Revista Colombiana de Estadística

Volumen 36. Número 1 - junio - 2013 **ISSN 0120 - 1751**

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

Revista Colombiana de Estadística

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

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

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

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

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

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

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

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

Revista Colombiana de Estadística	Bogotá	Vol. 36	Nº 1
ISSN 0120 - 1751	COLOMBIA	junio-2013	Págs. 1-192

Contenido

Editorial

Leonardo Trujillo[a]

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

---

"*Esencialmente todos los modelos están errados, pero algunos son útiles*"
George Box (1919-2013)

Bienvenidos a la primera edición del volumen 36 de la Revista Colombiana de Estadística. En este número hemos mantenido la característica de ser una revista publicada totalmente en idioma inglés de acuerdo a los requisitos por ser los ganadores de una convocatoria interna en la Universidad Nacional de Colombia entre otras revistas (ver editorial de Diciembre 2011). Estamos muy orgullosos de anunciar que hemos mantenido nuestra categorización como revista A1 según la clasificación de Publindex (Colciencias) que clasifica las revistas del país y siendo ésta la máxima categoría. Gracias a todos los Comités Editorial y Científico y a nuestra asistente de la Revista, Patricia Chavez, puesto que este resultado es fruto de la continua ayuda obtenida por parte de todos ellos. Más información se encuentra disponible en `http://201.234.78.173:8084/publindex/EnIbnPublindex/resultados.do`

Los temas de este número varían sobre diversas áreas de la estadística: cuatro artículos en Probabilidad de Martínez, Moreno y Vergara; Martínez, Vergara y González; Ozel; y de Shahzad y Asghar; dos artículos en Análisis Multivariado de Ortiz, Rivera y Melo y de Pardo, Ortiz y Becue; dos artículos en Análisis de Regresión de López y de Pérez y González - Farías; un artículo en Econometría de Ramírez y un artículo en Muestreo de Yadav y Kadilar.

El Departamento de Estadística de la Universidad Nacional de Colombia se encuentra organizando el Simposio Internacional de Estadística desde 1990. En la versión 2013, el evento se llevará a cabo en el Hotel AR en Bogotá y algunos invitados internacionales incluyen a Agustín Maravall (Banco de España), Heleno Bolfarine (Universidad de Sao Paulo, Brasil), Jeff Wu (Georgia Tech University, USA), Jon Rao (Carleton University, Canadá), Luigi Spezia (Biomathematics and Statistics Scotland, UK) and Peter Green (University of Bristol, UK). A través de los años, este evento ha unido a la comunidad estadística en diferentes regiones de Colombia y ha contado con la cooperación de diversas universidades. Algunas veces el Simposio se ha enfocado en un área de la estadística en particular: análisis Bayesiano, análisis de regresión, análisis multivariado, control de calidad, diseño de experimentos, estadística no paramétrica, muestreo, series de tiempo. En los años más recientes, se ha enfocado en más de un área. Para más información acerca del

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

evento se puede consultar en http://www.simposioestadistica.unal.edu.co o con Luz Mery González en lgonzalezg@unal.edu.co.

Otro evento que tendrá lugar pronto en Colombia es el XIII CLAPEM (Latin American Congress of Probability and Mathematical Statistics), el cual se ha mantenido creciendo en cuanto al número de instituciones participantes y su organización. Este evento tendrá lugar en Septiembre, 2014 por primera vez en Colombia en el Hotel Caribe de la ciudad de Cartagena (www.hotelcaribe.com) con la ayuda del Capítulo Latinoamericano de la Sociedad Bernoulli. CLAPEM es la mayor conferencia que reúne a científicos en las áreas de Probabilidad y Estadística Matemática en la región y tiene lugar cada dos o tres años. Ha tenido lugar anteriormente en Argentina, Brasil, Chile, Cuba, México, Perú, Uruguay and Venezuela. Las actividades del CLAPEM incluyen cursos dictados por investigadores invitados, reuniones satélites, sesiones de contribuciones orales y por poster, cursos cortos y sesiones temáticas. Los siguientes investigadores han confirmado su participación como expositores de cursos cortos o conferencias plenarias: Alison Etheridge (Universidad de Oxford, UK), Bin Yu (Universidad de California en Berkeley, USA), Paul Embrechts (ETH Zurich, Suiza), Carin Ludeña (Instituto Venezolano de Investigaciones Científicas, Venezuela), Gerard Biau (Universite Pierre et Marie Curie e Institut Universitaire, Francia), Roberto Imbuzeiro (IMPA, Brasil), Sourav Chatterjee (New York University y Stanford University, USA), Thomas Mikosch (Universidad de Copenhagen, Dinamarca), Víctor Rivero (CIMAT, México). El XIII CLAPEM es organizado por la Sociedad Bernoulli, Universidad Nacional de Colombia (sedes Bogotá y Medellín), Universidad de los Andes, Universidad de Cartagena, Universidad Industrial de Santander, Universidad Central, Universidad Santo Tomás, Universidad Sergio Arboleda, Universidad Pedagógica y Tecnológica de Colombia, EAFIT, Universidad del Norte, Universidad Antonio Nariño y la Universidad de Antioquia. Para más detalles se puede contactar a Ricardo Fraiman (presidente del XIII CLAPEM, fraimanricardo@gmail.com) o Leonardo Trujillo (ltrujilloo@unal.edu.co).

El Departamento de Estadística de la Universidad Nacional de Colombia acaba de aparecer en la lista de los 200 mejores departamentos en esta área en el mundo. Esto según la más reciente versión del QS World University Ranking (http://www.topuniversities.com/university-rankings/university-subject-rankings/2013/statistics-and-operational-research). La Universidad de California en Berkeley ocupa el primer lugar seguida por Massachusetts Institute of Technology (MIT), Stanford University, the Georgia Institute of Technology, la Universidad Nacional de Singapur y el Imperial College de Londres en Inglaterra. Solamente seis universidades latinoamericanas aparecen en esta lista con universidades de Brasil, Chile y Colombia. El QS World University Rankings by Subject 2013 evaluó 2,858 universidades alrededor del mundo y clasificó solamente a 678. En el análisis y clasificación se tuvieron en cuenta variables como la reputación de los programas entre empleadores y académicos, número de publicaciones y citaciones a nivel internacional en bases de datos de Scopus, entre otras. QS es la única clasificación que toma en cuenta la opinion de los empleadores de acuerdo con Ben Sowter, jefe de investigaciones de QS. Más información acerca de los resultados para Colombia en otras áreas del conocimien-

conocimiento se puede encontrar en http://www.eltiempo.com/vida-de-hoy/educacion/ARTICULO-WEB-NEW_NOTA_INTERIOR-12789070.html. También estamos muy orgullosos de anunciar que uno de nuestros exalumnos, Iván Díaz, obtuvo el Premio Erich L. Lehmann Citation por su tesis doctoral en estadística teórica en la Universidad de California en Berkeley (programa académico en el primer lugar del ranking mencionado anteriormente).

En marzo pasado, un estadístico muy eminente falleció lamentablemente: George Box, sin duda una de las grandes mentes estadísticas del siglo XX. Sin embargo, él se llamaba a sí mismo como un estadístico por accidente. Trabajó en áreas como análisis de series de tiempo, control de calidad, diseño de experimentos, inferencia bayesiana. Nació en Inglaterra pero desde 1960 se marchó para la Universidad de Wisconsin-Madison (USA) donde creó el Departamento de Estadística. Estaba casado con una de las hijas de Ronald Fisher (otro de los grandes contribuidores a la estadística). Su nombre está asociado a los modelos de Box-Jenkins, las transformaciones de Box-Cox y los diseños Box-Behnken. Una nueva autobiografía "The Accidental Statistician" se encuentra disponible en versión Kindle (http://www.amazon.com/An-Accidental-Statistician-Memories-ebook/dp/B00BU8Z3R6).

Editorial

Leonardo Trujillo[a]

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

---

"*Essentially all models are wrong, but some are useful*"
George Box (1919-2013)

Welcome to the first issue of the 36th volume of the Revista Colombiana de Estadistica (Colombian Journal of Statistics). We have kept, as in recent issues, the characteristic of being a Journal entirely published in English language as part of the requirements of being the winners (second year in a row) of an Internal Grant at the National University of Colombia among other many Journals (see editorial of December 2011). We are also very proud to announce that the Colombian Journal of Statistics have maintained its categorization as an A1 Journal by Publindex (Colciencias) which ranges the journals in the country, being A1 the maximum category. Thanks to all the Editorial and Scientific Committees and Patricia Chavez, our assistant in the Journal, as this is a result of the continuous help obtained from all of them. More information available at `http://201.234.78.173:8084/publindex/EnIbnPublindex/resultados.do`

The topics in this current issue range over diverse areas of statistics: four papers in Probability by Martínez, Moreno and Vergara; Martínez, Vergara and González; Ozel; and by Shahzad and Asghar; two papers in Multivariate Analysis by Ortiz, Rivera and Melo and by Pardo, Ortiz and Becue; two papers in Regression Analysis by Lopez and by Perez and Gonzalez-Farias; one paper in Econometrics by Ramirez and one paper in Survey Sampling by Yadav and Kadilar.

The Department of Statistics at the National University of Colombia is organizing the International Symposium in Statistics since 1990. In the 2013 version, the event is going to be held at the AR Hotel in the capital city of Bogota and some invited speakers include Agustin Maravall (Bank of Spain), Heleno Bolfarine (University of Sao Paulo), Jeff Wu (Georgia Tech), Jon Rao (Carleton University), Luigi Spezia (Biomathematics and Statistics Scotland) and Peter Green (University of Bristol). Through the years, this event has allowed the bonding of the statistical community in different regions of Colombia and has counted with the cooperation of other universities. Sometimes the Symposium has focused on a single topic of interest, such as regression analysis, time series, sampling, design of experiments, multivariate analysis, Bayesian analysis, nonparametric statistics, Statistical Quality Control and Industrial Statistics. For the most recent events,

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

the Symposium have spanned in more than one subject. You can find more information about this event at http://www.simposioestadistica.unal.edu.co/ or with Luz Mery Gonzalez at lgonzalezg@unal.edu.co.

Another event to be held in Colombia soon is the XIII CLAPEM (Latin American Congress of Probability and Mathematical Statistics), which keeps growing in the number of participant institutions and its organization. This event will be held in September, 2014 for the first time in Colombia at the Caribe Hotel (www.hotelcaribe.com) at the city of Cartagena with the help of the Latin American Chapter of the Bernoulli Society. CLAPEM is the largest conference gathering scientists in the particular areas of Probability and Mathematical Statistics in the region and takes place every two/three years. It has already been organized in Argentina, Brazil, Chile, Cuba, Mexico, Peru, Uruguay and Venezuela. The CLAPEM activities include lectures held by invited researchers, satellite meetings, sessions of oral and poster contributions, short courses, and thematic sessions.

The following researchers have confirmed their participation for short courses or plenary conferences: Alison Etheridge (University of Oxford, UK), Bin Yu (University of California at Berkeley, USA), Paul Embrechts (ETH Zurich, Switzerland), Carin Ludeña (Instituto Venezolano de Investigaciones Cientificas, Venezuela), Gerard Biau (Universite Pierre et Marie Curie and Institut Universitaire de France), Roberto Imbuzeiro (IMPA, Brazil), Sourav Chatterjee (New York University and Stanford University, USA), Thomas Mikosch (University of Copenhagen, Denmark), Victor Rivero (CIMAT, Mexico). The XIII CLAPEM is organized by the Bernoulli Society, Universidad Nacional de Colombia, Universidad de los Andes, Universidad de Cartagena, Universidad Industrial de Santander, Universidad Central, Universidad Santo Tomás, Universidad Sergio Arboleda, Universidad Pedagógica y Tecnológica de Colombia, EAFIT, Universidad del Norte, Universidad Antonio Nariño and Universidad de Antioquia. If you are interested you can also get more details with Ricardo Fraiman (president of the XIII CLAPEM, fraimanricardo@gmail.com) or Leonardo Trujillo (ltrujilloo@unal.edu.co).

The Department of Statistics at the National University of Colombia appears in the 200 best in the world among many other universities specifically for this subject, according to the more recent version of the QS World University Ranking (http://www.topuniversities.com/university-rankings/university-subject-rankings/2013/statistics-and-operational-research). The University of California at Berkeley is holding the first position followed by the Massachusetts Institute of Technology (MIT), Stanford University, the Georgia Institute of Technology, the National University of Singapore and the Imperial College of London at the UK. Only six Latinoamerican universities are also in this list with universities from Brazil, Chile and Colombia. The QS World University Rankings by Subject 2013 evaluated 2,858 universities around the globe and classified only 678. In the analysis and classification some variables that were taken into account were the reputation among employers and academics, number of publications and citations in an international scale through the Scopus database, among other criteria. QS is the only classification that takes into account the opinion of the employers according to Ben Sowter, chief of research at QS. More information can be found about Colombia s results in other subjects at http://www.eltiempo.com/

vida-de-hoy/educacion/ARTICULO-WEB-NEW_NOTA_INTERIOR-12789070.html. Also we are very proud to announce that one of our alumni, Ivan Diaz, just got the Erich L. Lehmann Citation Award for an outstanding PhD. Dissertation in theoretical statistics at the University of California at Berkeley (university at the first place in the ranking).

Last March, a very eminent statistician has passed away: George Box, without doubt one of the great statistical minds of the 20th century. However, he called himself as an accidental statistician. He worked in the areas of Bayesian inference, design of experiments, quality control and time series analysis. He was born in England but since 1960 he moved to the University of Wisconsin-Madison (USA) where he created the Department of Statistics. He was married to one of Ronald Fisher's daughters (another big one contributor to statistics). His name is associated to the Box-Jenkins models, Box-Cox transformations and Box-Behnken designs. A new autobiography "The Accidental Statistician" is available in the Kindle version to download (http://www.amazon.com/An-Accidental-Statistician-Memories-ebook/dp/B00BU8Z3R6).

Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 1 a 21

A Bayesian Approach to Parameter Estimation in Simplex Regression Model: A Comparison with Beta Regression

Un enfoque bayesiano para la estimación de los parámetros del modelo regresión Simplex: una comparación con la regresión Beta

Freddy Omar López[a]

Universidad de Valparaíso, Valparaíso, Chile

Abstract

Some variables are restricted to the open interval $(0, 1)$ and several methods have been developed to work with them under the scheme of the regression analysis. Most of research consider maximum likelihood methods and the use of Beta or Simplex distributions.

This paper presents the use of Bayesian techniques to estimate the parameters of the simplex regression supported on the implementation of some simulations and a comparison with Beta regression. We consider both models with constant variance and models with variance heterogeneity. Regressions are exemplified with heteroscedasticity.

***Key words*:** Beta distribution, Gibbs sampler, Heterogeneous, Proportions, Simplex distribution, Variances.

Resumen

Algunas variables están restringidas al intervalo abierto $(0, 1)$ y para trabajar con ellas se han desarrollado diversos métodos bajo el esquema del análisis de regresión. La mayoría de ellos han sido concebidos originalmente para ser estimados por métodos de máxima verosimilitud. Los más naturales parecen descansar especialmente sobre las distribuciones Beta o Simplex.

En este trabajo se presenta el uso de técnicas Bayesianas para la estimación de los parámetros de la regresión Simplex respaldada con la aplicación de algunas simulaciones y comparaciones con la regresión Beta. Se presentan resultados para modelos de varianza constante y de varianza heterogénea para cada individuo. Se presenta un ejemplo con datos reales.

***Palabras clave*:** distribución beta, distribución simplex, muestreador de Gibbs, proporciones, varianza heterogénea.

[a]PhD Student. E-mail: freddy.vate01@gmail.com

1. Introduction

Researchers frequently are dealing with situations where they are interested in modelling proportions, percentages or values within the open interval $(0, 1)$, according to one or several covariates, within the architecture of the regression models. This has usually been addressed with different approaches, including: linear regression, logistic regression, nonlinear regression, tobit regression, among others. However, most of them are not the natural way of working with such variables.

For this type of variable, the normal assumption, underlying in most of the mentioned techniques, it is not supported, invalidating conclusions that could be obtained from these results. Response variable's asymmetry and multicollinearity are two of the most frequent problems which the normal model cannot deal with.

In this situation, some alternatives have been developed such as Beta regression which take the general linear model advantages and the Simplex distribution, which is part of a more general class of models, the *dispersion models*.

These mentioned techniques have been developed to analyze variables that belong to the open interval $(0, 1)$ and not to $[0, 1]$. This distinction has been made by Kieschnick & McCullough (2003) in a comparative study as other authors. They recommended to use the Beta distribution or a quasi-likelihood based model when it is required to work with this type of variable.

As a comment to Paolino (2001), Buckley (2003) used the Bayesian paradigm to estimate the parameters from a Beta regression through the Metropolis-Hasting algorithm with non-informative previous distributions. This model contemplates the posibility to manage the heterogenity, besides the mean, by using two submodels corresponding to the location and dispersion submodels (Smithson & Verkuilen 2006). The research done by Paolino (2001) originally used a maximum likelihood method to estimate parameters. Ferrari & Cribari-Neto (2004) also apply this method.

Song, Qiu & Tan (2004) developed a similar model considering two submodels (one for a location parameter and another for a dispersion parameter) with a response simplex variable. The method to estimate the parameters by these authors was the generalized estimating equations (GEE).

In this work we consider a Bayesian approach for the estimation of the regression parameters and some simulations using the Gibbs sampler. Previous distributions to regression parameters have been normal with a high variance. Also, the estimation methods will be applied to a real dataset.

The main purpose of this work is to present the estimation by Bayesian methods of the simplex regression's parameters. Additionally, since Beta regression has the same objective of modelling proportions and rates, both methods will be compared some datasets generated by one or the other underlying model. We will be make emphasis on the details of the simplex distribution given the fact that the features of the beta distribution enjoy more fame in the literature than the simplex model.

This paper is structured as follows: in the Section 2 we present the simplex distribution, simplex regressions and the estimation method used in this investigation. Also, the beta regression and the comparison strategy in order to compare both models. In Section 3 we present some simulations and an application to real dataset. Finally in Section 4 some conclusions about this work.

2. Regression Models

2.1. Dispersion Models and Simplex Distribution

The simplex distribution is a distribution that belongs to the family of *dispersion models*, with location and dispersion parameters μ and σ^2, respectively (also abreviatted as $\mathsf{DM}(\mu, \sigma^2)$).

The *exponential dispersion* family density (ED) has the form

$$p(y;\theta,\phi) = \exp\left\{\frac{y\theta - \kappa(\theta)}{a(\theta)} + C(y,\phi)\right\},\ y \in \mathcal{C} \tag{1}$$

for some functions $a(\cdot), \kappa(\cdot)$ y $C(\cdot)$ with parameters $\theta \in \Theta$ and $\phi > 0$ and $\mathcal{C}$ is the support of the density. In particular, it is known that κ is the cumulant generating function. Note that ED is the classical *exponential family* of the random component in the GLM framework.

