $p \times K$ unknown constant loading matrix which indicates the impact of the kth factor over the ith variable, and $\boldsymbol{\nu}_t$ is a vector of idiosyncratic errors. $\boldsymbol{f}_t$ and $\boldsymbol{\nu}_t$ are assumed to satisfy

$$\begin{aligned}
&\mathbb{E}(\boldsymbol{f}_t \mid \Im_{t-1}) = \mathbf{0}, \quad \mathbb{E}(\boldsymbol{f}_t \boldsymbol{f}_t' \mid \Im_{t-1}) = \boldsymbol{\Phi}_t = \operatorname{diag}\{\phi_{1t}, \dots, \phi_{Kt}\}, \\
&\mathbb{E}(\boldsymbol{\nu}_t \mid \Im_{t-1}) = \mathbf{0}, \quad \mathbb{E}(\boldsymbol{\nu}_t \boldsymbol{\nu}_t' \mid \Im_{t-1}) = \boldsymbol{\Psi} = \operatorname{diag}\{\psi_1, \dots, \psi_p\}, \\
&\mathbb{E}(\boldsymbol{f}_t \boldsymbol{\nu}_t' \mid \Im_{t-1}) = \mathbf{0}.
\end{aligned}$$

where $\Im_{t-1}$ denotes the information set available at time $t-1$.

The covariance matrix of $\boldsymbol{r}_t$ is given by

$$\boldsymbol{\Sigma}_{F,t} = \mathbb{E}(\boldsymbol{r}_t \boldsymbol{r}_t' \mid \Im_{t-1}) = \boldsymbol{\Lambda \Phi}_t \boldsymbol{\Lambda}' + \boldsymbol{\Psi} = \sum_{k=1}^{K} \boldsymbol{\lambda}_k \boldsymbol{\lambda}_k' \phi_{kt} + \boldsymbol{\Psi} \tag{5}$$

where all the variance and covariance functions depend on the common movements of f_{kt}.

The multi-factor model which utilizes observed market returns as factors has been widely used both theoretically and empirically in economics and finance. It states that the excessive return of any asset r_{it} over the risk-free interest rate satisfies the equation above. Fama & French (1992) identified three key factors that capture the cross-sectional risk in the US equity market, which have been widely used. For instance, the Capital Asset Pricing Model −CAPM− uses a single factor to compare the excess returns of a portfolio with the excess returns of the market as a whole. But it oversimplifies the complex market. Fama and French added two more factors to CAPM to have a better description of market behavior. They proposed the “small market capitalization minus big” and “high book-to-price ratio minus low” as possible factors. These measure the historic excess returns of small caps over big caps and of value stocks over growth stocks, respectively. Another choice is macroeconomic factors such as: inflation, output and interest rates; and the third possibility are statistical factors which work under a purely dimension-reduction point of view.

The main advantage of statistical factors is that it is very easy to build the model. Fan et al. (2008) find that the major advantage of factor models is in the estimation of the inverse of the covariance matrix and demonstrate that the factor model provides a better conditioned alternative to the fully estimated covariance matrix. The main disadvantage is that there is no clear meaning for the factors. However, a lack of interpretability is not much of a handicap for portfolio optimization. Peña & Box (1987), Chan, Karceski & Lakonishok (1999), Peña & Poncela (2006), Pan & Yao (2008) and Lam & Yao (2010) among others have studied the covariance matrix estimate based on the factor model context.

2.5. Realized Covariance

More recently, with the availability of high frequency databases, the technique of realized volatility introduced by Andersen, Bollerslev, Diebold & Labys (2003) in a univariate setting has gain popularity. In a multivariate setting, Barndorff-Nielsen & Shephard (2004) proposed the realized covariance $-RCV-$, which is computed by adding the cross products of the intra-day returns of two assets. Dividing day t into M non-overlapping intervals of length $\Delta = 1/M$, the realized covariance between assets i and j can be obtained by

$$\mathbf{\Sigma}_{RCV,t}^{\Delta} = \sum_{m=1}^{M} r_{t,m}^{i} r_{t,m}^{j} \tag{6}$$

where $r_{t,m}^{i}$ is the continuously compounded return on asset i during the mth interval on day t.

The RCV based on the synchronized discrete observations of the latent process is a good proxy or representative of the integrated covariance matrix. Barndorff-Nielsen & Shephard (2004) showed that this is true in the low dimensional case. However, in the high dimensional case, i.e. when the dimension p is not small compared with T, it is in general not a good proxy (Zheng & Li 2010). This is a consequence of several issues related with non-synchronous trading, market microstructure noise and spurious intra-day dependence.

Indeed, estimating high dimensional integrated covariance matrix has been drawing more attention. Several solutions have been proposed that are robust to these frictions. Bannouh, Martens, Oomen & van Dijk (2010) propose a Mixed-Frequency Factor Model −MFFM− for estimating the daily covariance matrix for a vast number of assets, which aims to exploit the benefits of high-frequency data and a factor structure. They proposed to obtain the factor loadings in the conventional way by linear regression using daily stock information, and calculated the factor covariance matrix and residual variances with high precision from intra-day data. Using this approach they can avoid non-synchronicity problems inherent in the use of high frequency data for individual stocks.