The general form of a dispersion model is

$$p(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\},\ y \in \mathcal{C} \tag{2}$$

where $\mu \in \Omega, \sigma > 0$ and $a \geq 0$ is a normalizer term, independent of μ. Function d is known as the *unit deviance* and is defined in $(y,\mu) \in (\mathcal{C},\Omega)$ and it must satisfy some additional properties (Song 2007).

A simple advantage over the classical exponential family parametrization in (1) is that both, mean and dispersion parameters, μ and σ^2, are explicitly in the density expression (2) whereas in (1), $\mu = E(Y) = \kappa'(\theta)$.

More precisely the parameter $\mu = E(Y)$ and $\mathrm{Var}(Y) = \frac{\sigma^2}{V(\mu)}$, where $V(\mu)$ is directly related with $d(\cdot;\cdot)$, i.e.

$$V(\mu) = \frac{2}{\left.\frac{\partial^2 d(y;\mu)}{\partial \mu^2}\right|_{y=\mu}},\ \mu \in \Omega$$

This function is known as the "unit variance function".

Specifically, if y follows a simplex distribution, that is $y \sim S^-(\mu;\sigma^2)$, then (2) takes the form

$$p(y;\mu,\sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\},\ y \in (0,1),\ \mu \in (0,1) \tag{3}$$

In particular, where

$$a(y;\sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-\frac{1}{2}}$$

and

$$d(y;\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2},\ y \in (0,1),\ \mu \in (0,1)$$

It follows that $\mathsf{E}\{d(Y;\mu)\} = \sigma^2$, $\mathsf{E}\{d'(Y;\mu)\} = 0$, $\mathsf{Var}\{d(Y;\mu)\} = 2(\sigma^2)^2$. These and others features can be studied in detail at Song (2007). Other inferential properties can be studied in the seminal paper by Barndorff-Nielsen & Jørgensen (1991).

The distribution can have one or two modes and can take the approximate shape of a bell, U, J, or L (also known as reverse-J) for different combinations of its parameters. It is important to note that the simplex distribution cannot emulates a flat distribution as the uniform distribution on the interval $(0, 1)$.

Figure 1 presents several examples: simplex distributions with mean values: $0.1, 0.25, 0.50, 0.75$ and 0.90 with different dispersion parameters. Note that when the second parameter is increased, the curves are becoming flatter.

2.2. Simplex Regression Model

2.2.1. Introduction

Let be $Y_1, \ldots, Y_n$ independent random variables following the distribution in equation (3) with mean μ_i and dispersion parameter σ_i^2, and let be $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{iq})$, $i = 1, \ldots, n$, vectors of covariate information. It is important to note that covariables $\mathbf{x}$ and $\mathbf{w}$ can be identical or they could be subsets of each other. We want to model the mean value μ_i and the dispersion parameter σ_i^2.

Similar to Cepeda & Gamerman (2001), Smithson & Verkuilen (2006) and Song et al. (2004), two link functions, g and h will be considered one for each parameter in the simplex distribution.

A convenient function g for the mean is the logit function, which ensures the parameter μ is in the open interval $(0, 1)$. More specifically

$$g(\mu_i) = \log \frac{\mu_i}{1-\mu_i} = \mathbf{x}_i^\top \boldsymbol{\beta} \tag{4}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ is a vector of unknown parameters. Equation (4) is also known as the *location submodel*.

The logit function has an extensive application in the statistic field. This transformation helps to give answers in terms of the *odds ratio*. This is because the odd ratio between the predictive variable and its response variable can be found by using the relation $\mathsf{OR} = \exp(\beta_k)$, $k = 1, \ldots, p$.

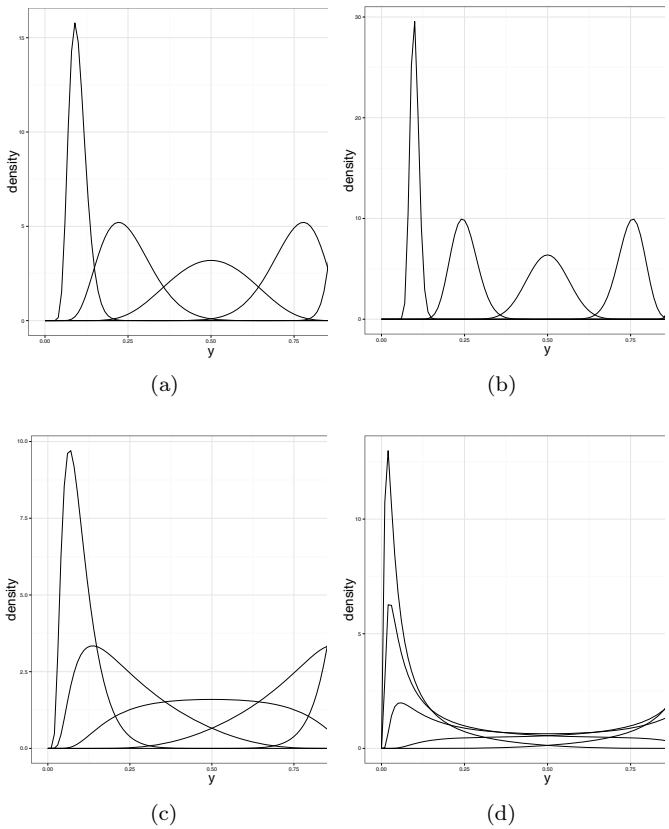


FIGURE 1: Different shapes for the simplex distribution. The distributions have as the mean value parameter $\mu = 0.1, 0.25, 0.5, 0.75, 0.9$ and different values for dispersion. (a) $\sigma = 1$; (b) $\sigma = 0.5$; (c) $\sigma = 2$ and (d) $\sigma = 5$.

On the other hand, the dispersion parameter σ_i^2 must be positive and a function h that enjoys this property is the logarithm function. So

$$h(\sigma_i^2) = \log(\sigma_i^2) = \mathbf{w}_i^\top \boldsymbol{\delta} \tag{5}$$

where $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_q)$ is a vector of unknown parameters that must be estimated. The equation (5) is known as the *dispersion submodel*.

2.2.2. Parameter Estimation

Maximum Likelihood

The classical theory of maximum likelihood estimation for the exponential family models (McCullagh & Nelder 1989) is very related with the maximum likelihood estimation for dispersion models as a special case. In the specific case

of simplex distribution and the general linear model the score equation (derivative of the likelihood with respect to parameters) is a given by

$$\sum_{i=1}^{n} \mathbf{x}_i\{\mu_i(1-\mu_i)\}\delta(y_i;\mu_i) = 0 \tag{6}$$

where

$$\delta(y;\mu) = \frac{y-\mu}{\mu(1-\mu)}\left\{d(y;\mu) + \frac{1}{\mu^2(1-\mu)^2}\right\}$$

Equation (6) is solved using Newton-Raphson or quasi-Newton algorithm.

In particular, it is necessary to introduce an estimation of the dispersion parameter σ^2. In this situation it is common to replace σ^2 with

$$\widehat{\sigma}^2 = \frac{1}{(n-p+1)}\sum_{i=1}^{n} d(y_i;\widehat{\mu}_i)$$

Interested readers are referred to Jørgensen (1997) and Song (2007) for more details. In this paper the maximum likelihood method is not considered.

Markov Chain Monte Carlo Sampling

With the aim of estimating the parameters of equations (4) and (5), we specify the likelihood function

$$L(\boldsymbol{\beta},\boldsymbol{\delta}) = \prod_{i=1}^{n} a(y_i; h^{-1}(\mathbf{w}_i^\top\boldsymbol{\delta}))\exp\left\{-\frac{1}{2h^{-1}(\mathbf{w}_i^\top\boldsymbol{\delta})}d(y_i; g^{-1}(\mathbf{x}_i^\top\boldsymbol{\beta}))\right\} \tag{7}$$

which posterior distribution is expressed as

$$p((\boldsymbol{\beta},\boldsymbol{\delta}) \mid \mathbf{y}) \propto L(\boldsymbol{\beta},\boldsymbol{\delta})p(\boldsymbol{\beta},\boldsymbol{\delta}) \tag{8}$$

where $p(\boldsymbol{\beta},\boldsymbol{\delta}) = p(\boldsymbol{\beta})p(\boldsymbol{\delta})$ are the previous distribution of parameters under the assumption that they are independent to each other. In this work it is assumed that each parameter $\beta_i, i = 1,\ldots,p$ and $\delta_j, j = 1,\ldots,q$ follow a non informative distribution centered at 0 and a large variance (about $1,000$). With this information, it is possible to use several Bayesian mechanisms in order to estimate the parameters. We have chosen a Gibbs sampling approach due to because the relative ease to be implemented.

In order to define the Bayesian regression modelling framework, we specify

$$\begin{aligned} &y_i \mid \mu_i,\sigma_i^2 \sim S^-(\mu_i,\sigma_i^2)\\ &g(\mu_i) = \mathbf{x}_i^\top\boldsymbol{\beta}\\ &h(\sigma_i^2) = \mathbf{w}_i^\top\boldsymbol{\delta} \end{aligned} \tag{9}$$

It is important to note that the models in this section are applicable to response variables y which range strictly in the open interval $(0,1)$. However, in some

situations, it is possible to have data where $y = 0$ or $y = 1$ (for instance, it can be the case where none person support the candidate's management; or that 100% of individuals under observation in a clinical trial have had reacted positively to certain stimuli). This situation can be addressed with different strategies. One of them is to replace all values 0 by a very small quantity $\epsilon > 0$ and all 1 values by $1 - \epsilon$ respectively. In other situations, when the theorical maximum and minimum values, β and α, are known the followings can be used

$$y^{\mathsf{new}} = \frac{(n-1)(y-\alpha)}{(\beta-\alpha)n} + \frac{1}{2n} \tag{10}$$

where n is the length of y. These approximations have been considered in the context of Beta regression by Smithson & Verkuilen (2006), Zimprich (2010), Verkuilen & Smithson (2011) and Eskelson, Madsen, Hagar & Temesgen (2011). This approach is not considered in this work.

2.3. Comparison to the Beta Regression Model

Beta regression has been studied with much interest on the last years (Ferrari & Cribari-Neto 2004, Ospina & Ferrari 2010, Cribari-Neto & Zeileis 2010, Cepeda & Garrido 2011, Cepeda 2012). In order to model proportions and rates.

The probability density function of a Beta distribution is given by

$$p(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1},\ 0 < y < 1$$

where Γ is the gamma function.

Considering $\mu = \frac{p}{p+q}$ and $\phi = p + q$ this produces $p = \mu\phi$ and $q = (1-\mu)\phi$. This will be the parametrization used in this work. A different parametrization based on mean and variance is studied by Cepeda (2012).

The shape of this distribution could have a variety of options. At most, it could have a single mode or a single antimode; it can show a bell-shaped, J and L-shaped and, among its particular cases, are the triangular distribution, uniform distribution and power function distribution (Johnson, Kotz & Balakrishnan 1994).

Beta regression is the most adequate model to be compared to the simplex regression because it is possible to model individual dispersion on the data (Cribari-Neto & Zeileis 2010).

It has been estimated traditionally using maximum likelihood methods but also Bayesian methods (Buckley 2003, Branscum, Johnson & Thurmond 2007, Cepeda & Garrido 2011, Cepeda 2012). In this work Bayesian methods will be used in order to estimate the parameters for Simplex and Beta regressions.

2.4. Model Comparison

2.4.1. Deviance Information Criterion

A way to compare models from the Bayesian perspective is through the DIC measure (Spiegelhalter, Best, Carlin & van der Linde 2002, Gelman, Carlin, Stern & Rubin 2003). This measure uses the *deviance* which is defined in its general form as

$$D(y, \theta) = -2 \log p(y|\theta)$$

where $p(y \mid \theta)$ is the likelihood of the data and θ are the parameters of the model. This measure depend both upon θ as y.

A measure which depend only of data y is $D_{\widehat{\theta}}(y) = D(y, \widehat{\theta}(y))$, which uses a point estimator of θ and is computed from simulations. The average over the posterior distribution is given by $D_{\text{avg}} = \text{E}(D(y, \theta) \mid y)$, whose estimator is

$$\widehat{D}_{\text{avg}}(y) = \frac{1}{n} \sum_{i=1}^{n} D(y, \theta^i)$$

Another important measure, known as the *effective number of parameters* is defined as

$$p_D = \widehat{D}_{\text{avg}}(y) - D_{\widehat{\theta}}(y)$$

Finally, the *deviance information criterion* (DIC) is defined by

$$\text{DIC} = 2\widehat{D}_{\text{avg}}(y) - D_{\widehat{\theta}}(y)$$

with smaller values suggesting a better-fitting model.

2.4.2. Comparison of Ordered Simulated Data Against Ordered Observed Data

A strategy to compare the performance of the models is simulate replicated data y^{rep}, and compare it with the real data, y. The comparison can be done ordering the simulated values, $y^{\text{rep}}_{(i)}$, and displaying it against the real ordered data, $y_{(i)}$. If at the moment of plotting, they are close to an identity function, then we have evidences of a good model. Moreover, we can appreciate values that can be outliers.

To create simulated data, y^{rep}, samples are taken following a model with the parameters $\widehat{\theta}$, estimated using real data (in this case, it will be sampled from Simplex and Beta distribution). To gain precision, it is usual to simulate several datasets and at the moment of plotting, to display empirical confidence intervals for each point of the observed data $y_{(i)}$.

3. Data Analysis

The following sections will show the performance of the simplex and beta regression. The simulation was followed using a similar scheme like the one by Song et al. (2004).

In each Section of 3.1 two types of dataset will be simulated. One, keeping a constant dispersion and another varying the dispersion cross the individuals. In Section 3.1.1 all data follow a simplex distribution and simplex and beta models are considered. In a similar way, data in the Section 3.1.2 lie under a beta distribution and the models used to these data are beta and simplex.

All simulations and computations were done using the R software (R Development Core Team 2011). Bayesian estimation was done using the Gibbs sampling using the `R2OpenBUGS` and `rjags` libraries (Sturtz, Ligges & Gelman 2005, Martyn 2011). All chains have the minimum requirements to think they have converged (i.e. Geweke diagnostic, Gelman-Rubin diagnostic, autocorrelation).

3.1. Simulation Study

3.1.1. Simulating Simplex Data

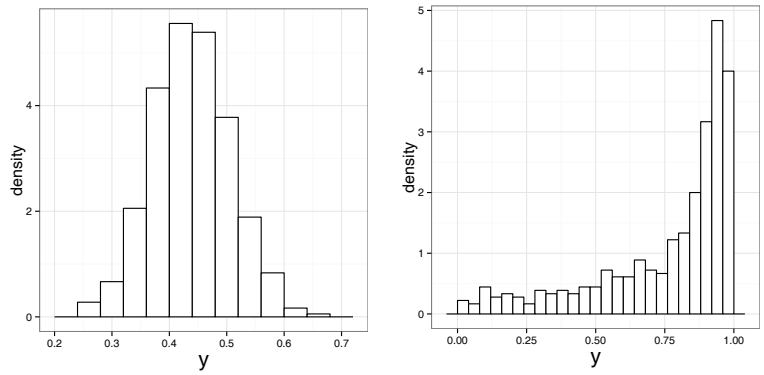
Firstly 450 independent observation $y_i, i = 1, \ldots, 450$ were obtained, belonging to a Simplex distribution with parameters (μ_i, σ^2) with the following specifications

$$\begin{cases} \text{logit}(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 S_i \\ \log(\sigma^2) = \delta_0 \end{cases} \tag{11}$$

where the variable $T \in \{-1, 0, 1\}$ emulates the level of some drug and $S \in \{0, \ldots, 6\}$ suggests the illness severity. To each level of T 150 individuals were taken and from S a random sample based on a binomial distribution was taken with parameters $n = 7$ y $p = 0.5$.

Parameters of equation (11) have been fixed to emulate various shapes of y (for instance: bell-shaped, J, L, U). Some of these shapes are plotted on figure 2.

After applying the model strategy in (9) the results can be appreciated in Table 1 and some of its realizations can be seen in Figure 3. All parameters were estimated with a four-chain run of 30,000 iterations length. Four chains of 30,000 length each were estimated and there its first 15,000 values were discarded from each one of them. It is important to note that in general, simplex estimation of parameters is close to real values, however, it seems there is a tendency when δ_0 increases then β_j, $j = 0, 1, 2$ are distant from real values. Moreover, we note that when y variable is bell-shaped then the estimated location parameters using beta or simplex model are very similar. Coefficients marked with a † symbol means that its Bayesian confidence interval includes the 0 value.



(a) (b)

(c) (d)

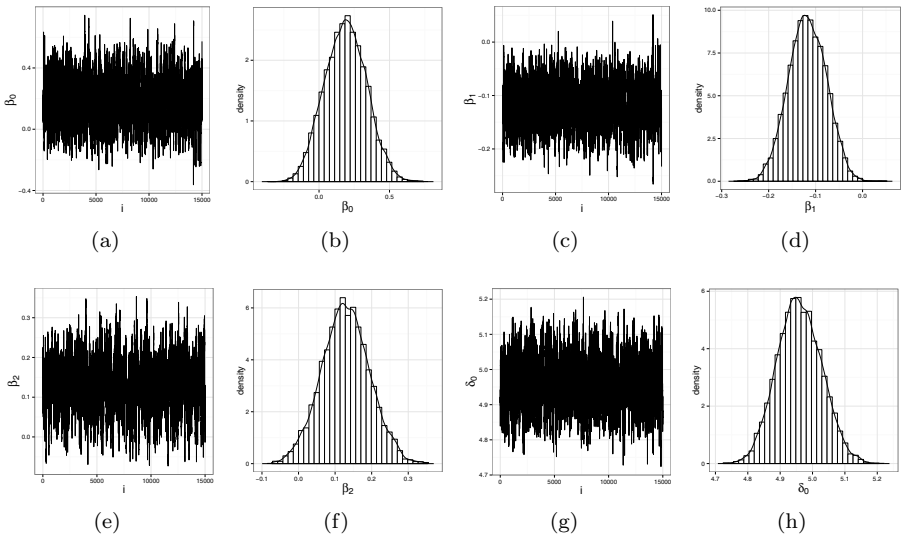
FIGURE 2: Simulations under homogeneous simplex models: (a) Bell-shaped ($\beta_0 = 0.1$, $\beta_1 = -0.1$, $\beta_2 = 0.1$, $\sigma = 0.5$); (b) J-shaped ($\beta_0 = -0.5$, $\beta_1 = 0.5$, $\beta_2 = -0.5$, $\sigma = \sqrt{15}$); (c) L-shaped ($\beta_0 = 0.1$, $\beta_1 = -0.1$, $\beta_2 = 0.1$, $\sigma = 0.5$); and, (d) U-shaped ($\beta_0 = 0.1$, $\beta_1 = -0.1$, $\beta_2 = 0.1$, $\sigma = 0.5$).