Considering the same linear factor structure specified in (4), the covariance matrix can be defined as before:

$$\mathbf{\Sigma}_{MFFM} = \mathbf{\Lambda\Pi\Lambda}' + \mathbf{\Theta} \tag{7}$$

where $\mathbf{\Pi} = \mathbb{E}(\mathcal{F}\mathcal{F}')$ is the realized covariance matrix obtained using $\mathcal{F}$ high-frequency factor return observations. $\mathbf{\Lambda}$ denotes the factor loadings, and $\mathbf{\Theta}$ the idiosyncratic residuals, which are obtained using $\boldsymbol{\nu} = \mathcal{R} - \mathbf{\Lambda}\mathcal{F}$ where $\mathcal{R}$ denotes the high-frequency matrix return observations.

This methodology has several advantages over the realized covariance matrix. First, the advantages of dimension reduction in the context of the factor model based purely on daily data continue to hold in the MFFM. Second, the MFFM makes efficient use of high-frequency factor data while bypassing potentially severe biases induced by microstructure noise for the individual assets. Third, we can easily expand the number of assets in the MFFM approach while this is more difficult with the RC matrix for which the inverse does not exist when the number of assets exceeds the number of return observations per asset. For additional details see Bannouh et al. (2010).

Wang & Zou (2009) also develop a methodology for estimating large volatility matrices based on high frequency data. The estimator proposed is constructed in two stages: first, they propose to calculate the average of the realized volatility matrices constructed using tick method and pre sampling frequency, which is called ARVM estimator. Then, regularize ARVM estimator to yield good consistent estimators of the large integrated volatility matrix. Other proposal have been introduced by Barndorff-Nielsen, Hansen, Lunde & Shephard (2010), Zheng & Li (2010), among others.

3. Portfolio Variance Minimization Problem with the Gross Exposure Constraint

In this section, we start recalling the portfolio variance minimization problem proposed by Fan et al. (2009). The noteworthy innovation in their proposal is to relax the gross exposure constraint in order to enlarge the pools of admissible portfolios generating more diversified portfolios.[8] Moreover, they showed that there is no accumulation of estimation errors thanks to the gross exposure constraint. We also present, in a different subsection, the LARS algorithm developed by Efron, Hastie, Johnstone & Tibshirani (2004), which permits to find efficiently the solution paths to the constrained variance minimization problem.

3.1. The Variance Minimization Problem with Gross Exposure Constraint

Following the proposal of Fan et al. (2009), we suppose a portfolio with p assets and corresponding returns $\boldsymbol{r} = (r_1, \dots, r_p)'$ to be managed. Let $\boldsymbol{\Sigma}$ be its associated covariance matrix, and $\boldsymbol{w}$ be its portfolio allocation vector. as a consequence, the variance of the portfolio return $\boldsymbol{w'r}$ is given by $\boldsymbol{w'\Sigma w}$. Considering the variance minimization problem with gross-exposure constraint as follows:

$$\begin{aligned} &\min_{\boldsymbol{w}} \Gamma(\boldsymbol{w}, \boldsymbol{\Sigma}) = \boldsymbol{w'\Sigma w}, \\ &\text{subject to: } \boldsymbol{w'\imath} = 1 \qquad \text{(Budget constraint)} \\ &\qquad \|\boldsymbol{w}\|_1 \leq c \qquad \text{(Gross exposure constraint)} \end{aligned} \tag{8}$$

where $\|\boldsymbol{w}\|_1$ is the L_1 norm. The constraint $\|\boldsymbol{w}\|_1 \leq c$ prevents extreme positions in the portfolio. Notice that when $\|\boldsymbol{w}\|_1 = 1$, ie $c = 1$, no short sales are allowed as studied by Jagannathan & Ma (2003); when $c = \infty$, there is no constraint on

[8] The portfolio optimization with the gross-exposure constraint bridges the gap between the optimal no-short-sale portfolio studied by Jagannathan & Ma (2003) and no constraint on short-sale in the Markowitz's framework.

short sales as in Markowitz (1952). Thus, the proposal of Fan et al. (2009) is a generalization to the work of them.[9]

The solution to the optimization problem $\boldsymbol{w}^*$ depends sensitively on the input vectors $\boldsymbol{\Sigma}$ and its accumulated estimation errors, but under the gross-exposure constraint, with a moderate value of c, the sensitive of the problem is bounded and these two problems disappear. The upper bounds on the approximation errors is given by

$$\left|\Gamma\left(\boldsymbol{w}, \boldsymbol{\Sigma}\right) - \Gamma\left(\boldsymbol{w}, \widehat{\boldsymbol{\Sigma}}\right)\right| \leq 2a_n c^2 \tag{9}$$

where $\Gamma\left(\boldsymbol{w}, \boldsymbol{\Sigma}\right)$ and $\Gamma(\boldsymbol{w}, \widehat{\boldsymbol{\Sigma}})$ correspond to the theoretical and empirical portfolio risks, $a_n = ||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_\infty$ and $\widehat{\boldsymbol{\Sigma}}$ is an estimated covariance matrix based on the data with sample size T.

They point out that this holds for any estimation of covariance matrix. However as long as each element is estimated precisely, the theoretical minimum risk and the empirical risk calculated from the data should be very close, thanks to the constraint on the gross exposure.

3.2. The Optimization Methodology

The risk minimization problem described in the equation (8) takes the form of the Lasso problem developed by Tibshirani (1996). For a complete study of Lasso (Least Absolute Shrinkage and Selection Operator) method see Buhlmann & van de Geer (2011). The connection between Markowitz problem and Lasso is conceptually and computationally useful. The Lasso is a constrained version of ordinary least squares −OLS−, which minimize a penalized residual sum of squares. Markowitz problem also can be viewed as a penalized least square problem given by

$$\begin{gathered}\boldsymbol{w}^*_{\text{Lasso}} = \arg\min \sum_{t=1}^{T}\left(y_t - b - \sum_{j=1}^{p-1} x_{tj} w_j\right)^2 \\ \text{subject to } \sum_{j=1}^{p-1} |w_j| \leq d \qquad (L_1 \text{ penalty})\end{gathered} \tag{10}$$

where $y_t = r_{tp}$, $x_{tj} = r_{tp} - r_{tj}$ with $j = 1, \dots, p-1$ and $d = c - \left|1 - \sum_{j=1}^{p-1} w_j^*\right|$. Thus, finding the optimal weight $\boldsymbol{w}$ is equivalent to finding the regression coeffcient $\boldsymbol{w}^* = (w_1, \dots, w_{p-1})'$ along with the intercept b to best predict y.

Quadratic programming techniques can be used to solve (8) and (10). However, Efron et al. (2004) proposed to compute the Lasso solution using the LARS algorithm which uses a simple mathematical formula that greatly reduced the

[9] Let w^+ and w^- be the total percent of long and short positions, respectively. Then, under $w^+ - w^- = 1$ and $w^+ - w^- \leq c$, we have $w^+ = (c+1)/2$ and $w^- = (c-1)/2$. These correspond to percentage of long and short positions allowed. The constraint on $\|\boldsymbol{w}\|_1 \leq c$ is equivalent to the constraint on w^-, which is binding when the portfolio is optimized.

computational burden. Fan et al. (2009) showed that this algorithm provides an accurate solution approximation of problem (8).

The LARS procedure works roughly as follows. Given a collection of possible predictors, we select the one having largest absolute correlation with the response y, say x_{j_1} ,and perform simple linear regression of y on x_{j1}. This leaves a residual vector orthogonal to x_{j_1}, which now is considered to be the response. We project the other predictors orthogonally to x_{j1} and repeat the selection process. Doing the same procedure after s steps this produce a set of predictors $x_{j_1}, x_{j_2}, \ldots, x_{j_s}$ that are then used in the usual way to construct a s-parameter linear model (Efron et al. (2004)). For more details, the LARS algorithm steps are summarized in the Appendix A

The LARS algorithm applied to the problem (10) produces the whole solution path $\boldsymbol{w}^*(d)$, for all $d \geq 0$. The number of non-vanishing weights varies as d ranges from 0 to 1. It recruits successively one stock, two stocks, and gradually all the stocks of the portfolio. When all stocks are recruited, the problem is the same as the Markowitz risk minimization problem, since no gross-exposure constraint is imposed when d is large enough (Fan et al. 2009).

4. Comparison of Minimum Variance Optimal Portfolios

In this section, we compare the minimum variance optimal portfolio constructed using five different estimation methods for the covariance matrix: the sample covariance, RiskMetrics, factor model, mixed frequency factor model and shrinkage method.

4.1. Dataset

We use a simulated return of p stocks considering 1 and 3 years of daily data, this is $T = 252, 756$. The simulated Fama-French three factor model is used to generate the returns of $p = 200$ and $p = 500$ stocks, using the specification in (4) and following the procedure employed by Fan et al. (2008). We carry out the following steps:

1. Generate p factor loading vectors $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p$ as a random sample of size p from the trivariate normal distribution $N(\boldsymbol{\mu}_{\boldsymbol{\lambda}}, \mathrm{cov}_{\boldsymbol{\lambda}})$. This is kept fixed during the simulation.

2. Generate a random sample of factors $\boldsymbol{f}_1$, $\boldsymbol{f}_2$ and $\boldsymbol{f}_3$ of size T from the trivariate normal distribution $N(\boldsymbol{\mu}_{\boldsymbol{f}}, \mathrm{cov}_{\boldsymbol{f}})$.

3. Generate p standard deviations of the errors $\psi_1, \ldots, \psi_p$ as a random sample of size p from a gamma distribution with parameters $\alpha = 3.3586$ and $\beta = 0.1876$. This is also kept fixed during the simulation.

4. Generate a random sample of p idiosyncratic noises $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_p$ with size T from the p-variate normal distribution $N(0, \boldsymbol{\Psi})$, and also from Student's t distribution t-Stud$(6, \boldsymbol{\Psi})$.
5. Calculate a random sample of returns $\boldsymbol{r}_t$, $t = 1, \ldots, T$ using the model (4) and the information generated in steps 1, 2 and 4.
6. By means of this simulated returns we calculated the following covariance matrix using: the sample covariance, RiskMetrics, factor model and shrinkage method, as was discussed in Section 2.

The parameters used in steps 1, 2 and 3, were taken from Fan et al. (2008) who fit three-factor model using the three-year daily data of 30 Industry Portfolios from May 1, 2002 to August 29, 2005, available at the Kenneth French website. They calculated the sample means and sample covariance matrices of $\boldsymbol{f}$ and $\boldsymbol{\lambda}$ denoted by $(\boldsymbol{\mu}_{\boldsymbol{f}}, \mathrm{cov}_{\boldsymbol{f}})$ and $(\boldsymbol{\mu}_{\boldsymbol{\lambda}}, \mathrm{cov}_{\boldsymbol{\lambda}})$. These values are reported in Appendix B, Table 4.