Additionally, DIC measures suggests both models are very competitive. Values estimated for the location submodels reach the greatest differences from real values when the shape of data y have form of U; in all cases the parameter of dispersion was estimated with high precision.

Second, several models were estimated varying the dispersion submodel according to the following specifications

$$\begin{cases} \text{logit}(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 S_i \\ \log(\sigma_i^2) = \delta_0 + \delta_1 T_i \end{cases} \tag{12}$$

where the parameters value β_j, $j = 0, 1, 2$ have been kept as in the previous exercise and δ_j, $j = 0, 1$ have been varied as shows Table 2 to preserve shapes similar to those shown in Figure 2.



(a) (b) (c) (d)

(e) (f) (g) (h)

FIGURE 3: Simulation of some chains for the homogeneous simplex model with U shaped: $\beta_0 = 0.1$, $\beta_1 = -0.1$, $\beta_2 = 0.1$, $\sigma^2 = 150$: (a) and (b) are summaries for parameter β_0; (c) and (d) for β_1; (e) and (f) for β_2; (g) and (h) for δ_0.

In the same way, the results from a four-chain run of 30,000 iterations (15,000 burn-in) are presented in Table 2. Additionally, when the shape of the distribution is like a bell, estimated parameters of the location submodel in simplex and beta model are extremely similar and according to DIC, the superiority of a model over the other is not pronounced. However, these estimated values are clearly distant from its true values. When the shape of the distribution is like a J or L then the estimated location parameters are closer to true values. Estimation of dispersion parameters were also close to its true values.

3.1.2. Simulating Beta Data

Also, several models following equations (11) and (12) were considered where the support distribution is beta. The structures were estimated with beta and simplex models and results are shown in Tables 3 y 4.

It can be appreciated in Table 3 that in some cases, when beta distribution is bell-shaped, some estimations (beta and simplex) tend to be similar in its location submodel. The beta estimation seems, however, to be more distant from its true parameters values; for instance, when the distribution has shape of U given that most of its location parameters include the 0 value inside its empirical highest posterior density.

The heterogeneous case (see Table 4) was not very different. Estimated parameters are more distant from its true values in most of the cases (shapes). In several of them, the DIC measure point out that the preferred model is the simplex one.

TABLE 1: Homogeneous simplex models: Results after fitting Simplex and Beta regression models.

Bell − shaped								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.1)	0.13	0.13	β_0 (0.1)	0.10	0.10	β_0 (0.1)	0.11	0.10
β_1 (-0.1)	-0.11	-0.11	β_1 (-0.1)	-0.10	-0.10	β_1 (-0.1)	-0.10	-0.10
β_2 (0.1)	0.10	0.10	β_2 (0.1)	0.10	0.10	β_2 (0.1)	0.11	0.11
$\delta_0(\log 0.1)$	-2.32	10.30	$\delta_0(\log 0.01)$	-4.64	14.93	$\delta_0(\log 0.25)$	-1.45	8.59
DIC	-1673	-1677	DIC	-2712	-2713	DIC	-1293	-1290
J								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (-0.5)	-0.44	-0.45	β_0 (-0.5)	-0.52	-0.34	β_0 (-0.5)	-0.60	$-0.23^{\dagger}$
β_1 (0.5)	0.49	0.49	β_1 (0.5)	0.51	0.43	β_1 (0.5)	0.54	0.38
β_2 (-0.5)	-0.51	-0.49	β_2 (-0.5)	-0.46	-0.40	β_2 (-0.5)	-0.59	-0.44
$\delta_0(\log 1)$	$0.03^{\dagger}$	6.90	$\delta_0(\log 5)$	1.60	4.19	$\delta_0(\log 15)$	2.77	2.63
DIC	-1298	-1142	DIC	-748.40	-611	DIC	-685	-481
L								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.5)	0.60	0.56	β_0 (0.5)	0.45	0.25	β_0 (0.5)	0.37	$0.04^{\dagger}$
β_1 (-0.5)	-0.53	-0.51	β_1 (-0.5)	-0.49	-0.41	β_1 (-0.5)	-0.45	-0.32
β_2 (0.5)	0.47	0.44	β_2 (0.5)	0.47	0.42	β_2 (0.5)	0.48	0.37
$\delta_0(\log 1)$	$0.00^{\dagger}$	6.91	$\delta_0(\log 5)$	1.62	4.48	$\delta_0(\log 15)$	2.72	2.74
DIC	-1246	-1123	DIC	-815	-699	DIC	-598	-457
U								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.1)	0.46	0.39	β_0 (0.1)	0.19	$0.13^{\dagger}$	β_0 (0.1)	0.18	$0.06^{\dagger}$
β_1 (-0.1)	-0.22	-0.18	β_1 (-0.1)	-0.11	$-0.07^{\dagger}$	β_1 (-0.1)	-0.12	$-0.08^{\dagger}$
β_2 (0.1)	0.10	$0.12^{\dagger}$	β_2 (0.1)	0.13	$0.09^{\dagger}$	β_2 (0.1)	0.13	0.12
$\delta_0(\log 50)$	3.89	0.58	$\delta_0(\log 100)$	4.51	0.08	$\delta_0(\log 150)$	4.96	$-0.31^{\dagger}$
DIC	-300	-125	DIC	-386	-201	DIC	-688	-385

3.2. Example with Real Data

In this section we study the relationship between the amount people of in poverty and the government form they have elected in some geographical region. We want to determine if some variables, traditionally indicators of poverty (number of people indeed poverty, suicide rate, Human Development Index) are associated with a political option in electoral preferences terms.

The relationship between these variables has been studied previously. For instance, it is documented that for some countries, suicide rates increases when a specific political party is in the government. Blakely & Collings (2002) commented that "suicide rates were indeed higher during periods of conservative government" for the investigation done with Australian data carried out by Page, Morrell & Taylor (2002). Shaw, Dorling & Smith (2002) analyzed data from England and Wales and reached similar conclusions to the point to add the subtitle to their investigation: *Do conservative governments make people want to die?*

Also there have been findings there exists out a significant association between general mortality and political preferences (Smith & Dorling 1996).

Data analyzed in this paper correspond to 322 of 335 municipalities in Venezuela (the position of Amazonas' Governor and others municipalities were not available for that election date). These data were taken from the website of the National Electoral Council, (CNE 2008) and the National Statistical Office, (INE 2008).

TABLE 2: Heterogeneous simplex models: Results after fitting Simplex and Beta regression models.

Bell − shaped								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.1)	-0.05†	-0.06†	β_0 (0.1)	0.12†	0.12†	β_0 (0.1)	-0.03†	-0.03†
β_1 (-0.1)	-0.06	-0.06	β_1 (-0.1)	-0.10	-0.10	β_1 (-0.1)	-0.07	-0.07
β_2 (0.1)	0.07†	0.07†	β_2 (0.1)	0.12	0.12	β_2 (0.1)	0.06	0.06
$\delta_0(1)$	1.06	4.13	$\delta_0(0.1)$	0.15	5.57	$\delta_0(0.3)$	0.29	5.33
$\delta_1(1)$	1.19	-1.90	$\delta_1(0.1)$	0.03†	-0.13†	$\delta_0(0.2)$	0.20	-0.36
DIC	-394	-374	DIC	-639	-634	DIC	-588	-584
J								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (-0.5)	-0.42	-0.41	β_0 (-0.5)	-0.71	-0.48	β_0 (-0.5)	-0.47	-0.05†
β_1 (0.5)	0.49	0.48	β_1 (0.5)	0.54	0.45	β_1 (0.5)	0.49	0.34
β_2 (-0.5)	-0.45	-0.45	β_2 (-0.5)	-0.49	-0.52	β_2 (-0.5)	-0.57	-0.59
$\delta_0(1)$	0.99	5.58	$\delta_0(2)$	2.01	3.95	$\delta_0(3)$	3.01	2.65
$\delta_1(1)$	0.99	-2.29	$\delta_1(1)$	1.01	-2.15	$\delta_0(1)$	1.07	-2.00
DIC	-994	-896	DIC	-783	-647	DIC	-808	-601
L								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.5)	0.60	0.50	β_0 (0.5)	0.33	0.23	β_0 (0.5)	0.30	-0.03†
β_1 (-0.5)	-0.52	-0.47	β_1 (-0.5)	-0.46	-0.41	β_1 (-0.5)	-0.43	-0.30
β_2 (0.5)	0.49	0.51	β_2 (0.5)	0.52	0.55	β_2 (0.5)	0.51	0.50
$\delta_0(1)$	1.02	5.28	$\delta_0(2)$	1.99	4.08	$\delta_0(3)$	3.04	2.49
$\delta_1(1)$	1.03	-2.34	$\delta_1(1)$	1.12	-2.39	$\delta_0(1)$	1.01	-1.67
DIC	-946	-818	DIC	-782	-668	DIC	-623	-483
U								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.1)	0.25	0.32	β_0 (0.1)	0.13†	0.18†	β_0 (0.1)	0.26†	0.19†
β_1 (-0.1)	-0.13	-0.13	β_1 (-0.1)	-0.10	-0.09	β_1 (-0.1)	-0.13	-0.12
β_2 (0.1)	0.14	0.18	β_2 (0.1)	0.12	0.12†	β_2 (0.1)	0.07†	0.03†
$\delta_0(3)$	2.98	1.59	$\delta_0(4)$	4.04	0.58	$\delta_0(5)$	4.98	-0.22
$\delta_1(1)$	1.20	-1.36	$\delta_1(1)$	1.00	-0.89	$\delta_0(1)$	1.06	-0.75
DIC	-188	-98	DIC	-252	-142	DIC	-728	-446

The response variable is the proportion of people who support with their votes the political proposal lead by Hugo Chávez.

Several models were adjusted to these data and the results can be seen in Table 5. In this Table, three models for the two underlying distributions were considered. The first of them (m_{s_0} and m_{b_0}) are the saturated models and m_{s_1} and m_{b_1} are the null models. Searching over additive structures in function of DIC give us as best models those labeled as m_{s_2} y m_{b_2}. For both, the same variables are significant for location and dispersion submodels. Note that, in general terms, coefficients for location submodels are very similar. This can be expected because the shape of the variable `% Chávez` is symmetric (see Figure 5 (b)).

Table 3: Homogeneous beta models: Results after fitting Beta and Simplex regression models.

Bell − shaped								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (0.1)	-0.07†	-0.15†	β_0 (0.1)	0.05†	0.07†	β_0 (0.1)	0.17	0.17
β_1 (-0.1)	-0.06	-0.05†	β_1 (-0.1)	-0.09	-0.09†	β_1 (-0.1)	-0.13	-0.13
β_2 (0.1)	0.13	0.15	β_2 (0.1)	0.12	0.12	β_2 (0.1)	0.09	0.09
$\delta_0(\log 30)$	3.29	1.65	$\delta_0(\log 50)$	3.80	1.25	$\delta_0(\log 200)$	5.36	0.30
DIC	-205	-176	DIC	-286	-285	DIC	-597	-593
J								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (-0.5)	-0.09	-0.30†	β_0 (-0.5)	-0.25†	-0.99	β_0 (-0.5)	-0.20†	-0.16†
β_1 (0.5)	0.28	0.35	β_1 (0.5)	0.35	0.61	β_1 (0.5)	0.38	0.44
β_2 (-0.5)	-0.24	-0.07†	β_2 (-0.5)	-0.35	-0.27	β_2 (-0.5)	-0.42	-0.55
$\delta_0(\log 1)$	0.52	5.12	$\delta_0(\log 5)$	1.51	4.46	$\delta_0(\log 15)$	2.81	3.60
DIC	-690	-861	DIC	-533	-467	DIC	-538	-347
L								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (0.5)	0.11†	0.79	β_0 (0.5)	0.24†	0.12†	β_0 (0.5)	0.25†	0.36†
β_1 (-0.5)	-0.31	-0.59	β_1 (-0.5)	-0.40	-0.45	β_1 (-0.5)	-0.42	-0.51
β_2 (0.5)	0.24	0.34	β_2 (0.5)	0.44	0.60	β_2 (0.5)	0.40	0.46
$\delta_0(\log 1)$	0.71	5.02	$\delta_0(\log 5)$	1.85	4.59	$\delta_0(\log 15)$	2.74	3.78
DIC	-752	-949	DIC	-625	-472	DIC	-587	-399
U								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (0.1)	0.03†	-0.02†	β_0 (0.1)	0.08†	-0.01†	β_0 (0.1)	-0.13†	-0.20†
β_1 (-0.1)	-0.09†	-0.08	β_1 (-0.1)	-0.07†	-0.04†	β_1 (-0.1)	-0.03†	-0.01†
β_2 (0.1)	0.06	0.14†	β_2 (0.1)	0.12†	0.08†	β_2 (0.1)	-0.02†	-0.02†
$\delta_0(\log 1)$	0.29	5.21	$\delta_0(\log 0.5)$	-0.25	5.62	$\delta_0(\log 0.25)$	-0.60	5.81
DIC	-177	67	DIC	-352	-289	DIC	-571	-756

A sample of predicted values for all models can be appreciated in Figure 5 (a). Note that the models give a *linear* prediction, that is, crossing the approximate mean of data for each value of variable `Mortality` according to its linear nature. Both models are quite similar and its fitting is displayed in Figure 5 (a). Figures 5 (c) and (d) show the average predicted values (and its empirical error bar) for each y_i point. There were simulated 100 datasets.

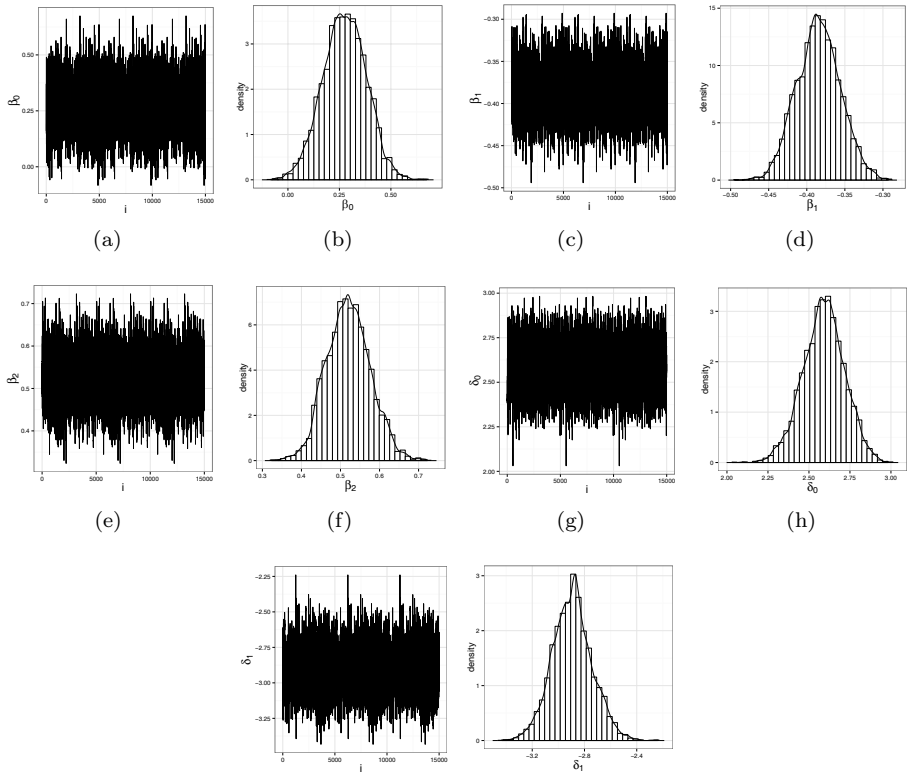


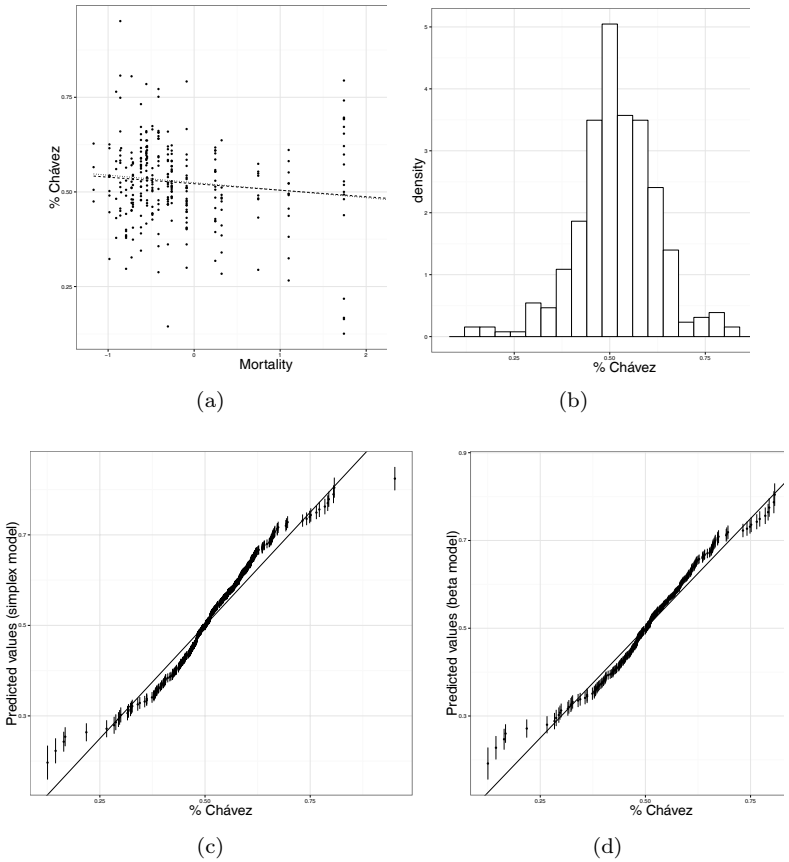
FIGURE 4: Simulation of some chains for the heterogeneous beta model with L shaped: $\beta_0 = 0.5$, $\beta_1 = 0.5$, $\beta_2 = 0.5$, $\delta_0 = 3$, $\delta_1 = 2$: (a) and (b) describes results for parameter β_0; (c) and (d) for β_1; (e) and (f) for β_2; (g) and (h) for δ_0; (i) and (j) for δ_1.

Table 4: Heterogeneous beta models: Results after fitting Beta and Simplex regression models.