Additionally, to implement the Mixed-Frequency Factor Model we simulated, as proposed by Bannouh et al. (2010), five minutes high frequency factor data $\mathcal{F}$ from a trivariate Gaussian distribution, $N(0, \mathrm{cov}_f)$ and high frequency idiosyncratic noises from a p-variate normal distribution $N(0, \boldsymbol{\Psi})$. In practice high-frequency financial asset prices bring problems such as non-synchronous trading and are contaminated by market microstructure noise.

We implement non-synchronous trading by assuming trades arrive following a Poisson process with an intensity parameter equal to the average number of daily trades for the S&P500.[10] Also, we include a microstructure noise component in the model, $\boldsymbol{\eta} \sim N(0, \boldsymbol{\Delta})$ where $\boldsymbol{\Delta} = (1/4\tau)(\boldsymbol{\Lambda\Pi\Lambda}' + \boldsymbol{\Theta})$ with τ the high frequency sample size returns. Using this we also calculate the random sample of high frequency returns $\mathcal{R} = \boldsymbol{\Lambda}\mathcal{F} + \boldsymbol{\nu} + \boldsymbol{\eta}$ and by means of these returns we calculate (7).

Finally, from the estimated covariance matrices obtained using the different methods, we find an approximately optimal solution to problem (8) using the LARS algorithm. For this calculation, we take the no short sale constraint optimal portfolio as dependent variable in (10). Thus, having the optimal portfolio weights and the estimated covariance matrix we calculate the theoretical and empirical minimum variance optimal risk. In this paper, the risk of each optimal portfolio is referring to the standard deviation of the quantities $\Gamma(\boldsymbol{w}, \boldsymbol{\Sigma})$ and $\Gamma\left(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\Sigma}}\right)$, calculated as the square-root thereof.

4.2. Simulation results

Fan et al. (2009) showed that the unknown theoretical minimum risk, $\Gamma(\boldsymbol{w}, \boldsymbol{\Sigma})$, and the empirical minimum risk, $\Gamma\left(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\Sigma}}\right)$, of the invested portfolio are approximately the same as long as: i) the c is not too large and ii) the accuracy of

[10] This value corresponds to 19.385 which is the average number of daily trades over the period November 2006 through May 2008 (Bannouh et al. (2010)Z).

estimated covariance matrix is not too low. Based on this result, we are going to compare the theoretical solution path of the minimum variance optimal portfolios with the solution path obtained using five different estimation methods for high dimensional covariance matrix: the sample covariance, RiskMetrics, factor model, shrinkage and mixed frequency factor model.

We first examine the results in case of $p = 200$ with 100 replications. In Table 1, we present the mean value of the minimum variance optimal portfolio in three cases: i) when no short sales are allowed, that is $c = 1$, as studied by Jagannathan & Ma (2003), ii) under a gross exposure constraint equal to $c = 1.6$ as proposed by Fan et al. (2009), which correspond to a typical choice and iii) when $c = \infty$, that is, no constraint on short sales as in Markowitz (1952).[11]

The results show that the empirical minimum portfolio risk obtained using the covariance matrix estimated from mixed frequency factor model method has the smaller difference with respect to the theoretical risk. Thus, the MFFM method produces the better relative estimation accuracy among the competing estimators. The gains come from the fact that this model exploits the advantages of both high frequency data and the factor model approach. The factor model also permits a precise estimation of the covariance matrix, which is closer to the MFFM. The accuracy of the covariance matrix estimated from the shrinkage method is also fairly similar to the factor models and slightly superior to the sample covariance matrix.[12] Finally, all estimation methods overcome the RiskMetrics, especially when no short sales are allowed. We have the same results when we used three years of daily returns, presented at the bottom of Table 1.

Table 1: Theoretical and empirical risk of the minimum variance optimal portfolio High dimensional case ($p = 200$).

True covariance matrix		Competing estimators				
c	$\mathbf{\Sigma}$	$\widehat{\mathbf{\Sigma}}$	$\mathbf{\Sigma}_{RM}$	$\mathbf{\Sigma}_S$	$\mathbf{\Sigma}_F$	$\mathbf{\Sigma}_{MFFM}$
$T = 252$						
1	21.23	19.16	16.65	19.72	19.88	20.12
1.6	7.64	6.76	5.92	7.29	7.35	7.42
∞	1.32	0.88	0.85	0.93	1.03	1.01
$T = 756$						
1	19.84	18.53	15.53	19.01	19.15	19.45
1.6	5.85	3.82	3.05	4.95	5.05	5.53
∞	1.25	0.69	0.61	0.87	0.94	0.98

As we can see in Table 1, in all cases the theoretical risk is greater than the empirical risk, although in some cases the difference is slim. The intuition of

[11] The corresponding values for parameter d in each case is: 0, 0.7, and 12.8.

[12] We used as target matrix the identity which works well as was shown by Ledoit & Wolf (2003) and also the shrinkage target actually proposed by them. The practical problem in applying the shrinkage method is to determine the shrinkage intensity. Ledoit & Wolf (2003) showed that it behaves like a constant over the sample size and provide a way to consistently estimate it. Following the Ledoit & Wolf (2003) proposal we found $\alpha^* = 0.7895$. However, we check the stability of the results using different values for α chosen ad hoc. The results show that the as long as the shrinkage intensity is lower than α^* the methods tends to underestimate a little bit more the risk. However, this method maintains his superiority with respect to sample and RiskMetrics estimated covariance matrices. Detailed results are available upon request.

these results is that having to estimate the high dimensional covariance matrix at stake here leads to risk underestimation. In other words, in general the covariance matrix estimation leads to overoptimistic conclusions about the risk. The most dramatic case occurs with the RiskMetrics portfolio, which shows the lower risk.

Additionally, the results show that constrained short sale portfolios are not diversified enough, as also was found by Fan et al. (2009). For instance, the risks can be improved by relaxing the gross exposure constraint, which implies allowing some short positions. However, allowing the possibility of extreme short or long positions in the portfolio we can get a lower optimal risk; extremely negative weights are difficult to implement in practice. Actually, practical portfolio choices always involve constraints on individual assets such as the allocations are no larger than certain percentages of the median daily trading volume of an asset. This result is true no matter what method is used to estimate the covariance matrix and which sample size is used.

Figure 1, shows the whole path solution of the risk for a selected portfolio as a function of LARS steps. The path solution was calculated for each of the five competing methods and the true covariance matrix, using the LARS algorithm. This figure illustrates the decrease in optimal risk when we move from a portfolio with no short sale to allowed short sale portfolio, which is more diversified and therefore less risky. In other words, the graph suggests that the optimal risk decreases as soon as in each step the parameter d is growing. This occurs as long as the LARS algorithm progresses.[13] This implies that the higher value of optimal risk is reached in the case of no short sale.

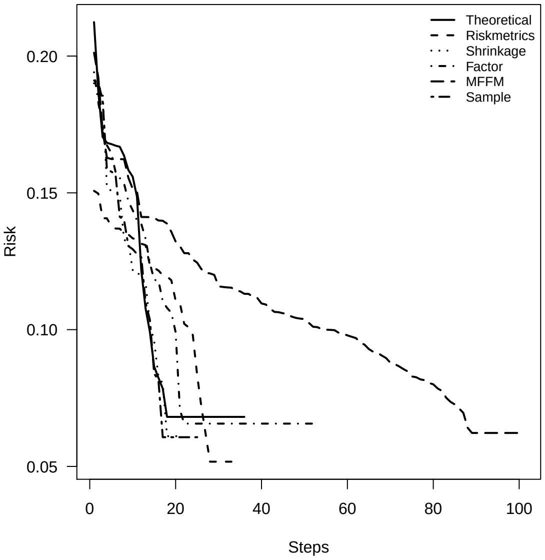

Figure 1: LARS solution path of the optimal risk for each minimum variance portfolio

[13] The number of steps required to complete the algorithm and have the entire solution path can be different in each case

In consequence, once the gross exposure constraint is relaxed the number of selected stocks increases and the portfolio becomes more diversified. In fact, at the first step when d is relaxed the LARS algorithm identifies the stock that permit reduction of the minimum optimal risk under no short sale restriction, permitting this stock to enter into the optimal portfolio allocation with a weight that can be positive or negative. This process is continued until the entire set of stocks are examined and as result in each step you will have a decreasing optimal risk but increasing short percentage. This process is illustrated in Figure 2. Each graph in the panel corresponds to a profile of optimal portfolio weights obtained solving the problem (10) using the true covariance matrix and each estimated covariance matrix.