Bell-shaped								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (0.1)	0.13†	0.34	β_0 (0.1)	0.08†	0.08†	β_0 (0.1)	0.11	0.11
β_1 (-0.1)	-0.11	-0.19	β_1 (-0.1)	-0.10	-0.10	β_1 (-0.1)	-0.10	-0.10
β_2 (0.1)	0.01†	0.03†	β_2 (0.1)	0.15	0.16	β_2 (0.1)	0.09	0.09
$\delta_0(3)$	2.84	2.23	$\delta_0(5)$	5.09	0.48	$\delta_0(10)$	9.97	-2.12
$\delta_1(1)$	0.97	-0.93	$\delta_1(1)$	1.08	-0.69	$\delta_1(5)$	5.12	-2.66
DIC	-152	-34	DIC	-544	-542	DIC	-1605	-1608
J								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (-0.5)	-0.17†	-0.35	β_0 (-0.5)	-0.55	-0.46	β_0 (-0.5)	-0.34	-0.41
β_1 (0.5)	0.37	0.51	β_1 (0.5)	0.44	0.41	β_1 (0.5)	0.46	0.49
β_2 (-0.5)	-0.50	-0.45	β_2 (-0.5)	-0.23	-0.09†	β_2 (-0.5)	-0.49	-0.44
$\delta_0(1)$	1.56	4.55	$\delta_0(1)$	2.29	3.98	$\delta_0(3)$	3.45	3.32
$\delta_1(1)$	0.18	-0.55	$\delta_1(5)$	2.73	-2.76	$\delta_1(2)$	1.50	-1.80
DIC	-703	-757	DIC	-884	-980	DIC	-783	-678
L								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (0.5)	-0.17†	-5.72	β_0 (0.5)	0.55	0.87	β_0 (0.5)	0.58	-3.16
β_1 (-0.5)	0.37	-1.06	β_1 (-0.5)	-0.43	-0.55	β_1 (-0.5)	-0.55	-0.58
β_2 (0.5)	-0.50	7.68	β_2 (0.5)	0.15	0.18	β_2 (0.5)	0.62	4.49
$\delta_0(1)$	1.56	36.77	$\delta_0(1)$	2.18	3.94	$\delta_0(3)$	3.27	16.11
$\delta_1(1)$	0.18	-27.85	$\delta_1(5)$	2.87	-2.64	$\delta_1(2)$	1.80	-12.65
DIC	-4121	11948	DIC	-824	-924	DIC	-1408	3936
U								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (0.1)	0.20†	-0.62	β_0 (0.1)	0.20†	0.51	β_0 (0.1)	0.01†	1.37
β_1 (-0.1)	-0.13	-0.21	β_1 (-0.1)	-0.14	-0.22	β_1 (-0.1)	-0.04†	-0.45
β_2 (0.1)	0.14	0.61	β_2 (0.1)	0.24	0.19	β_2 (0.1)	0.13	0.43†
$\delta_0(0.1)$	0.36	7.98	$\delta_0(0.1)$	0.40	4.92	$\delta_0(0.01)$	-0.11†	9.06
$\delta_1(0.1)$	0.13†	-1.97	$\delta_1(0.5)$	0.21†	-0.44	$\delta_1(0.05)$	0.02†	0.08†
DIC	-169	1358	DIC	-201	-63	DIC	-274	1450

Table 5: Parameter estimates using simplex and Bbeta regression for venezuelan election data (2008).

	Simplex model			Beta model		
	m_{s_0}	m_{s_1}	m_{s_2}	m_{b_0}	m_{b_1}	m_{b_2}
Location submodel						
Intercept	0.91	0.08	0.10	0.90	0.08	0.09
Suicides	0.03			0.02		
General Mortality	-0.10		-0.08	-0.07		-0.07
Households in poverty	-0.03			0.08		
IDH	-1.00			-1.04		
Dispersion submodel						
Intercept	-9.21	0.13	-12.23	15.53	5.84	24.31
Suicides	-0.18		-0.19	0.42		0.27
General Mortality	-0.03			-0.22		
Households in poverty	-1.69			3.44		
IDH	12.01		15.11	-13.03		-22.62
DIC	-469.03	-435.31	-476.23	-505.49	-489.87	-511.87



(a) (b)

(c) (d)

FIGURE 5: (a) Adjusted values for Simplex and Beta models; lines are nearly superimposed; (b) histogram of proportion of percentage of people that support Chávez; (c) ordered Chavism vs. ordered prediction based on Simplex model; (d) ordered Chavism vs. ordered prediction based on Beta model

4. Conclusions

This paper has shown how the Bayesian estimation can be applied on simplex model regression and, in addition, several simulations were performed to compare Simplex and Beta regressions. It was found that the estimation strategy produces better results when the true model is homogeneous. In particular, when the true model is homogeneous simplex, the estimates are closer to the true value parameters than the Beta model. Similar situations were found with the heterogeneous models. Most of the time, dispersion submodel parameters were estimated quite well even in the case where none parameter for the location submodel was near to its true value. Methodology was exemplified with a real dataset. For this, point estimates were pretty similar for both models: Simplex and Beta.

Further research could consider the natural extension to the (longitudinal) mixed models similar to those presented by Verkuilen & Smithson (2011) and Zimprich (2010) from the Bayesian perspective and supported by underlying simplex distribution assumption. Song et al. (2004) propose a simplex longitudinal data analysis in its marginal version.

Although, in the applications considered here, all data were inside the open interval $(0, 1)$; it is possible to model variables inside the closed interval $[0, 1]$ and there exist more adequate models such as those proposed by Cook, Kieschnick & McCullough (2008) and Ospina & Ferrari (2010).

Furthermore, it is important to investigate another alternatives for the link functions. As pointed out by Eskelson et al. (2011), the logit transformation is used because it offers an easy interpretation in terms of odds ratio but it is also possible to use the non-transformed variable. In relation with the beta regression, Giovanetti (2007) explores another alternatives to link functions and studies the empirical consequences having an incorrect specification.

In relation with Simplex regression residuals, Santos (2011) considers the situation when the parameters are estimated using the maximum likelihood method. Miyashiro (2008) proposes some diagnostic measures and performs comparisons with two real datasets estimating its parameters under Beta and Simplex assumptions. Results for those particular cases are very similar for location submodels. In that investigation, Miyashiro only studied homogeneous models using maximum likelihood.

Acknowledgments

I am in debt to Lisbeth Mora and Professor Daniel Paredes for reading the earlier stages of this research and for suggesting invaluable improvements. I also thank two anonymous referees for their comments that helped to increase the quality of this paper.

[Recibido: marzo de 2012 — Aceptado: enero de 2013]

References

Barndorff-Nielsen, O. E. & Jørgensen, B. (1991), 'Some parametric models on the simplex', *Journal of Multivariate Analysis* **39**, 106–116.

Blakely, T. & Collings, S. (2002), 'Is there a causal association between suicide rates and the political leanings of government?', *Journal of Epidemiology and Community Health* **56**(10), 722.

Branscum, A. J., Johnson, W. O. & Thurmond, M. C. (2007), 'Bayesian Beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth desease viruses', *Australian & New Zealand Journal of Statistics* **49**(3), 287–301.

Buckley, J. (2003), 'Estimation of models with Beta-distributed dependent variables: A replication and extension of Paolino's study', *Political Analysis* **11**, 204–205.

Cepeda, E. (2012), Beta regression models: Joint mean and variance modeling, Technical report, Universidad Nacional de Colombia.

Cepeda, E. & Gamerman, D. (2001), 'Bayesian modeling of variance heterogeneity in normal regression models', *Brazilian Journal of Probability and Statistics* **14**, 207–221.

Cepeda, E. & Garrido, L. (2011), Bayesian Beta regression models: Joint mean and precision modeling, Technical report, Universidad Nacional de Colombia.

CNE (2008), 'Consejo Nacional Electoral', `http://www.cne.gob.ve`.

Cook, D. O., Kieschnick, R. & McCullough, B. D. (2008), 'Regression analysis of proportions in finance with self selection', *Journal of Empirical Finance* **15**, 860–867.

Cribari-Neto, F. & Zeileis, A. (2010), 'Beta regression in R', *Journal of Empirical Finance* **34**(2), 1–24.

Eskelson, N. I., Madsen, L., Hagar, J. C. & Temesgen, H. (2011), 'Estimating Riparian understory vegetation cover with Beta regression and copula models', *Forest Science* **57**(3), 212–221.

Ferrari, S. L. P. & Cribari-Neto, F. (2004), 'Beta regression for modeling rates and proportions', *Journal of Applied Statistics* **31**(7), 799–815.

Gelman, A., Carlin, B. P., Stern, H. S. & Rubin, D. B. (2003), *Bayes and Empirical Bayes Methods for Analysis*, 2 edn, Chapman & Hall/CRC.

Giovanetti, A. C. (2007), Efeitos da especificaçao incorreta da funçao de ligaçao no modelo de regressão beta, Master's thesis, USP, Sao Paulo.

INE (2008), 'Instituto Nacional de Estadística', `http://www.ine.gob.ve`.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994), *Continuous Univariate Distributions*, Vol. 2, 2 edn, John Wiley & Sons.

Jørgensen, B. (1997), *The Theory of Dispersion Models*, Monographs on Statistics and Applied Probability, Taylor & Francis.

Kieschnick, R. & McCullough, B. D. (2003), 'Regression analysis of variates observed on (0,1): Percentages, proportions and fractions', *Statistical Modelling* **3**, 193–213.

Martyn, P. (2011), *rjags: Bayesian graphical models using MCMC*. R package version 3-5.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models, Second Edition*, number 37 *in* 'Monographs on Statistics and Applied Probability', London: Chapman & Hall.

Miyashiro, E. S. (2008), Modelos de regressão Beta e simplex para análise de proporçoes no modelo de regressão Beta, Master's thesis, USP, Sao Paulo.

Ospina, R. & Ferrari, S. L. (2010), 'Inflated beta distributions', *Statistical Papers* **51**, 111–126.

Page, A., Morrell, S. & Taylor, R. (2002), 'Suicide and political regime in New South Wales and Australia during the 20th century', *Journal of Epidemiology and Community Health* **56**(10), 766–772.

Paolino, P. (2001), 'Maximum likelihood estimation of models with Beta-distributed dependent variables', *Political Analysis* **9**, 325–346.

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

Santos, L. A. (2011), Modelos de Regressão Simplex: Resíduos de Pearson Corrigidos e Aplicações, PhD thesis, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

Shaw, M., Dorling, D. & Smith, G. D. (2002), 'Mortality and political climate: How suicide rates have risen during periods of Conservative government, 1901–2000', *Journal of Epidemiology and Community Health* **56**(10), 723–725.

Smith, G. D. & Dorling, D. (1996), "'I'm all right, John": Voting patterns and mortality in England and Wales', *British Medical Journal* **313**(21), 1573–1577.

Smithson, M. & Verkuilen, J. (2006), 'A better lemon squeezer? Maximum-likelihood regression with Beta-distributed dependent variables', *Psychological Methods* **11**, 54–71.

Song, X. K. (2007), *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer, New York.

Song, X., Qiu, Z. & Tan, M. (2004), 'Modelling heterogeneous dispersion in marginal models for longitudinal proportional data', *Biometrical Journal* **5**, 540–553.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002), 'Bayesian measures of model complexity and fit (with discussion)', *Statistical Methodology Series B* **64**(4), 583–639.

Sturtz, S., Ligges, U. & Gelman, A. (2005), 'R2WinBUGS: A package for running WinBUGS from R', *Journal of Statistical Software* **12**(3), 1–16.

Verkuilen, J. & Smithson, M. (2011), 'Mixed and mixture regression models for continuous bounded responses using the Beta distribution', *Journal of Educational and Behavioral Statistics* **000**, 1–32.

Zimprich, D. (2010), 'Modeling change in skewed variables using mixed Beta regression models', *Research in Human Development* **7**(1), 9–26.

Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 23 a 42

A Multi-Stage Almost Ideal Demand System: The Case of Beef Demand in Colombia

Sistema casi ideal de demanda multinivel: el caso de la demanda de carne de res en Colombia

Andrés Ramírez[a]

Departamento de Economía, Universidad EAFIT, Medellín, Colombia

Abstract

The main objective in this paper is to obtain reliable long-term and short-term elasticities estimates of the beef demand in Colombia using quarterly data since 1998 until 2007. However, complexity on the decision process of consumption should be taken into account, since expenditure on a particular good is sequential. In the case of beef demand in Colombia, a Multi-Stage process is proposed based on an Almost Ideal Demand System (AIDS). The econometric novelty in this paper is to estimate simultaneously all the stages by the Generalized Method of Moments to obtain a joint covariance matrix of parameter estimates in order to use the Delta Method for calculating the standard deviation of the long-term elasticities estimates. Additionally, this approach allows us to get elasticity estimates in each stage, but also, total elasticities which incorporate interaction between stages. On the other hand, the short-term dynamic is handled by a simultaneous estimation of the Error Correction version of the model; therefore, Monte Carlo simulation exercises are performed to analyse the impact on beef demand because of shocks at different levels of the decision making process of consumers. The results indicate that, although the total expenditure elasticity estimate of demand for beef is 1.78 in the long-term and the expenditure elasticity estimate within the meat group is 1.07, the total short-term expenditure elasticity is merely 0.03. The smaller short-term reaction of consumers is also evidenced on price shocks; while the total own price elasticity of beef is -0.24 in the short-term, the total and within meat group long-term elasticities are -1.95 and -1.17, respectively.

***Key words*:** Cointegration, Delta method, Demand system, Generalized method of moments, Monte Carlo Simulation.

[a]Associate professor. E-mail: aramir21@eafit.edu.co

Resumen

El objetivo más importante de este artículo es obtener estimaciones confiables de las elasticidades de la demanda de carne de res en Colombia para el largo y corto plazo utilizando información trimestral desde 1998 hasta 2007. Sin embargo, las decisiones que toman los consumidores se enmarcan en un ambiente complejo, puesto que el gasto en un bien particular se realiza de forma secuencial. En el caso particular de la demanda de carne de res en la economía colombiana, se propone un Sistema Casi Ideal de Demanda Multinivel. La novedad econométrica en este artículo es estimar simulatáneamente todos los niveles del modelo mediante el Método Generalizado de los Momentos; esto permite obtener una matriz conjunta de covarianzas de todos los parámetros, y así utilizar el Método Delta para calcular las desviaciones estándar de las elasticidades estimadas de largo plazo. Adicionalmente, este enfoque nos permite obtener estimaciones de las elasticidades en cada nivel, pero también, elasticidades totales que incorporan la interacción entre los niveles. Por otra parte, la dinámica de corto plazo se estudia a través de la estimación conjunta de la versión en Corrección de Errores del modelo; de esta forma, ejercicios de simulación Monte Carlo son reaizados para analizar el impacto sobre la demanda de carne de res debido a perturbaciones en diferentes niveles del proceso de toma de decisiones de los consumidores. Los resultados indican que aunque en el largo plazo la elasticidad estimada de la demanda de carne de res con respecto al gasto total es 1.78, y la elasticidad estimada de la demanda con respecto al gasto en cárnicos es 1.07, la elasticidad de la demanda con respecto al gasto total en el corto plazo es solo 0.03. La reducida reacción en el corto plazo también está presente ante perturbaciones en el precio; mientras que la elasticidad precio propia total de la demanda de carne de res es -0.24 en el corto plazo, las elasticidades total y al interior del grupo de cárnicos para el largo plazo son -1.95 y -1.17, respectivamente.

***Palabras clave*:** cointegración, método delta, método generalizado de los momentos, simulación Monte Carlo, sistema de demanda.

1. Introduction

Colombian beef demand is important for a number of reasons. Historically consumers have generally preferred beef to other types of meat. Beef accounted for approximately 60% of the total meat budget, compared to only 30% for poultry and 10% for pork. In addition, the beef sector is an important component of the Colombian economy, accounting for 3.4% of Gross Domestic Product in 2007 and providing 1.4 million jobs (DANE 2007). Moreover, the beef sector is a significant component of the Colombian exports to Venezuela, one of Colombia's most important trading partners. Approximately, 15% of Colombian beef production is exported to Venezuela. Recently, the Venezuelan Government decided to stop imports from Colombia as a result of political tensions. This trade restriction policy of Venezuela has generated preoccupation among specialists due to its consequences for the beef sector. Additionally, Colombia is currently negotiating international trade agreements with the United States and the European Union.

The implication is that the Colombian beef sector would have international competition from countries with high subsidies, as a consequence, given the trading conditions, the internal beef price would decrease. On the other hand, there is an asymmetric aspect that is necessary to take into account, the Colombian beef sector does not have international certification on phytosanitary aspects while the United States and the European Union accomplish this requirement. This implies that Colombia cannot export beef while the latter countries can do it. All these changes would, in turn, affect internal beef demand. So on the whole, understanding beef demand is necessary for the Colombian agricultural policy.

Although the beef sector is important for the Colombian economy, little effort has been made to estimate demand elasticities and simulate different scenarios that impact on the sector. Therefore from an economic point of view, the objective of this study is to obtain reliable estimates of Colombian meat demand, and make some simulation exercises in order to evaluate the impact of different shocks on beef demand. Given that policy evaluations and simulations require reliable estimates of demand responsiveness to price and expenditure (Wahl, Hayes & Williams 1991), the methodology used to estimate elasticities is the Almost Ideal Demand System (AIDS), because

> "... gives an arbitrary first-order approximation to any demand system; it satisfies the axioms of choice exactly; it aggregates perfectly over consumers without invoking parallel linear Engel curves; it has a functional form which is consistent with known household-budget data; it is simple to estimate, largely avoiding the need for non-linear estimation; and it can be used to test the restriction of homogeneity and symmetry through linear restrictions on fixed parameters."
>
> *(Deaton & Muellbauer 1980*a*, pp 312)*

Specifically, we use a Multi-Stage AIDS model due to consumers following multiple steps when acquiring goods in the market. This approach allows us to estimate long-term elasticities in each stage, and also, total elasticities which incorporate interaction between levels. Additionally from an econometric perspective, it is well known that the level of uncertainty associated with elasticities estimates is very important; therefore, a simultaneous estimation procedure permits us to estimate a joint covariance matrix which can be used to calculate the standard deviation of the elasticities through the Delta Method. This is the methodological novelty of our paper. In particular, we use the Generalized Method of Moments to estimate the complete system.

Referring to short-term dynamics, we estimate an Error Correction version of the Multi-Stage Almost Ideal Demand System, and then, we simulate shocks at different levels of the decision making process of the consumers and measure their impacts. This strategy allows us to calculate, the short-term impact on beef demand associated with changes in the consumer's total expenditure and prices of beef, poultry and pork.