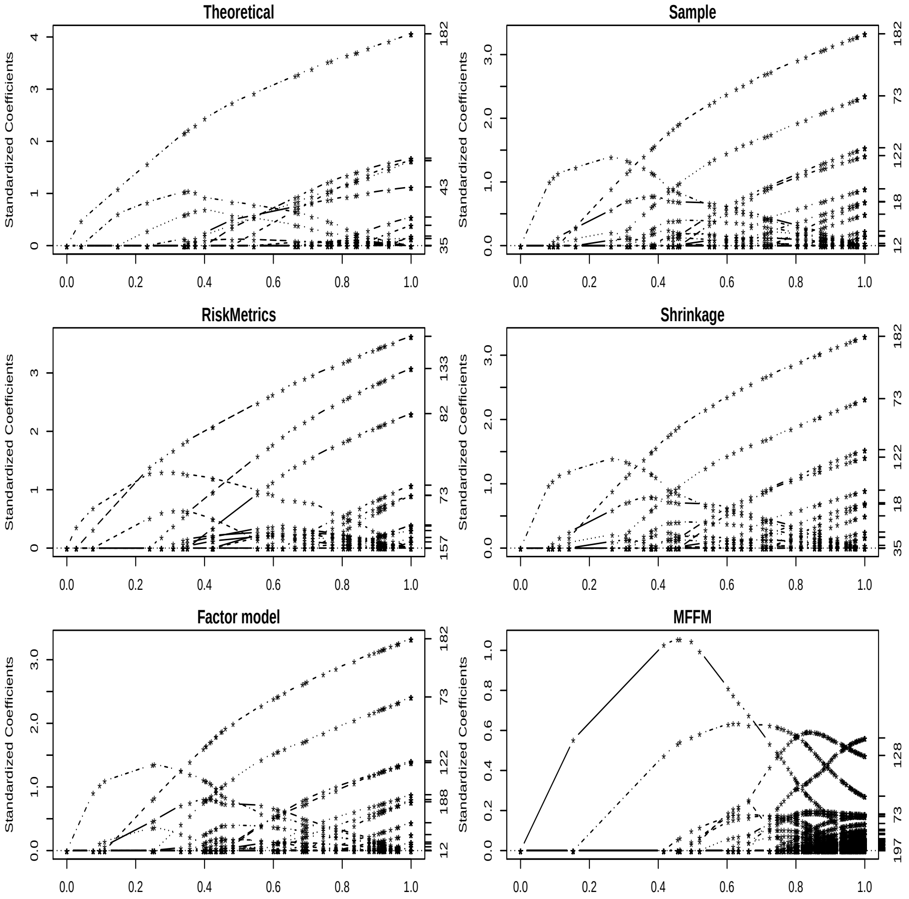

FIGURE 2: Estimated optimal portfolio weights via the Lasso. The abscissae correspond to the standardized Lasso parameter, $s = d/\sum_{j=1}^{p-1} |w_j|$.

The figure shows the optimal portfolio weights as a function of the standardized Lasso parameter $s = d/\sum_{j=1}^{p-1} |w_j|$. Each curve represents the optimal weight of a particular stock in the portfolio as s is varied. We start with no short sale portfolio at $s = 0$. The stocks begin to enter in the active set sequentially as d increases, allowing us to have a more diversified portfolio. Finally, at $s = 1$, the graph shows the stocks that are included in the active stock set where short sales are allowed with no restriction. The number of some of them are labeled on the right side in each graph.[14]

We now examine the results in case of $p = 500$, again with 100 replications. The results, considering this very high dimensional case, are presented in Table 2. Similarly, this table contains the mean value of the minimum variance optimal portfolio risk using different estimation methods for covariance matrix. First of all, as we can see, sampling variability for the case with 500 stocks is smaller than the case involuing 200 stocks. These are due to the fact that with more stocks, the selected portfolio is generally more diversified and hence the risks are generally smaller. This result is according with the founded results by Fan et al. (2009).

Table 2: Theoretical and empirical risk of the minimum variance optimal portfolio Very high dimensional case ($p = 500$).

True covariance matrix		Competing estimators				
c	$\mathbf{\Sigma}$	$\widehat{\mathbf{\Sigma}}$	$\mathbf{\Sigma}_{RM}$	$\mathbf{\Sigma}_{S}$	$\mathbf{\Sigma}_{F}$	$\mathbf{\Sigma}_{MFFM}$
$T = 252$						
1	15.49	13.89	12.85	14.28	14.07	14.16
1.6	4.91	1.89	1.22	4.17	4.04	4.14
∞	1.21	0.40	0.38	1.11	0.98	1.09
$T = 756$						
1	14.04	13.03	12.23	13.71	13.11	13.55
1.6	3.58	1.32	1.00	3.05	1.55	3.68
∞	1.01	0.17	0.01	0.89	0.64	0.78

Additionally, simulation results show that the shrinkage method offers an estimated covariance matrix with superior estimation accuracy. This is reflected in the fact that the minimum optimal portfolio risk using this method is just a little different with respect to the theoretical risk. The mixed frequency factor model and the factor model using daily data also have a high accuracy. However, as can be seen, the factor model, the MFFM and shrinkage method offer a quite close estimation accuracy of the covariance matrix. Finally, all estimation methods overcome the sample covariance matrix, however, its performance is quite similar to the RiskMetrics.

4.3. Empirical Results

In the same way than Fan et al. (2009), data from Kenneth French was obtained is website from January 2, 1997 to December 31, 2010. We use the daily returns of 100 industrial portfolios formed on size and book to market ratio, to estimate according to four estimators, the sample covariance, RiskMetrics, factor model and

[14] The active stock set refers to the stocks with weight different from zero. This set changes as the LARS algorithm progresses. Actually, it can increase or decrease in each step depending if a particular stock is added or dropped from the active set. This is the reason why in Figure 2, some curves at the last step are at zero.

the Shrinkage, the covariance matrix of the 100 assets using the past 12 months' daily returns data.[15] These covariance matrices, calculated at the end of each month from 1997 to 2010, are then used to construct optimal portfolios under three different gross exposure constraints. The portfolios are then held for one month and rebalanced at the beginning of the next month. Different characteristics of these portfolios are presented in Table 3.

Table 3: Returns and Risks based on Fama French Industrial Portfolios, $p = 100$.

c	Mean	Standard deviation	Sharpe ratio	Min weight	Max weight
Sample covariance					
1	20.89	12.03	1.80	0.00	0.30
1.6	22.36	8.06	2.22	−0.05	0.28
∞	15.64	7.13	1.86	−0.11	0.25
Factor model					
1	21.49	12.09	1.82	0.00	0.29
1.6	22.56	8.26	2.24	−0.04	0.24
∞	16.73	7.40	1.90	−0.11	0.22
Shrinkage					
1	21.34	11.90	1.79	0.00	0.29
1.6	22.46	8.06	2.23	−0.05	0.23
∞	15.94	7.16	1.88	−0.11	0.22
RiskMetrics					
1	17.07	9.23	1.43	0.00	0.46
1.6	18.89	7.83	1.56	−0.07	0.44
∞	15.80	6.87	1.48	−0.13	0.42

We found that the optimal no short sale portfolio is not diversified enough. It is still a conservative portfolio that can be improved by allowing some short positions. In fact, when $c = 1$, the risk is greater than when we allowed short positions. These results hold using all covariance matrices measures. Also, we found that the portfolios selected by using the RiskMetrics have lower risk which coincides with Fan et al. (2009) results. Thus, according our simulation and empirical results, RiskMetrics give us the most overoptimistic conclusions about the risk.

Finally, the Sharpe ratio is a more interesting characterization of a security than the mean return alone. It is a measure of risk premium per unit of risk in an investment. Thus the higher the Sharpe Ratio the better. Because of the low returns showed by Riskmetrics, it has also a lower Sharpe ratio. Although differences between the other three methods are not important, the factor model has the higher Sharpe ratio. This result indicates that the return of the portfolio better compensates the investor for the risk taken.

5. Conclusions

When p is small, an estimate of the covariance matrix and its inverse can easily obtained. However, when p is closer or larger than T, the presence of

[15] We do not include the mixed frequency factor model because of the impossibility to have access to high frequency data.

many small or null eigenvalues makes the covariance matrix not positive definite any more and it can not be inverted as it becomes singular. That suggests that serious problems may arise if one naively solves the high-dimensional Markowitz problem. This paper evaluates the performance of the different methods in terms of their precision to estimate a covariance matrix in the high dimensional minimum variance optimal portfolios allocation context. Five methods were employed for the comparison: the sample covariance, RiskMetrics, factor model, shrinkage and realized covariance.

The simulated Fama-French three factor model was used to generate the returns of $p = 200$ and $p = 500$ stocks over a period of 1 and 3 years of daily and intraday data. Thus using the Monte Carlo simulation we provide evidence than the mixed frequency factor model and the factor model using daily data show a high accuracy when we have portfolios with p closer or larger than T. This is reflected in the fact that the minimum optimal portfolio risk using these methods is just a little different with respect to the theoretical risk. The superiority of the MFFM, comes from the fact that this model offers a more efficient estimation of the covariance matrix being able to deal with a very large number of stocks (Bannouh et al. 2010).

Simulation results also show that the accuracy of the covariance matrix estimated from shrinkage method is also fairly similar to the factor models with slightly superior estimation accuracy in a very high dimensional situation. Finally, as have been found in the literature all these estimation methods overcome the sample covariance matrix. However, RiskMetrics shows a low accuracy and in both studies (simulation and empirical) leads to the most overoptimistic conclusions about the risk.

Finally, we discuss the construction of portfolios that take advantage of short selling to expand investment opportunities and enhance performance beyond that available from long-only portfolios. In fact, when long only constraint is present we have an optimal portfolio with some associated risk exposure. When shorting is allowed, by contrast, a less risky optimal portfolio can be achieved.

Acknowledgements

We are grateful to the anonymous referees and the editor of the *Colombian Journal of Statistics* for their valuable comments and constructive suggestions.

[Recibido: septiembre de 2010 — Aceptado: marzo de 2011]

References

Andersen, T., Bollerslev, T., Diebold, F. & Labys, P. (2003), 'Modeling and forecasting realized volatility', *Econometrica* **71**(2), 579–625.

Anderson, H., Issler, J. & Vahid, F. (2006), 'Common features', *Journal of Econometrics* **132**(1), 1–5.

Bannouh, K., Martens, M., Oomen, R. & van Dijk, D. (2010), Realized mixed frequency factor models for vast dimensional covariance estimation, Discussion paper, Econometric Institute, Erasmus Rotterdam University.

Barndorff-Nielsen, O., Hansen, P., Lunde, A. & Shephard, N. (2010), 'Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading', *Journal of Econometrics* **162**(2), 149–169.

Barndorff-Nielsen, O. & Shephard, N. (2004), 'Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics', *Econometrica* **72**(3), 885–925.

Bickel, P. & Levina, E. (2008), 'Regularized estimation of large covariance matrices', *The Annals of Statistics* **36**(1), 199–227.

Bollerslev, T. (1990), 'Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH approach', *Journal of Portfolio Management* **72**(3), 498–505.

Bollerslev, T., R., E. & Wooldridge, J. (1988), 'A capital asset pricing model with time varying covariances', *Journal of Political Economy* **96**(1), 116–131.

Buhlmann, P. & van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics, Springer. Berlin.

Chan, L., Karceski, J. & Lakonishok, J. (1999), 'On portfolio optimization: Forecasting covariances and choosing the risk model', *Review of Financial Studies* **12**(5), 937–974.

Chopra, V. & Ziemba, W. (1993), 'The effect of errors in means, variance and covariances on optimal portfolio choice', *Journal of Portfolio Management* **19**(2), 6–11.

Dempster, A. (1979), 'Covariance selection', *Biometrics* **28**(1), 157–175.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *The Annals of Statistics* **32**(2), 407–499.

Engle, R. & Kroner, K. (1995), 'Multivariate simultaneous generalized ARCH', *Econometric Theory* **11**(1), 122–150.

Engle, R., Shephard, N. & Sheppard, K. (2008), Fitting vast dimensional time-varying covariance models, Discussion Paper Series 403, Department of Economics, University of Oxford.

Fama, E. & French, K. (1992), 'The cross-section of expected stock returns', *Journal of Financial Economics* **47**(2), 427–465.

Fan, J., Fan, Y. & Lv, J. (2008), 'High dimensional covariance matrix estimation using a factor model', *Journal of Econometrics* **147**(1), 186–197.