There is extensive empirical literature on the demand for meat. In most of this literature, the demand is estimated using the AIDS methodology (Asatryan 2003,

Clark 2006, Fuller 1997, Galvis 2000, Holt & Goodwin 2009, Sulgham & Zapata 2006). Even though there have been efforts in Colombia to determine beef demand elasticities (Caraballo 2003, Galvis 2000) most of the literature is focused on North America and Asia. Due undoubtedly to widely varying economic conditions across countries, the estimates of the elasticities of demand vary greatly. For example, the expenditure elasticity of beef consumption varies between 0.23 and 1.68. In the wealthier countries in the West, it is often below 1.0 (Barreira & Duarte 1997, Clark 2006, MAFF 2000, Sulgham & Zapata 2006), while in the poorer countries in the East it is generally above 1.0 (Liu, Parton, Zhou & Cox 2008, Chern, Ishibashi, Taniguchi & Tokoyama 2003, Ma, Huang, Rozelle & Rae 2003, Rastegari & Hwang 2007). The own-Marshallian price demand elasticity is between -1.19 and -0.10, usually less than -1 (Fousekis & Revell 2000, Galvis 2000, Golan, Perloff & Shen 2000). The compensated price elasticities show that changes in price does not affect the demand for beef as much.

In the specific case of Colombia, Galvis (2000) estimated the elasticities of demand for beef, poultry, and pork using the Seemingly Unrelated Regression (SUR) technique. He estimated an expenditure elasticity of demand for beef between 0.67 and 0.79, while the Marshallian (own price) elasticity is between -1.19 and -1.41. The cross-price elasticity of poultry prices on beef demand is between 0.27 and 0.96, and the cross-price elasticity of pork on beef demand is between 1.08 and 1.37. However, Galvis (2000) did not perform unit root tests, so the regressions might be spurious in the event that the variables are not cointegrated.

The empirical results in this article indicate that the long-term total and within meat group uncompensated price elasticities are -1.95 and -1.17, respectively. The total and within group compensated price elasticities are -1.78 and -0.52, and the total consumer expenditure elasticity of demand is 1.78. The results also indicate that consumers substitute beef for poultry, but not beef for pork. The short-term elasticities, calculated through Monte Carlo simulations, are smaller. They indicate that an increase of 1% in the price of beef decreases its demand by 0.24%, while increasing total expenditure by 1% has no significant impact on the demand for beef in Colombia.

The paper is organized as follows. Section 2 provides the methodology, Section 3 presents the long-term results, Section 4 presents some Monte Carlo simulation exercises, and Section 5 concludes.

2. Methodology

The methodology used in this paper is based on a Multi-Stage model which replicates the decision making process of the consumers when they buy beef (Gao, Eric, Gail & Cramer 1996, Michalek & Keyzer 1992, Shenggen, Wailes & Cramer 1995). Necessary and sufficient conditions for estimating a Multi-Stage budgeting process are that the direct utility function must be additively separable and the specific satisfaction functions in each stage should be homogeneous. Gorman (1957) provided conditions for this procedure to be optimal subject to the condition that must have more than two groups in each stage. Blackorby &

Russell (1997), extends Gorman's classic result to encompass the two-group cases that he did not take into account. These conditions are very restrictive, and must be in general considered implausible. However, Edgerton (1997) showed that a Multi-Stage budgeting process will lead to an approximately correct allocation if preferences are weakly separable and the group price indices being used do not vary too greatly with utility level. This means that a change in price of a commodity in one group affects the demand for all commodities in another group in the same manner. Also that the group price indices do not vary too greatly with expenditure level.

In particular, we estimate a Multi-stage Ideal Demand System of three levels to obtain the long-term elasticities in each level, and also, the total elasticities. The complete system is estimated using the Generalized Method of Moments. Following this strategy, the resulting three problems will be smaller and more tractable from an empirical point of view than the original problem, because including all goods prices in each of the equations is often faced with the problem of having too many variables (Segerson & Mount 1985). The long-term estimation is based on equation (1).

In order to simulate shocks in the short-term at different levels of the decision making process of consumers, we estimate the Error Correction version of the Multi-Stage AIDS model. This strategy allow us to calculate by Monte Carlo simulations, the short-term impact on beef demand associated with changes in the consumer's total expenditure and the prices of beef, poultry and pork. This estimation is based on equation (11).

This strategy considers the complex decision process through which an individual makes consumption decisions. Specifically, there are three levels: The upper one determines the aggregate level of food consumption; the middle one, conditioned by the upper one, determines the consumption of meat, and the lower level, conditioned by the other two, determines the beef, poultry, and pork demand.

In order to handle each stage budgeting process, an Almost Ideal Demand System is introduced (Deaton & Muellbauer 1980*a*). The mathematical specification of the AIDS model is the following,

$$w_{it} = \alpha_i + \sum_{j=1}^{N} \gamma_{ij} ln(p_{jt}) + \beta_i ln(X_t/P_t) + e_{it} \tag{1}$$

for $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$ where N is the number of goods, T is the temporal length, and the share in the total expenditure of the good i (w_{it}) is a function of the prices (p_{jt}), real expenditure(X_t/P_t) and an error (e_{it}). The general price index is usually represented by a nonlinear equation which is, in most cases, replaced by the Stone price index

$$ln(P_t^S) = \sum_{i=1}^{N} w_{it} ln(p_{it}) \tag{2}$$

However, the Stone index typically used in estimating Linear AIDS is not invariant to changes in units of measurement, which may seriously affect the approximation

properties of the model and can result in biased parameter estimates (Pashardes 1993, Moschini 1995). To overcome this problem other specifications for the price index can be used, such as the Paasche (3) or Laspeyres (4) index:

$$ln(P_t^P) = \sum_{i=1}^{N} w_{it} ln(p_{it}/p_i^0) \tag{3}$$

$$ln(P_t^L) = \sum_{i=1}^{N} w_i^0 ln(p_{it}) \tag{4}$$

where the superscript represents a base period.

It is worth noting the constraints (additivity, homogeneity and symmetry) that are imposed by the microeconomic theory:

$$\sum_{i=1}^{N} \alpha_i = 1, \quad \sum_{i=1}^{N} \gamma_{ij} = 0, \quad \sum_{i=1}^{N} \beta_i = 0 \tag{5}$$

$$\sum_{j=1}^{N} \gamma_{ij} = 0 \tag{6}$$

$$\gamma_{ij} = \gamma_{ji} \tag{7}$$

From the above specification the following long-term elasticities in each level can be calculated:

$$\eta_{it} = 1 + \beta_i / w_{it} \tag{8}$$

$$\epsilon_{ijt}^M = -I_A + \gamma_{ij}/w_{it} - \beta_i(w_{jt}/w_{it}) \tag{9}$$

$$\epsilon_{ijt}^H = -I_A + \gamma_{ij}/w_{it} + w_{jt} \tag{10}$$

where $I_A = 1$ if $i = j$.

Where $\eta_{it}, \epsilon_{ijt}^M$ and ϵ_{ijt}^H are expenditure, Marshallian (uncompensated) and Hicksian (compensated) elasticities, respectively.

It is required to investigate the time series properties of the data used in order to specify the most appropriate dynamic form of the model and to find out if the long-term demand relationships provided by equation (1) are economically meaningful or they are merely spurious. If all variables in equation (1) are cointegrated, the Error Correction Linear AIDS is given by the following form:

$$\Delta w_{it} = \sum_{j=1}^{N} \delta_{ij} \Delta w_{jt-1} + \sum_{j=1}^{N} \gamma_{ij} \Delta ln(p_{jt}) + \beta_i \Delta ln(X_t/P_t) + \lambda \hat{e}_{i,t-1} + \mu_{it}, \tag{11}$$

for $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, N$ y $t = 1, 2, \ldots, T$, where Δ refers to the difference operator, $\hat{e}_{i,t-1}$ represents the estimated residuals from the cointegrated equation (1), $-1 < \lambda < 0$ is the velocity of convergence, and μ_{it} is the error term. Intertemporal consistency requires that $\sum_{i=1}^{N} \delta_{ij} = 0$ (Anderson & Blundell 1983) and identification of the lagged budget shares requires $\sum_{j=1}^{N} \delta_{ij} = 0$ (Edgerton 1997).

Once the cointegrated equations are estimated, we can calculate the long-term total demand elasticities. Edgerton (1997) provide expressions to get elasticities associated with the lower level and we adapt these equations as follows:

$$\eta_{it}^{(T)} = \eta_{it} \times \eta_{Meat,t} \times \eta_{Food,t} \tag{12}$$

$$\epsilon_{ijt}^{M(T)} = \epsilon_{ijt}^{H} + w_{jt} \times \eta_{it} \times \epsilon_{Meat,t}^{H} + w_{jt} \times w_{Meat,t} \times \eta_{it} \times \eta_{Meat,t} \times \epsilon_{Food,t}^{M} \tag{13}$$

$$\epsilon_{ijt}^{H(T)} = \epsilon_{ijt}^{H} + w_{jt} \times \eta_{it} \times \epsilon_{Meat,t}^{H} + w_{jt} \times w_{Meat,t} \times \eta_{it} \times \eta_{Meat,t} \times \epsilon_{Food,t}^{H} \tag{14}$$

where superscript, i, j = beef, pork, poultry.

The total expenditure elasticity of beef demand, $\eta_{it}^{(T)}$, is a product of the expenditure elasticity of food, the food expenditure elasticity of meat and the meat expenditure elasticity of beef. The total price elasticities, $\epsilon_{ijt}^{M(T)}$ and $\epsilon_{ijt}^{H(T)}$, are the result of a direct effect within the meat group, but also of the reallocation effects of meat within food, and food within total consumption. Finally, we obtain standard deviations for the total elasticities with the Delta Method where this method establishes that given $Z = (Z_1, Z_2, \ldots, Z_k)$, a random vector with mean $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, if $g(Z)$ is a differentiable function, we can approximate its variance by

$$Var_\theta g(Z) \approx \sum_{i=1}^{k} (g_i'(\theta))^2 Var_\theta(Z_i) + 2 \sum_{i>j} g_i'(\theta) g_j'(\theta) Cov_\theta(Z_i, Z_j)$$

where $g_i'(\theta) = \frac{\partial}{\partial z_i} g(z)|_{z_1=\theta_1, z_2=\theta_2, \ldots, z_k=\theta_k}$.

Let $g(Z) = \eta_i^{(T)} = \eta_i \times \eta_{Meat} \times \eta_{Food}$, the total expenditure elasticity in the lower stage. We approximate its variance by

$$\begin{aligned}
Var_\theta \eta_i^{(T)} \approx & \left(\frac{1}{w_i}(\eta_{Meat}\eta_{Food})\right)^2 Var(\beta_i) \\
+ & \left(\frac{1}{w_{Meat}}(\eta_i\eta_{Food})\right)^2 Var(\beta_{Meat}) \\
+ & \left(\frac{1}{w_{Food}}(\eta_i\eta_{Meat})\right)^2 Var(\beta_{Food}) \\
+ & 2\left(\frac{1}{w_i w_{Meat}}(\eta_i\eta_{Meat})(\eta_{Food})^2\right) Cov(\beta_i, \beta_{Meat}) \\
+ & 2\left(\frac{1}{w_i w_{Food}}(\eta_i\eta_{Food})(\eta_{Meat})^2\right) Cov(\beta_i, \beta_{Food}) \\
+ & 2\left(\frac{1}{w_{Meat} w_{Food}}(\eta_{Meat}\eta_{Food})(\eta_i)^2\right) Cov(\beta_{Meat}, \beta_{Food})
\end{aligned}$$

where $\theta = (\beta_i, \beta_{Meat}, \beta_{Food})$, and i, j = beef, pork, poultry.

It must be observed that we need the covariance between the expenditure parameters at different stages. Therefore, we have to estimate the three levels simultaneously.

Now let $g(Z) = \epsilon_{ijt}^{M(T)} = \epsilon_{ijt}^{H} + w_{jt} \times \eta_{it} \times \epsilon_{Meat,t}^{H} + w_{jt} \times w_{Meat,t} \times \eta_{it} \times \eta_{Meat,t} \times \epsilon_{Food,t}^{M}$ i.e.,

$$\begin{aligned}
\epsilon_{ij}^{M(T)} &= (-I_A + \gamma_{ij}/w_i + w_j) \\
&+ w_j(1 + \beta_i/w_i)(-1 + \gamma_{Meat}/w_{Meat} + w_{Meat}) \\
&+ w_j w_{Meat}(1 + \beta_i/w_i)(1 + \beta_{Meat}/w_{Meat})(-1 + \gamma_{Food}/w_{Food} - \beta_{Food})
\end{aligned}$$

We can approximate the variance of the Marshallian total price demand elasticity by

$$\begin{aligned}
Var_\theta \epsilon_{ij}^{M(T)} &\approx \left(\frac{1}{w_i}\right)^2 Var(\gamma_{ij}) \\
&+ \left(\frac{w_j}{w_i}\epsilon_{Meat}^{H} + \frac{w_j w_{Meat}}{w_i}\eta_{Meat}\epsilon_{Food}^{M}\right)^2 Var(\beta_i) \\
&+ \left(\frac{w_j}{w_{Meat}}\eta_i\right)^2 Var(\gamma_{Meat}) \\
&+ \left(w_j \eta_i \epsilon_{Food}^{M}\right)^2 Var(\beta_{Meat}) \\
&+ \left(\frac{w_j w_{Meat}}{w_{Food}}\eta_i \eta_{Food}\right)^2 Var(\gamma_{Food}) \\
&+ \left(-w_j w_{Meat} \eta_i \eta_{Food}\right)^2 Var(\beta_{Food}) \\
&+ 2\left(\frac{w_j}{(w_i)^2}\epsilon_{Meat}^{H} + \frac{w_j w_{Meat}}{(w_i)^2}\eta_{Meat}\epsilon_{Food}^{M}\right) Cov(\gamma_{ij}, \beta_i) \\
&+ 2\left(\frac{w_j}{w_i w_{Meat}}\eta_i\right) Cov(\gamma_{ij}, \gamma_{Meat}) \\
&+ 2\left(\frac{w_j}{w_i}\eta_i \epsilon_{Food}^{M}\right) Cov(\gamma_{ij}, \beta_{Meat}) \\
&+ 2\left(\frac{w_j w_{Meat}}{w_i w_{Food}}\eta_i \eta_{Food}\right) Cov(\gamma_{ij}, \gamma_{Food}) \\
&+ 2\left(\frac{-w_j w_{Meat}}{w_i}\eta_i \eta_{Food}\right) Cov(\gamma_{ij}, \beta_{Food}) \\
&+ 2\left(\frac{w_j}{w_i}\epsilon_{Meat}^{H} + \frac{w_j w_{Meat}}{w_i}\eta_{Meat}\epsilon_{Food}^{M}\right)\left(\frac{w_j}{w_{Meat}}\eta_i\right) Cov(\beta_i, \gamma_{Meat}) \\
&+ 2\left(\frac{w_j}{w_i}\epsilon_{Meat}^{H} + \frac{w_j w_{Meat}}{w_i}\eta_{Meat}\epsilon_{Food}^{M}\right)\left(w_j \eta_i \epsilon_{Food}^{M}\right) Cov(\beta_i, \beta_{Meat}) \\
&+ 2\left(\frac{w_j}{w_i}\epsilon_{Meat}^{H} + \frac{w_j w_{Meat}}{w_i}\eta_{Meat}\epsilon_{Food}^{M}\right)\left(\frac{w_j w_{Meat}}{w_{Food}}\eta_i \eta_{Food}\right) Cov(\beta_i, \gamma_{Food})
\end{aligned}$$

$$
\begin{aligned}
+ \quad & 2\left(\frac{w_j}{w_i}\epsilon^H_{Meat} + \frac{w_j w_{Meat}}{w_i}\eta_{Meat}\epsilon^M_{Food}\right)(-w_j w_{Meat}\eta_i\eta_{Food})\, Cov(\beta_i, \beta_{Food}) \\
+ \quad & 2\left(\frac{w_j}{w_{Meat}}\eta_i\right)\left(w_j\eta_i\epsilon^M_{Food}\right) Cov(\gamma_{Meat}, \beta_{Meat}) \\
+ \quad & 2\left(\frac{w_j}{w_{Meat}}\eta_i\right)\left(\frac{w_j w_{Meat}}{w_{Food}}\eta_i\eta_{Food}\right) Cov(\gamma_{Meat}, \gamma_{Food}) \\
+ \quad & 2\left(\frac{w_j}{w_{Meat}}\eta_i\right)(-w_j w_{Meat}\eta_i\eta_{Food})\, Cov(\gamma_{Meat}, \beta_{Food}) \\
+ \quad & 2\left(w_j\eta_i\epsilon^M_{Food}\right)\left(\frac{w_j w_{Meat}}{w_{Food}}\eta_i\eta_{Food}\right) Cov(\beta_{Meat}, \gamma_{Food}) \\
+ \quad & 2\left(w_j\eta_i\epsilon^M_{Food}\right)(-w_j w_{Meat}\eta_i\eta_{Food})\, Cov(\beta_{Meat}, \beta_{Food}) \\
+ \quad & 2\left(\frac{w_j w_{Meat}}{w_{Food}}\eta_i\eta_{Food}\right)(-w_j w_{Meat}\eta_i\eta_{Food})\, Cov(\gamma_{Food}, \beta_{Food})
\end{aligned}
$$

where $\theta = (\gamma_{ij}, \beta_i, \gamma_{Meat}, \beta_{Meat}, \gamma_{Food}, \beta_{Food})$. Again, we ought to estimate the three levels simultaneously because we need the covariances between parameters at different stages.