Fan, J., Zhang, J. & Yu, K. (2009), Asset allocation and risk assessment with gross exposure constraints for vast portfolios, Technical report, Department of Operations Research and Financial Engineering, Princeton University. Manuscrit.

Furrer, R. & Bengtsson, T. (2006), 'Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants', *Journal of Multivariate Analysis* **98**(2), 227–255.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer. New York.

Huang, J., Liu, N., Pourahmadi, M. & Liu, L. (2006), 'Covariance matrix selection and estimation via penalized normal likelihood', *Biometrika* **93**(1), 85–98.

Jagannathan, R. & Ma, T. (2003), 'Risk reduction in large portfolios: Why imposing the wrong constraints helps', *Journal of Finance* **58**(4), 1651–1683.

Johnstone, I. (2001), 'On the distribution of the largest eigenvalue in principal components analysis', *The Annals of Statistics* **29**(2), 295–327.

Lam, C. & Yao, Q. (2010), Estimation for latent factor models for high-dimensional time series, Discussion paper, Department of Statistics, London School of Economics and Political Science.

Lam, L., Fung, L. & Yu, I. (2009), Forecasting a large dimensional covariance matrix of a portfolio of different asset classes, Discussion Paper 1, Research Department, Hong Kong Monetary Autority.

Ledoit, O. & Wolf, M. (2003), 'Improved estimation of the covariance matrix of stock returns with an application to portfolio selection', *Journal of Empirical Finance* **10**(5), 603–621.

Ledoit, O. & Wolf, M. (2004), 'Honey, I shrunk the sample covariance matrix', *Journal of Portfolio Management* **30**(4), 110–119.

Markowitz, H. (1952), 'Portfolio selecction', *Journal of Finance* **7**(1), 77–91.

Morgan, J. P. (1996), Riskmetrics, Technical report, J. P. Morgan/Reuters. New York.

Pan, J. & Yao, Q. (2008), 'Modelling multiple time series via common factors', *Boimetrika* **95**(2), 365–379.

Peña, D. & Box, G. (1987), 'Identifying a simplifying structure in time series', *Journal of American Statistical Association* **82**(399), 836–843.

Peña, D. & Poncela, P. (2006), 'Nonstationary dynamic factor analysis', *Journal of Statistics Planing and Inference* **136**, 1237–1257.

Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *in* J. Neyman, ed., ‘Proceedings of the Third Berkeley Symposium on Mathematical and Statistical Probability’, Vol. 1, University of California, pp. 197–206. Berkeley.

Tibshirani, R. (1996), ‘Regression shrinkage and selection via the Lasso’, *The Journal of Royal Statistical Society, Series B* **58**(1), 267–288.

Voev, V. (2008), *Dynamic Modelling of Large Dimensional Covariance Matrices*, High Frequency Financial Econometrics, Springer-Verlag. Berlin.

Wang, Y. & Zou, J. (2009), ‘Vast volatility matrix estimation for high-frequency financial data’, *The Annals of Statistics* **38**(2), 943–978.

Wu, W. & Pourahmadi, M. (2003), ‘Nonparametric estimation of large covariance matrices of longitudinal data’, *Biometrika* **90**(4), 831–844.

Zheng, X. & Li, Y. (2010), On the estimation of integrated covariance matrices of high dimensional diffusion processes, Discusion paper, Business Statistics and Operations Management, Hong Kong University of Science and Technology.

Appendix A.

In this appendix we present the LAR algorithm with the Lasso modification proposed by Efron et al. (2004), which is an efficient way of computing the solution to any Lasso problem, especially when $T \ll p$.

Algorithm. LARS: Least Angle Regression algorithm to calculate the entire Lasso path

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $r = y - \bar{y}$, and $w_j = 0$ for $j = 1, \dots, p-1$.

2. Find the predictor x_j most correlated with r.

3. Move w_j from 0 towards its least-squares coefficient $\langle x_j, r\rangle$, until some other competitor x_k has as much correlation with the current residual as does x_j.

4. Move w_j and w_k in the direction defined by their joint least squares coefficient of the current residual on (x_j, x_k), until some other competitor x_l has as much correlation with the current residual. If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.

5. Continue in this way until all p predictors have been entered. After a number of steps no more than $\min(T-1, p)$, we arrive at the full least-squares solution.

Source: Hastie, Tibshirani & Friedman (2009)

Appendix B.

Table 4: Parameters used in the simulation.

Parameters for factor loadings				Parameters for factor returns			
$\boldsymbol{\mu}_{\boldsymbol{\lambda}}$	$\text{cov}_{\boldsymbol{\lambda}}$			$\boldsymbol{\mu}_{\boldsymbol{f}}$	$\text{cov}_{\boldsymbol{f}}$		
0.7828	0.029145			0.023558	1.2507		
0.5180	0.023873	0.053951		0.012989	−0.0349	0.31564	
0.4100	0.010184	−0.006967	0.086856	0.020714	−0.2041	−0.0022526	0.19303

Source: Fan et al. (2008).

Revista Colombiana de Estadística

Índice de autores del volumen 34, 2011

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. 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.
- 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].

[1] http://www.ctan.org/tex-archive/systems/win32/miktex/

[2] http://www.estadistica.unal.edu.co/revista

[3] http://www.statindex.org/CIS/homepage/keywords.html

- 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.

[4] http://tug.ctan.org/tex-archive/macros/latex/contrib/harvard