Finally, let $g(Z) = \epsilon^{H(T)}_{ijt} = \epsilon^H_{ijt} + w_{jt} \times \eta_{it} \times \epsilon^H_{Meat,t} + w_{jt} \times w_{Meat,t} \times \eta_{it} \times \eta_{Meat,t} \times \epsilon^H_{Food,t}$, i.e.,

$$
\begin{aligned}
\epsilon^{H(T)}_{ij} &= (-I_A + \gamma_{ij}/w_i + w_j) \\
&+ w_j(1 + \beta_i/w_i)(-1 + \gamma_{Meat}/w_{Meati} + w_{Meat}) \\
&+ w_j w_{Meat}(1 + \beta_i/w_i)(1 + \beta_{Meat}/w_{Meat})(-1 + \gamma_{Food}/w_{Food} + w_{Food})
\end{aligned}
$$

We can approximate the variance of the Hicksian total price elasticity by

$$
\begin{aligned}
Var_\theta \epsilon^{H(T)}_{ij} \approx \quad & \left(\frac{1}{w_i}\right)^2 Var(\gamma_{ij}) \\
+ \quad & \left(\frac{w_j}{w_i}\epsilon^H_{Meat} + \frac{w_j w_{Meat}}{w_i}\eta_{Meat}\epsilon^H_{Food}\right)^2 Var(\beta_i) \\
+ \quad & \left(\frac{w_j}{w_{Meat}}\eta_i\right)^2 Var(\gamma_{Meat}) \\
+ \quad & \left(w_j\eta_i\epsilon^H_{Food}\right)^2 Var(\beta_{Meat}) \\
+ \quad & \left(\frac{w_j w_{Meat}}{w_{Food}}\eta_i\eta_{Food}\right)^2 Var(\gamma_{Food}) \\
+ \quad & 2\left(\frac{w_j}{(w_i)^2}\epsilon^H_{Meat} + \frac{w_j w_{Meat}}{(w_i)^2}\eta_{Meat}\epsilon^H_{Food}\right) Cov(\gamma_{ij}, \beta_i)
\end{aligned}
$$

$$
\begin{aligned}
&+ \quad 2\left(\frac{w_j}{w_i w_{Meat}}\eta_i\right) Cov(\gamma_{ij}, \gamma_{Meat}) \\
&+ \quad 2\left(\frac{w_j}{w_i}\eta_i \epsilon^H_{Food}\right) Cov(\gamma_{ij}, \beta_{Meat}) \\
&+ \quad 2\left(\frac{w_j w_{Meat}}{w_i w_{Food}}\eta_i \eta_{Food}\right) Cov(\gamma_{ij}, \gamma_{Food}) \\
&+ \quad 2\left(\frac{w_j}{w_i}\epsilon^H_{Meat} + \frac{w_j w_{Meat}}{w_i}\eta_{Meat}\epsilon^H_{Food}\right)\left(\frac{w_j}{w_{Meat}}\eta_i\right) Cov(\beta_i, \gamma_{Meat}) \\
&+ \quad 2\left(\frac{w_j}{w_i}\epsilon^H_{Meat} + \frac{w_j w_{Meat}}{w_i}\eta_{Meat}\epsilon^H_{Food}\right)\left(w_j \eta_i \epsilon^H_{Food}\right) Cov(\beta_i, \beta_{Meat}) \\
&+ \quad 2\left(\frac{w_j}{w_i}\epsilon^H_{Meat} + \frac{w_j w_{Meat}}{w_i}\eta_{Meat}\epsilon^H_{Food}\right)\left(\frac{w_j w_{Meat}}{w_{Food}}\eta_i \eta_{Food}\right) Cov(\beta_i, \gamma_{Food}) \\
&+ \quad 2\left(\frac{w_j}{w_{Meat}}\eta_i\right)\left(w_j \eta_i \epsilon^H_{Food}\right) Cov(\gamma_{Meat}, \beta_{Meat}) \\
&+ \quad 2\left(\frac{w_j}{w_{Meat}}\eta_i\right)\left(\frac{w_j w_{Meat}}{w_{Food}}\eta_i \eta_{Food}\right) Cov(\gamma_{Meat}, \gamma_{Food}) \\
&+ \quad 2\left(w_j \eta_i \epsilon^H_{Food}\right)\left(\frac{w_j w_{Meat}}{w_{Food}}\eta_i \eta_{Food}\right) Cov(\beta_{Meat}, \gamma_{Food})
\end{aligned}
$$

3. Results

The model is estimated using quarterly data for the period 1998-2007. The time series data for prices and per-capita consumption of beef, poultry and pork are taken from Federación Colombiana de Ganaderos (FEDEGAN). Data for per-capita expenditures are obtained from the Colombian National Accounts (DANE 2007). Prices are built from the implicit price indices formed as the ratio between nominal and real expenditures, i.e., Paasche indices.

We should use the True Cost of Living index, but Deaton & Muellbauer (1980*b*) considered Taylor's expansion of the cost function to show that a first order approximation to the True Cost of Living index will be the Paasche like index (see equation 3). An empirical evidence that supports this argument is that most price indices are highly correlated (Edgerton 1997).

Table (1) indicates that food expenditure is 25% of per-capita expenditure, of which expenditure on meat is 30%, and finally beef expenditure is 60% of the latter. Thus, beef consumption accounts for 4.5% of per-capita expenditure.

Historical data indicate that meat budget shares of the various types of meat have not changed. Between 1998 and 2007 average quarterly consumption of beef declined from 5.75 to 4.44 kg/capita, while poultry consumption rose from 2.92 to 5.49 kg/capita and pork consumption increased from 0.63 to 0.92 kg/capita. It seems likely that this shift in consumption has been caused by changes in the

Table 1: Descriptive Statistics: Colombian beef demand, 1998:I-2007:IV.

Variable	Mean	Standard Deviation	Jarque-Bera Test*
	Upper level		
$X_{TotalExpenditure}$	880,923	228,820	0.27
w_{Food}	0.25	0.0068	0.09
p_{Food}	114.42	20.78	0.35
p_{NoFood}	112.53	20.10	0.31
	Middle level		
$X_{FoodExpenditure}$	218,812	50,975	0.33
w_{Meat}	0.30	0.02	0.09
p_{Meat}	128.28	29.87	0.36
$p_{OtherFood}$	104.01	16.28	0.33
	Lower level		
w_{Beef}	0.60	0.27	0.67
p_{Beef}	8,598	2,672	0.15
w_{Pork}	0.08	0.01	0.13
p_{Pork}	8,007	1,802	0.29
$w_{Poultry}$	0.32	0.02	0.71
$p_{Poultry}$	5,299	726	0.15

* p-value

Source: Author's Estimations

relative prices of the different kinds of meat, as the data indicate that over the period, the price index of beef rose by 200%, while the price index of poultry increased by only 47% and the index of pork 110% (see Figures 1 and 2).

Unit root tests (Kwiatkowski, Phillips, Schmidt & Shin 1992, Ng & Perron 2001) were carried out, which indicate that all of the data series are I(1) (See Table 2). In order to account for endogeneity, the Johansen (1988) cointegration test was carried out at each budgeting allocation level based on equations (1).[1] As can be seen in Table 3, we cannot reject the null hypothesis of one cointegration vector in each equation. On the other hand, we use Hayes, Wahl & Williams (1990) statistical tests for testing weak separability on the second stage, i.e. meat decision. We use a Wald test under the null hypothesis of weak separability, and we cannot reject it, the p-value is 0.17.

We estimate simultaneously long-term system equations (1) for the three stages through Generalized Method of Moments.[2] In all stages, the Laspeyres index is used to build moment conditions, because of endogeneity caused due to the Stone index uses shares in its construction and it is not invariant to changes in units of

[1]Information criteria was used to select VEC order and deterministic components of the cointegration test.

[2]Residuals are normal and homoscedastic, but because of autocorrelation, we estimate the covariance matrix through consistent process (Newey & West 1987). Outcomes can be seen in Table 4.

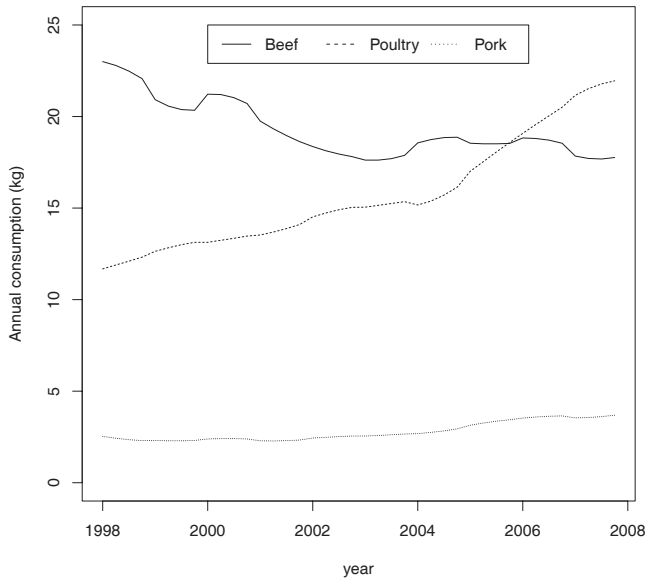


FIGURE 1: Meat per-capita annual consumption: Colombia, 1998:I-2007:IV.

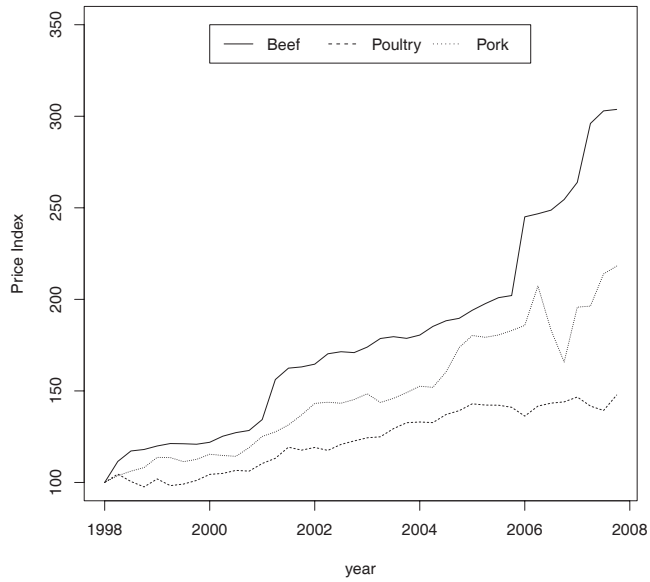


FIGURE 2: Meat's price index: Colombia, 1998:I-2007:IV.

measurement. We imposed homogeneity and symmetry conditions due to these conditions being important for demand theory, and not always being treated as verifiable conditions (Parikh 1988).

Table 2: Unit root tests: Colombian beef demand, 1998:I-2007:IV.

Variable	$KPSS^a$	Critical Value (5%)	$Ng-Perron^b$	Critical Value (5%)
		Upper level		
w_{Food}	0.625	0.463	-6.712	-8.100
Δw_{Food}	0.306	0.463	-14.062	-8.100
$Log(X/P)$	0.168	0.146	-2.835	-2.910
$\Delta Log(X/P)$	0.134	0.146	-2.116^c	-2.910
$Log(p_{Food}/P_{NoFood})$	0.192	0.146	-0.886	-2.910
$\Delta Log(p_{Food}/P_{NoFood})$	0.144	0.146	-2.919	-2.910
		Middle level		
w_{Meat}	0.482	0.463	-1.193	-8.100
Δw_{Meat}	0.288	0.463	-13.205	-8.100
$Log(X_{Food}/P_{Food})$	0.173	0.146	-0.363	-2.910
$\Delta Log(X_{Food}/P_{Food})$	0.100	0.146	-3.004	-2.910
$Log(p_{Meat}/p_{NoMeat})$	0.165	0.146	-2.619	-2.910
$\Delta Log(p_{Meat}/p_{NoMeat})$	0.075	0.146	-2.956	-2.910
		Lower level		
w_{Beef}	0.667	0.463	-3.629	-8.100
Δw_{Beef}	0.096	0.463	-18.851	-8.100
w_{Pork}	0.652	0.463	-2.552	-8.100
Δw_{Pork}	0.114	0.463	-13.998	-8.100
$Log(X_{Meat}/P_{Meat})$	0.185	0.146	-1.589	-2.910
$\Delta Log(X_{Meat}/P_{Meat})$	0.089	0.146	-2.713^c	-2.910
$Log(p_{Beef}/p_{Poultry})$	0.660	0.463	-1.760	-2.910
$\Delta Log(p_{Beef}/p_{Poultry})$	0.186	0.463	-3.079	-2.910
$Log(p_{Pork}/p_{Poultry})$	0.830	0.463	-3.566^c	-2.910
$\Delta Log(p_{Pork}/p_{Poultry})$	0.400	0.463	-4.116	-2.910

Notes: [a] Null hypothesis stationarity. [b] Null hypothesis unit root.
[c] We use the MZ_t^d statistic. However, the 5% critical value of the MPT^d statistic is 5.480 while its values are equal to 10.161, 6.205 and 3.754 for $\Delta Log(X/P)$, $\Delta Log(X_{Meat}/P_{Meat})$ and $Log(p_{Pork}/p_{Poultry})$, respectively. Additionally, the 5% critical value of the MSB^d statistic is 0.168 while its values are equal to 0.235, 0.182 and 0.136 for $\Delta Log(X/P)$, $\Delta Log(X_{Meat}/P_{Meat})$ and $Log(p_{Pork}/p_{Poultry})$, respectively.
Source: Author's Estimations

Long-term elasticities associated with each level are calculated using equations (8), (9) and (10). Equations (12), (13) and (14) are used to calculate total long-term elasticities. As can be seen in Table (5), beef, pork and poultry are luxuries, although this is not the result obtained for poultry if one only looked at within meat group elasticity. On the other hand, meat expenditure elasticity is 2.16, but its total expenditure elasticity is 1.65.[3] Although it is less than one, the food expenditure elasticity is still high at 0.76.

The partial beef expenditure elasticity is 1.07 in the Colombian economy (see Table 5). This value is smaller than the elasticity found in Mexico which is 1.30 (Golan et al. 2000). In general, the wealthier countries in the West have expenditure elasticities of beef below 1.0 (Clark 2006, Barreira & Duarte 1997, MAFF 2000, Sulgham & Zapata 2006), while the poorer countries in the East have elas-

[3]This is calculated as 2.16 (within expenditure elasticity) × 0.76 (food expenditure elasticity).

Table 3: Cointegration tests: Colombian beef demand, 1998:I-2007:IV.

Equation	Ho: CE(s)	Max. Eigenvalue[a]	Critical Value (5%)	Trace[b]	Critical Value (5%)
		Upper Level			
Food Demand[c]	r=0*	43.72	24.25	57.76	35.01
	r=1	9.85	17.14	14.03	18.39
	r=2*	4.18	3.84	4.18	3.84
		Middle Level			
Meat Demand[c]	r=0*	33.87	24.25	51.82	35.01
	r=1	11.98	17.14	17.94	18.39
	r=2*	5.95	3.84	5.95	3.84
		Lower Level			
Beef Demand[d]	r=0*	36.24	24.15	57.09	40.17
	r=1	13.08	17.79	20.84	24.27
	r=2	5.36	11.22	7.75	12.32
	r=3	2.39	4.12	2.39	4.12
Pork Demand[d]	r=0*	32.51	24.15	51.14	40.17
	r=1	11.86	17.79	18.62	24.27
	r=2	6.20	11.22	6.76	12.32
	r=3	0.56	4.12	0.56	4.12

[a] Null hypothesis: the number of cointegrating vectors is r against the alternative of $r+1$
[b] Null hypothesis: the number of cointegrating vectors is less than or equal to r against general alternative
[c] There is a constant and a deterministic trend in the cointegrated equations. Schwarz criterion supports these outcomes.
[d] There is not a constant nor a deterministic trend in the cointegrated equations. Schwarz criterion supports these outcomes.
* Denotes rejection of the hypothesis at the 5% level
Source: Author's estimations.

ticities above 1.0 (Chern et al. 2003, Liu et al. 2008, Ma et al. 2003, Rastegari & Hwang 2007).

Table 4: Residuals tests: Colombian beef demand, 1998:I-2007:IV.

Equation	Jarque-Bera[a]	Breusch-Pagan-Godfrey[b]	Breusch-Godfrey[c]
	Upper Level		
Food Demand	1.61	3.31	36.21*
	Middle Level		
Meat Demand	2.34	8.70	27.82*
	Lower Level		
Beef Demand	1.24	6.66	24.53*
Pork Demand	2.75	5.72	30.26*

[a] The null hypothesis is normality
[b] The null hypothesis is homocedasticity
[c] The null hypothesis is not autocorrelation
* Denotes rejection of the hypothesis at the 5% level
Source: Author's estimations.

As can be seen in Table (6), there is substitution of poultry for beef within the meat group, but this effect is not present if taking into account that a change

Table 5: Expenditure elasticities for the three levels: Colombian beef demand, 1998:I-2007:IV.

Upper level		
Food	Other goods	
0.76*	1.07*	
(0.034)	(0.011)	
Middle level		
Meat	Other food	
2.16*	0.48*	
(0.291)	(0.129)	
Lower level		
Within meat group		
Beef	Pork	Poultry
1.07*	1.78*	0.64*
(0.145)	(0.367)	(0.268)
Total		
Beef	Pork	Poultry
1.78*	2.95*	1.05*
(0.378)	(0.687)	(0.166)

Standard deviation are calculated with Delta method.
* Significant at 5%
Source: Author's estimations

of poultry price implies reallocation effects of meat within food and food within total consumption. With regard to total uncompensated and compensated own-price elasticities, we can see that beef is quite elastic, and the differences between within meat group and total elasticities are large. This fact can be misleading if the within elasticities are used for making policy judgements.[4]

The partial own-Marshallian price demand elasticity is -1.17 in Colombia (see Table 6). This value is similar to elasticities that are internationally found (Galvis 2000, Golan et al. 2000, Fousekis & Revell 2000). Usually, this elasticity is less than -1. With regard to the partial compensated price elasticity, it is found a value equal to -0.52 in the Colombian economy (see Table 6). The partial own-Hicksian price demand elasticity is -0.59 in Mexico (Golan et al. 2000). This elasticity internationally has a range between -0.23 and -1.63. The highest elasticity in absolute value is found in Nigeria (Osho & Nazemzadeh 2005), while the lowest is found in U.S. (Asatryan 2003).

[4]Uncompensated own-price elasticities of poultry and pork are -1.020 and -0.028, respectively.

Table 6: Uncompensated and compensated beef price elasticities: Colombian beef demand, 1998:I-2007:IV.

	Marshallian			Hicksian		
	Beef	Pork	Poultry	Beef	Pork	Poultry
Within	-1.17*	-0.04	0.14*	-0.52*	0.04*	0.47*
	(0.142)	(0.033)	(0.043)	(0.063)	(0.001)	(0.003)
Total	-1.95*	-0.09	-0.03	-1.78*	-0.08	8E-03
	(0.278)	(0.047)	(0.111)	(0.262)	(0.045)	(0.103)

Standard deviation are calculated with Delta method.
* Significant at 5%
Source: Author's estimations

Table 7: Short-term beef elasticities: Colombian beef demand, 1998:I-2007:IV.

Beef demand			
Total expenditure	Beef price	Pork price	Poultry price
0.034	-0.247	-0.025	0.103

Source: Author's estimations

4. Simulations

In order to calculate short-term elasticities, Seemingly Unrelated Regression Equations are used for estimating an Error Correction Linear AIDS with the three stages simultaneously. Monte Carlo simulation exercises are done based on the estimated model in order to analyse the short-term dynamics of beef demand. The algorithm used solves the model for each observation in the solution sample, using a recursive procedure to compute values for the endogenous variables. The model is solved repeatedly for different draws of the stochastic components (coefficients and errors). During each repetition, errors are generated for each observation in accordance with the residual uncertainty in the model. The three stages are linked by prices and expenditures; for example, a shock on consumption expenditure causes a direct effect on food demand, which implies an expenditure effect on meat demand, and as consequence a reallocation within the group. On the other hand, a change of beef price implies a direct effect within the meat group, but also affects meat within food and food within consumption.

The simulation results suggest a good fit for each equation in the model; during the period analysed observed data fell inside the 95% prediction interval (outcomes upon author's request).

We analyse transitory effects associated with a positive shock on total expenditure, and increases in beef, poultry and pork prices. We use our simulated model to measure the impact on beef demand by comparing in-sample forecasted beef demand with and without the shocks for the first quarter of 2007. Given that a comparison is being performed, the same set of random residuals is applied to both scenarios during each repetition. This is done so that the deviation between the different scenarios is based only on differences in the exogenous variables, not on differences in random errors.

The first exercise evaluates the short-term effect on beef demand associated with a positive shock on total expenditure. Specifically, we increase the consumer expenditure by 1%, and compare this scenario with the baseline scenario (without shock). We find that there is an increase in beef demand by only 0.034%. On the other hand, we evaluate the short-term effects in beef demand associated with transitory increases in beef, pork and poultry prices. It can be seen on Table 7, that an increase of 1% in beef price reduces its own demand by 0.24%. Finally, there is a substitution effect of poultry for beef, because an increases of 1% on poultry price causes an increase in beef demand by 0.1%, while an increase in pork price causes very little effect on beef demand.

5. Conclusions

The results in the long-term indicate that the expenditure elasticity of food is less than one, supporting the idea of a normal good. On the other hand, meat is a luxury good because its expenditure elasticity is greater than one. In the lower level, the cross price elasticities indicate that there is a bigger substitution effect of beef for poultry than beef for pork. Although the total expenditure elasticity of demand for beef is 1.78 in the long-term, the short-term expenditure elasticity is merely 0.034. The smaller short-term reaction of the consumers is also evidenced in price shocks; while the own price elasticity of beef is -0.24 in the short-term, the long-term total elasticity is -1.95. These differences between elasticities obey the small velocities of convergence in the three levels of the model. Specifically, the velocities of convergence are 2%, 10% and 17% on the beef, meat and food demand equations.

Colombian real per-capita total expenditure has grown at 2.1% per annum from 2000 to 2007; therefore, given a 1.5% population growth rate per annum, the total expenditure beef elasticity implies beef demand growing at 5.3% a year.[5] However, Colombian beef production has grown at -0.51% per annum in the same period, this difference has caused Colombian beef price to increase by 14.7% per annum. Recently, Colombia has been negotiating international trade agreements with the United States and the European Union. This implies that the Colombian beef sector would have international competition from countries with high subsidies, and as a consequence, the internal beef price would decrease. These facts would have important effects on domestic producers, which ought to improve productivity in order to stay as an important sector in the Colombian economy and make good use of the new market opportunities.

[Recibido: abril de 2012 — Aceptado: abril de 2013]

[5]This is calculated as 1.5% (population growth rate per annum) + 2.1% (per-capita total expenditure growth per annum) * 1.78 (total expenditure elasticity).

References

Anderson, G. & Blundell, R. (1983), 'Estimation and hypothesis testing in dynamic singular equations systems', *Econometrica* **50**, 1559–1571.

Asatryan, A. (2003), Data Mining of Market Information to Assess at-Home Pork Demand, PhD thesis, Texas AM University.

Barreira, M. & Duarte, F. (1997), An analysis of changes in Portuguese meat consumption., *in* B. W. et al., ed., 'Agricultural Marketing and Consumer Behaviour in a Changing World', Kluwer Academic Publishers, pp. 261–273.

Blackorby, C. & Russell, R. (1997), 'Two-stage budgeting: An extension of Gorman's theorem', *Economic Theory* **9**, 185–193.

Caraballo, L. J. (2003), '¿Cómo estimar una funcion de demanda? Caso: Demanda de carne de res en Colombia', *Geoenseñanza* **8**(2), 95–104.

Chern, W., Ishibashi, K., Taniguchi, K. & Tokoyama, Y. (2003), Analysis of the Food Consumption of Japanese Households, Technical Report 152, FAO Economic and Social Development.

Clark, G. (2006), Mexican Meat Demand Analysis: A Post-NAFTA Demand Systems Approach, Master's thesis, Texas Tech University.

DANE (2007), 'Cuentas nacionales', `http://www.dane.gov.co/daneweb_V09/index.php?option=com_content&view=article&id=128&Itemid=85`. [Online; accessed March-2010].

Deaton, A. & Muellbauer, J. (1980*a*), 'An almost ideal demand system', *The American Economic Review* **70**(3), 312–326.

Deaton, A. & Muellbauer, J. (1980*b*), *Economics and Consumer Behaviour*, Cambridge: Cambridge University Press.

Edgerton, D. (1997), 'Weak separability and the estimation of elasticities in multistage demand systems', *American Journal of Agricultural Economics* **79**(1), 62–79.

Fousekis, P. & Revell, B. (2000), 'Meat demand in the UK: A differential approach', *Journal of Agriculture and Applied Economics* **32**(1), 11–19.

Fuller, F. (1997), Policy and Projection Model for the Meat Sector in the People's Republic of China, Technical report 97-tr 36, Center for Agricultural and Rural Development - Iowa State University.

Galvis, L. A. (2000), 'La demanda de carnes en Colombia: un análisis econométrico', *Documentos de Trabajo sobre Economía Regional* **13**. Centro de Estudios Económicos Regionales, Banco de la República.

Gao, X., Eric, J., Gail, W. & Cramer, L. (1996), 'A two-stage rural household demand analysis: Microdata evidence from Jiangsu Province, China', *American Journal of Agricultural Economics* **78**(3), 604–613.

Golan, A., Perloff, J. & Shen, E. (2000), 'Estimating a demand system with non-negativity constraints: Mexican meat demand', *The Review of Economics and Statistics* **83**, 541–550.

Gorman, W. (1957), 'Separable utility and aggregation', *Econometrica* **27**(3), 469–481.

Hayes, D., Wahl, T. & Williams, G. (1990), 'Testing restrictions on a model of Japanese meat demand', *American Journal of Agricultural Economics* **72**(3), 556–566.

Holt, M. & Goodwin, B. (2009), The almost ideal and translog demand systems, Technical Report 15092, Munich Personal RePEc Archive - MPRA. Online at http://mpra.ub.uni-muenchen.de/15092/.

Johansen, S. (1988), 'Statistical analysis of cointegration vectors', *Journal of Economic Dynamics and Control* **12**(2-3), 231–254.

Kwiatkowski, D., Phillips, P., Schmidt, P. & Shin, Y. (1992), 'Testing the null hypothesis of stationarity against the alternative of a unit root', *Journal of Econometrics* **54**, 159–178.

Liu, H., Parton, K., Zhou, Z. & Cox, R. (2008), 'Meat consumption in the home in China: An empirical study', *Australian Journal of Agricultural and Resource Economics* .
*Online: aede.osu.edu/programs/anderson/trade/57HongboLiu.pdf

Ma, H., Huang, J., Rozelle, S. & Rae, A. (2003), Livestock Product Consumption Patterns in Urban and Rural China, Working paper, Research in Agricultural Applied Economics.

MAFF (2000), National Food Survey 1999, Annual report, Ministry of Agriculture, Forestry and Fisheries of United Kingdom.

Michalek, J. & Keyzer, M. (1992), 'Estimation of a two-stage LES-AIDS consumer demand system for eight EC countries', *European Review of Agricultural Economics* **19**(2), 137–163.

Moschini, G. (1995), 'Units of measurement and the stone index in demand system estimation', *American Journal of Agriculture Economics* **77**, 63–68.

Newey, W. & West, K. (1987), 'A simple positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**, 703–708.

Ng, S. & Perron, P. (2001), 'Lag length selection and the construction of unit root tests with good size and power', *Econometrica* **69**(6), 1519–1554.

Osho, G. & Nazemzadeh, A. (2005), 'Consumerism: Statistical estimation of Nigeria meat demand', *Journal of International Business Research* **4**(1), 69–79.

Parikh, A. (1988), 'An econometric study on estimation of trade shares using the almost ideal demand system in the world link', *Applied Economics* **20**, 1017–1079.

Pashardes, P. (1993), 'Bias in estimating the almost ideal demand system with the stone index approximation', *The Economic Journal* **103**(419), 908–915.

Rastegari, S. & Hwang, S. (2007), 'Meat demand in South Korea: An application of the restricted source-differentiated almost ideal demand system model', *Journal of Agriculture and Applied Economics* **39**(1), 47–60.

Segerson, K. & Mount, D. (1985), 'A non-homothetic two-stage decision model using AIDS', *The Review of Economics and Statistics* **67**(4), 630–639.

Shenggen, F., Wailes, E. & Cramer, G. (1995), 'Household demand in rural China: A two-stage LES-AIDS model', *American Journal of Agricultural Economics.* **77**(1), 54–62.

Sulgham, A. & Zapata, H. (2006), A dynamic approach to estimate theoretically consistent US meat demand system, Research paper, Annual Meeting of American Agricultural Economics Association.

Wahl, T., Hayes, D. & Williams, G. (1991), 'Dynamic adjustment in the Japanese livestock industry under beef import liberalization', *American Agricultural Economics Association* **73**(1), 118–132.

Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 43 a 57

The Family of Log-Skew-Normal Alpha-Power Distributions using Precipitation Data

La familia de distribuciones alfa-potencia log-skew-normal usando datos de precipitación

Guillermo Martínez-Flórez[1,a], Sandra Vergara-Cardozo[2,b], Luz Mery González[2,c]

[1]Departamento de Matemáticas y Estadística, Universidad de Córdoba, Montería, Colombia

[2]Departamento de Estadística, Facultad de Ciencias, Univesidad Nacional de Colombia, Bogotá D.C, Colombia

Abstract

We present a new set of distributions for positive data based on a skew-normal alpha-power (PSN) model including a new parameter which in turn makes the log-skew-normal alpha-power (LPSN) model more flexible than both the log-normal (LN) model and log-skew-normal (LSN) model. The LPSN model contains the LN model and LSN model as special cases. Furthermore, it models positive data with asymmetry and kurtosis larger than the one permitted by the LN distribution. Precipitation data illustrates the usefulness of the LPSN model being less influenced by outliers.

***Key words*:** Asymmetry, Fisher information matrix, Kurtosis, Likelihood ratio test, Maximum likelihood estimator.

Resumen

Presentamos una nueva familia de distribuciones para datos positivos basada en el modelo skew-normal alpha-power (PSN), incluyendo un nuevo parámetro el cual hace el modelo log-skew-normal alpha-power (LPSN) más flexible que los modelos log-normal (LN) y log-skew-normal (LSN). El modelo LPSN contiene el modelo LN y el modelo LSN como casos particulares. Además, modela datos positivos con asimetría y curtosis más allá de lo permitido por la distribución LN. Datos de precipitación ilustran la utilidad del modelo LPSN siendo menos influenciado por outliers.

***Palabras clave*:** asimetría, curtosis, estimador máxima verosimilitud, matriz de información de Fisher, test de razón de verosimilitud.

[a]Professor. E-mail: gmartinez@correo.unicordoba.edu.co
[b]Assistant professor. E-mail: svergarac@unal.edu.co
[c]Assistant professor. E-mail: lgonzalezg@unal.edu.co

1. Introduction

The log-normal (LN) distribution obtained as a transformation of the normal distribution has been widely used to model different types of information including income in economics and material lifetimes. In of different fields of knowledge, asymmetry and kurtosis of the data are outside of the range allowed by the LN distribution so it is necessary to use another distribution that can take into account these issues. In the same way that Azzalini (1985), we introduce the skew-normal (SN) distribution to conform data with a range of asymmetry and kurtosis outside the range allowed by the normal distribution, Lin & Stoyanov (2009) present the log-skew-normal (LSN) distribution which is an extension for positive data of the LN distribution in order to conform data with asymmetry and kurtosis outside the range allowed by the LN distribution. The probability density function of this model is given by

$$\begin{aligned}\varphi_{LSN}(y;\xi,\eta,\lambda) &= \frac{2}{\eta y}\phi\left(\frac{\log(y)-\xi}{\eta}\right)\left\{\Phi\left(\lambda\frac{\log(y)-\xi}{\eta}\right)\right\} \\ &= \frac{1}{y}\phi_{SN}\left(\log(y);\xi,\eta,\lambda\right),\ y\in\mathbb{R}^{+}\end{aligned} \tag{1}$$

where

$$\phi_{SN}(x;\xi,\eta,\lambda) = \frac{2}{\eta}\phi\left(\frac{x-\xi}{\eta}\right)\left\{\Phi\left(\lambda\frac{x-\xi}{\eta}\right)\right\}$$

denotes the density function of the SN distribution with parameters of location (ξ), scale (η), and shape (λ). The LSN model $[Y \sim LSN(\xi,\eta,\lambda)]$ given by (1) contains the parameters of location (ξ), scale (η), and shape (λ) that control the asymmetry of the data. $\phi(.)$ and $\Phi(.)$ denote the density and cumulative distribution function of standard normal distribution, N(0,1). Based on the SN of Azzalini (1985) and generalized Gaussian (PN) of Durrans (1992), Martínez-Flórez (2011) introduce and studies the main features of the asymmetric distribution called skew-normal alpha-power (PSN) distribution with probability density function given by

$$\phi_{PSN}(z;\lambda,\alpha) = \alpha\phi_{SN}(z;\lambda)\left\{\Phi_{SN}(z;\lambda)\right\}^{\alpha-1} \tag{2}$$

where $z,\lambda\in\mathbb{R}$, $\alpha\in\mathbb{R}^{+}$, $\phi_{SN}(z;\lambda)=\phi_{SN}(z;0,1,\lambda)$ as defined in (1) and $\Phi_{SN}(z;\lambda)$ in (3). The PSN model $[X \sim PSN(\lambda,\alpha)]$ given by (2) considers parameters of shape λ and α with

$$\Phi_{SN}(z;\lambda) = \int_{-\infty}^{z}\phi_{SN}(t;\lambda)dt = \Phi(z) - 2T(z,\lambda) \tag{3}$$

being the cumulative distribution function of skew-normal distribution, Azzalini (1985), and $T(.,\lambda)$ the Owen's (1956) function.

In (2), $\lambda = 0$ and $\alpha = 1$ corresponds to the standard normal case, *i.e.*, $\phi_{SN}(.;0,1,0) = \phi_{PSN}(z;0,1) = \phi_{SN}(.;0) = \phi(.)$ and $\Phi_{SN}(.;0) = \Phi(.)$. The model is an extension of the PN model, Durrans (1992) and the Gupta & Gupta (2008) exponential model

$$\varphi_\alpha(z;\alpha) = \alpha\phi(z)\{\Phi(z)\}^{\alpha-1}, \quad z \in \mathbb{R} \tag{4}$$

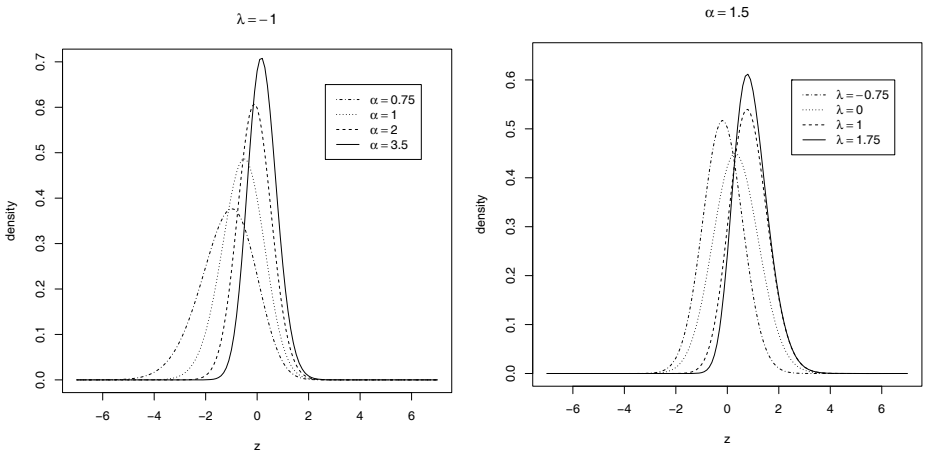
replacing the normal density by the skew-normal density.

Martínez-Flórez (2011) demonstrate that the expected information matrix of the PSN model is nonsingular in the neighborhood of the skewness parameters $\lambda = 0$ and $\alpha = 1$ contrary to the case of Azzalini (1985) whose expected information matrix is singular in the neighborhood of $\lambda = 0$. Table 1 shows the intervals of asymmetry and kurtosis coefficients for the PSN, SN, and PN models. The PSN model has greater asymmetry and the distribution is more platikurtic or leptokurtic than the Azzalini (1985) and Durrans (1992) models. This shown an advantage of the model (2) over the $\phi_{SN}(z;\lambda)$ and $\varphi_\alpha(z;\alpha)$ models.

TABLE 1: Intervals of asymmetry ($\sqrt{\beta_1}$) and kurtosis (β_2) coefficients, defined in (7), for the PSN, SN, and PN models given by Martínez-Flórez, G. (2011).

Model	$\sqrt{\beta_1}$	β_2
Skew-normal alpha-power (PSN) model	[-1.4676 ; 0.9953)	[1.4672 ; 5.4386]
Skew-normal (SN) model	(-0.9953 ; 0.9953)	[3 ; 3.8692)
Generalized gaussian (PN) model	[-0.6115 ; 0.9007]	[1.7170 ; 4.3556]

Figures 1(a) show corresponding and 1(b), the parameters λ and α of asymmetry and kurtosis of the PSN distribution a more flexible model than Azzalini (1985) and Durrans (1992) yielding.



(a) (b)

FIGURE 1: Probability density function of the skew-normal alpha-power distribution.

Other work on this type of distribution was studied by Arnold & Beaver (2002) and Gupta & Gupta (2004). We present a new set of distributions based on the PSN distribution that corresponds to the log-skew-normal alpha-power (LPSN) distribution.

In Section 2, we describe the LPSN distribution, its observed information matrix and the expected information matrix. We also perform an application of the proposed model to data by IDEAM (2006) in which the coefficients of skewness and kurtosis of the model justify the use of the LPSN model. We conclude with a brief discussion in Section 3.

2. Log-Skew-Normal Alpha-Power (LPSN) Model

The LPSN distribution is a new alternative for family distribution of positive data with a range of asymmetry and/or kurtosis outside of the range permitted by the LN and LSN distributions.

Definition 1. The positive random variable Y in the $\mathbb{R}^+$ has a univariate log-skew-normal alpha-power distribution with parameters λ and α if the transformed variable $Z = \log(Y)$ has a PSN distribution with parameters λ and α. This is denoted by $Y \sim LPSN(\lambda, \alpha)$. The probability density function of a random variable Y with distribution $LPSN(\lambda, \alpha)$ is given by

$$\varphi_{LPSN}(y; \lambda, \alpha) = \frac{\alpha}{y}\phi_{SN}(\log(y); \lambda)\left\{\Phi_{SN}(log(y); \lambda)\right\}^{\alpha-1},\ y,\ \alpha \in \mathbb{R}^+ \text{ and } \lambda \in \mathbb{R}$$

The cumulative distribution function of the LPSN model is given by

$$\mathcal{F}_Y(y; \lambda, \alpha) = \{\Phi_{SN}(log(y); \lambda)\}^{\alpha},\ y \in \mathbb{R}^+ \tag{5}$$

According to equation (5), the inversion method can be used to generate a random variable with distribution $LPSN(\lambda, \alpha)$. Thus, if U is a uniform random variable in (0,1) the random variable $Y = exp\{\Phi_{ISN}(U^{1/\alpha}; \lambda)\}$ has LPSN distribution of the parameters λ and α where Φ_{ISN} represents the inverse function of the SN distribution, $\Phi_{SN}(.; \lambda)$, whose values can be obtained in many statistical packages (R Development Core Team 2011).

When $\alpha = 1$, the LPSN distribution is identical to the LSN distribution $[\varphi_{LPSN}(y; \lambda, 1) = \varphi_{LSN}(y; 0, 1, \lambda)]$ and when $\lambda = 0$ and $\alpha = 1$, the LPSN distribution is identical to the log-normal (LN) distribution. So LPSN distribution is more flexible than LN and LSN distributions (see, for example, Figures 2(a) and 2(b)).

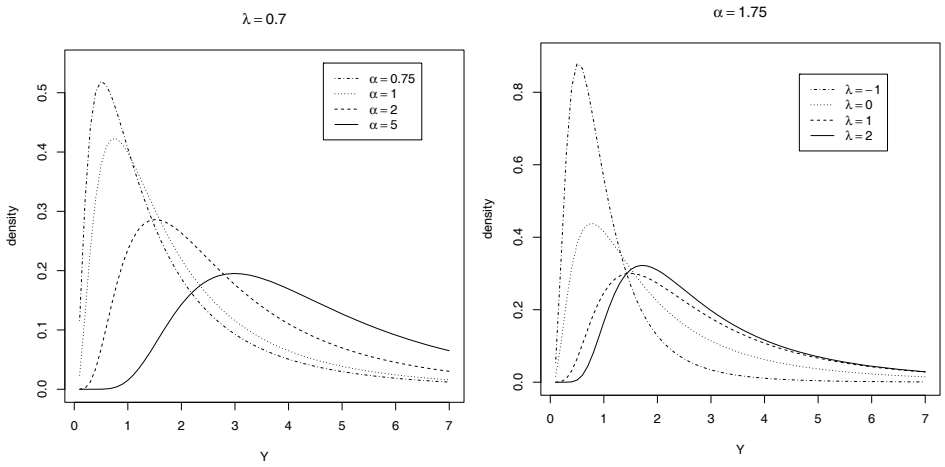
2.1. Moments of the Distribution

The *r-th* moment of the random variable Y with LPSN distribution can be written as,

$$\mu_r = \mathbb{E}(Y^r) = \alpha \int_0^1 \left\{\exp\left[r\Phi_{ISN}(y; \lambda)\right]\right\} y^{\alpha-1} dy \tag{6}$$

Let $\mu'_r = \mathbb{E}(Y - \mathbb{E}(Y))^r$, $r = 2, 3, 4$,

$$\mu'_2 = \mu_2 - \mu_1^2,\quad \mu'_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 \text{ and } \mu'_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4$$



(a) (b)

FIGURE 2: Probability density function of the log-skew-normal alpha-power distribution.

The variance, coefficient of variation, skewness and kurtosis are given by:

$$\sigma_Y^2 = Var(Y) = \mu_2', \quad CV = \frac{\sqrt{\sigma_Y^2}}{\mu_1}, \quad \sqrt{\beta_1} = \frac{\mu_3'}{[\mu_2']^{3/2}} \quad \text{and} \quad \beta_2 = \frac{\mu_4'}{[\mu_2']^2} \tag{7}$$

2.2. Scale-Location

Let $PSN(\xi, \eta, \lambda, \alpha)$ denotes a location-scale transformation of $PSN(\lambda, \alpha)$ where $\xi \in \mathbb{R}$, $\eta \in \mathbb{R}^+$ and $Y = \xi + \eta Z$.

Definition 2. If X has a distribution of localization-scale parameters $PSN(\xi, \eta, \lambda, \alpha)$ then the extension of scale-location to the LPSN distribution follows the transformation $X = \log(Y)$, where $\xi \in \mathbb{R}$ and $\eta \in \mathbb{R}^+$. Then, the density of Y is given by

$$\varphi_{LPSN}(y; \xi, \eta, \lambda, \alpha) = \alpha \varphi_{LSN}(y; \xi, \eta, \lambda) \left\{ \Phi_{SN}\left(\frac{\log(y) - \xi}{\eta}; \lambda \right) \right\}^{\alpha - 1} \tag{8}$$

$$y, \ \alpha \in \mathbb{R}^+, \text{ and } \lambda \in \mathbb{R}$$

where $\varphi_{LSN}(y; \xi, \eta, \lambda)$ is defined in (1) and $\Phi_{SN}(.; \lambda)$, in (3)

We use the notation $Y \sim LPSN(\xi, \eta, \lambda, \alpha)$. So $LPSN(\lambda, \alpha) = LPSN(0, 1, \lambda, \alpha)$.

A special case in the model (8) is when $\lambda = 0$, obtaining the density,

$$\varphi_{LPSN}(y; \xi, \eta, 0, \alpha) = \frac{\alpha}{\eta y} \phi\left(\frac{\log(y) - \xi}{\eta} \right) \left\{ \Phi\left(\frac{\log(y) - \xi}{\eta} \right) \right\}^{\alpha - 1}, \ y \in \mathbb{R}^+$$

This is denoted $Y \sim LPSN_{\lambda=0}(\xi, \eta, \alpha)$. Like the model LSN model, this distribution is also a generalization of the LN model which we will call the generalized LN distribution.

The following result is an extension of the LN and LSN distributions.

Theorem 1. *For any $\lambda \in \mathbb{R}$ and $\alpha \in \mathbb{R}^+$, the random variable $Y \sim LPSN(\xi, \eta, \lambda, \alpha)$ does not have a moment generating function (MGF).*

Proof. As $\lambda = 0$ and $\alpha = 1$ in the LPSN model, we have the case of the LN distribution, which does not have a moment generating function. Since MGF satisfies the property,

$$M_{aY+b}(t) = \exp(bt) M_Y(at)$$

then it is sufficient to consider the standard case $LPSN(\lambda, \alpha)$. □

For fixed values $\alpha = \alpha_0 > 0$ and $\lambda = \lambda_0$, the MGF of Y can be written as

$$\begin{aligned} M_Y(t) &= \mathbb{E}(e^{ty}) \\ &= \int_0^\infty e^{ty} \varphi_{LPSN}(y; \lambda_0, \alpha_0) dy \\ &= \int_0^\infty \frac{\alpha_0}{y} e^{ty} \phi_{SN}(log(y); \lambda_0) \left\{\Phi_{SN}(log(y); \lambda_0)\right\}^{\alpha_0 - 1} dy \\ &= \int_0^\infty h(y, t, \lambda_0, \alpha_0) g(y, \lambda_0, \alpha_0) dy, \qquad y \in \mathbb{R}^+ \end{aligned}$$

with

$$h(y, t, \lambda_0, \alpha_0) = \frac{2\alpha_0}{y} e^{ty} \phi(\log(y)) \{\Phi(\lambda_0 \log(y))\} > 0$$

and

$$g(y, \lambda_0, \alpha_0) = \left\{\Phi_{SN}(log(y); \lambda_0)\right\}^{\alpha_0 - 1}$$

to all $y > 0$.

When $t > 0$ is fixed, we prove that

$$J_{(\lambda_0, \alpha_0)} = \int_0^\infty h(y, t, \lambda_0, \alpha_0) g(y, \lambda_0, \alpha_0) dy = \infty$$

for all $\lambda_0 \in \mathbb{R}$ and $\alpha_0 \in \mathbb{R}^+$.

If $\lambda_0 > 0$ according to Lin & Stoyanov (2009)

$$\liminf_{y \to \infty} \{\Phi(\lambda_0 \log(y))\} \geq \frac{1}{2}$$

therefore $h(y, t, \lambda_0, \alpha_0) \to \infty$ when $y \to \infty$. Now, $g(y, \lambda_0, \alpha_0) \to 1$ when $y \to \infty$, then we conclude that $J_{(\lambda_0, \alpha_0)} \to \infty$ when $y \to \infty$.

According to Lin & Stoyanov (2009), if $\lambda_0 < 0$ then

$$\lim_{y \to \infty} \frac{-\log\left(\Phi(-y)\right)}{y^2} = \frac{1}{2}$$

Therefore, when $y \to \infty$, we have the asymptotic approximation,

$$\log\left(\Phi(\lambda_0 \log(y))\right) \approx \frac{1}{2}\left(\lambda_0 \log(y)\right)^2$$

Then, we assume that $log(\alpha_0) < \infty$, where $y \to \infty$ must be

$$\log\left(h(y,t,\lambda_0,\alpha_0)\right) - \log(\alpha_0) \approx \frac{1}{2}\log\left(\frac{2}{\pi}\right) - \log(y) + ty - \frac{1}{2}(\lambda_0^2+1)(\log(y))^2 \to \infty$$

Now, since $g(y,\lambda_0,\alpha_0) \to 1$, when $y \to \infty$, then we conclude that $J_{(\lambda_0,\alpha_0)} \to \infty$ when $y \to \infty$.

2.3. Inference

The maximum likelihood estimation and observed and expected matrix information for the parameters of the $LPSN(\xi,\eta,\lambda,\alpha)$ model are studied. For a random sample of size n, $Y_1, Y_2, \ldots, Y_n$, with $Y_i \sim LPSN(\xi,\eta,\lambda,\alpha)$, the log-likelihood function of $\theta = (\xi,\eta,\lambda,\alpha)'$ given $\boldsymbol{Y}$, can be expressed by

$$\ell(\theta,\mathbf{Y}) = n\left(\log(\alpha) - \log(\eta)\right) - \sum_{i=1}^{n}\log(y_i) - \frac{1}{2}\sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n}\log\left\{\Phi(\lambda z_i)\right\} + (\alpha-1)\sum_{i=1}^{n}\log\left\{\Phi_{SN}(z_i;\lambda)\right\}$$

where $z_i = \frac{\log(y_i)-\xi}{\eta}$. The elements of the score function are given by

$$U(\xi) = \frac{1}{\eta}\sum_{i=1}^{n} z_i - \frac{\lambda}{\eta}\sum_{i=1}^{n} w_i - \frac{\alpha-1}{\eta}\sum_{i=1}^{n} w_{1i}$$

$$U(\eta) = -\frac{n}{\eta} + \frac{1}{\eta}\sum_{i=1}^{n} z_i^2 - \frac{\lambda}{\eta}\sum_{i=1}^{n} z_i w_i - \frac{\alpha-1}{\eta}\sum_{i=1}^{n} w_{1i} z_i$$

$$U(\lambda) = \sum_{i=1}^{n} z_i w_i - \sqrt{\frac{2}{\pi}}\frac{(\alpha-1)}{1+\lambda^2}\sum_{i=1}^{n} w_i(\lambda)$$

and

$$U(\alpha) = \frac{n}{\alpha} + \sum_{i=1}^{n}\log\left\{\Phi_{SN}(z_i;\lambda)\right\}$$

where $w = \frac{\phi(\lambda z)}{\Phi(\lambda z)}$, $w_1 = \frac{\phi_{SN}(z)}{\Phi_{SN}(z;\lambda)}$ and $w(\lambda) = \frac{\phi\left(\sqrt{1+\lambda^2}z\right)}{\Phi_{SN}(z;\lambda)}$. The score equations are obtained by equating these partial derivatives to zero. The maximum likelihood estimators (MLEs) are the solutions to the score equations. These solutions are usually obtained by iterative numerical methods.

2.3.1. Observed Information Matrix

The elements of the observed information matrix are defined without the second derivative of the log-likelihood function with respect to parameter denoted by $j_{\xi\xi}, j_{\eta\xi}, \ldots, j_{\alpha\alpha}$ which can be written as

$$j_{\xi\xi} = \frac{n}{\eta^2} + \frac{\lambda^2}{\eta^2}\sum_{i=1}^{n}\lambda z_i w_i + \frac{\lambda^2}{\eta^2}\sum_{i=1}^{n} w_i^2 + \frac{\alpha-1}{\eta^2}\sum_{i=1}^{n} w_{1i}(z_i + w_{1i}) - \sqrt{\frac{2}{\pi}}\frac{\lambda(\alpha-1)}{\eta^2}\sum_{i=1}^{n} w_i(\lambda)$$

$$j_{\eta\xi} = \frac{2}{\eta^2}\sum_{i=1}^{n} z_i + \frac{\lambda^3}{\eta^2}\sum_{i=1}^{n} z_i^2 w_i + \frac{\lambda^2}{\eta^2}\sum_{i=1}^{n} z_i w_i^2 - \frac{\lambda}{\eta^2}\sum_{i=1}^{n} w_i - \sqrt{\frac{2}{\pi}}\frac{\lambda(\alpha-1)}{\eta^2}\sum_{i=1}^{n} z_i w_i(\lambda) + \frac{\alpha-1}{\eta^2}\sum_{i=1}^{n} w_{1i}(-1 + z_i^2 + z_i w_{1i})$$

$$j_{\lambda\xi} = \frac{1}{\eta}\sum_{i=1}^{n}\left[w_i - \lambda^2 z_i^2 w_i - \lambda z_i w_i^2\right] + \sqrt{\frac{2}{\pi}}\frac{\alpha-1}{\eta}\sum_{i=1}^{n} w_i(\lambda)\left[z_i + \frac{1}{1+\lambda^2} w_{1i}\right], \quad j_{\alpha\xi} = \frac{1}{\eta}\sum_{i=1}^{n} w_{1i}$$

$$j_{\eta\eta} = -\frac{n}{\eta^2} + \frac{3}{\eta^2}\sum_{i=1}^{n} z_i^2 - \frac{2\lambda}{\eta^2}\sum_{i=1}^{n} z_i w_i + \frac{\lambda^3}{\eta^2}\sum_{i=1}^{n} z_i^3 w_i + \frac{\lambda^2}{\eta^2}\sum_{i=1}^{n} z_i^2 w_i^2 - \sqrt{\frac{2}{\pi}}\frac{\lambda(\alpha-1)}{\eta^2}\sum_{i=1}^{n} z_i^2 w_i(\lambda) + \frac{\alpha-1}{\eta^2}\sum_{i=1}^{n} z_i w_{1i}\left[-2 + z_i^2 + z_i w_{1i}\right]$$

$$j_{\lambda\eta} = \frac{1}{\eta}\sum_{i=1}^{n} z_i w_i - \frac{\lambda^2}{\eta}\sum_{i=1}^{n} z_i^3 w_i - \frac{\lambda}{\eta}\sum_{i=1}^{n} z_i^2 w_i^2 + \sqrt{\frac{2}{\pi}}\frac{\alpha-1}{\eta}\sum_{i=1}^{n} z_i w_i(\lambda)\left[z_i + \frac{1}{1+\lambda^2} w_{1i}\right]$$

$$j_{\lambda\lambda} = \sum_{i=1}^{n} z_i^2(\lambda z_i w_i + w_i^2) - \sqrt{\frac{2}{\pi}}\frac{2\lambda(\alpha-1)}{(1+\lambda^2)^2}\sum_{i=1}^{n} w_i(\lambda) + 2(\alpha-1)\sum_{i=1}^{n}\left[-\sqrt{\frac{1}{2\pi}}\frac{\lambda}{1+\lambda^2} z_i^2 w_i(\lambda) + \frac{1}{\pi}\frac{1}{(1+\lambda^2)^2} w_i^2(\lambda)\right]$$

and

$$j_{\alpha\eta} = \frac{1}{\eta}\sum_{i=1}^{n} z_i w_{1i}, \qquad j_{\alpha\lambda} = \sqrt{\frac{2}{\pi}}\frac{1}{1+\lambda^2}\sum_{i=1}^{n} w_i(\lambda), \qquad j_{\alpha\alpha} = \frac{n}{\alpha^2}$$

2.3.2. Expected Information Matrix

The elements of the expected information matrix are the expected values of the elements of the observed information matrix; let $i_{\xi\xi}, i_{\eta\xi}, \ldots, i_{\alpha\alpha}$ be the elements of the observed information matrix multiplied by n^{-1}, calling $a_{jk} = \mathbb{E}(z^j w^k)$, $a_{1jk} = \mathbb{E}(z^j w_1^k)$ and $a_{jk}(\lambda) = \mathbb{E}(z^j w^k(\lambda))$. The elements of the expected information matrix can be written as

$$i_{\xi\xi} = \frac{1}{\eta^2} + \frac{\lambda^3}{\eta^2}a_{11} + \frac{\lambda^2}{\eta^2}a_{02} - \sqrt{\frac{2}{\pi}}\frac{\lambda(\alpha-1)}{\eta^2}a_{01}(\lambda) + \frac{\alpha-1}{\eta^2}(a_{111}+a_{102})$$

$$i_{\eta\xi} = \frac{2}{\eta^2}a_{10} + \frac{\lambda^3}{\eta^2}a_{21} + \frac{\lambda^2}{\eta^2}a_{12} - \frac{\lambda}{\eta^2}a_{10} - \sqrt{\frac{2}{\pi}}\frac{\lambda(\alpha-1)}{\eta^2}a_{11}(\lambda) + \frac{\alpha-1}{\eta^2}(-a_{101}+a_{121}+a_{112})$$

$$i_{\lambda\xi} = \frac{1}{\eta}\left[a_{01} - \lambda^2 a_{21} - \lambda a_{12}\right] + \sqrt{\frac{2}{\pi}}\frac{\alpha-1}{\eta}\Big[a_{11}(\lambda) + \frac{1}{1+\lambda^2}\mathbb{E}(w_1 w(\lambda))\Big], \quad i_{\alpha\xi} = \frac{1}{\eta}a_{101}$$

$$i_{\eta\eta} = -\frac{1}{\eta^2} + \frac{3}{\eta}a_{20} - \frac{2\lambda}{\eta^2}a_{11} + \frac{\lambda^3}{\eta}a_{31} + \frac{\lambda^2}{\eta^2}a_{22} - \sqrt{\frac{2}{\pi}}\frac{\lambda(\alpha-1)}{\eta^2}a_{21}(\lambda) + \frac{\alpha-1}{\eta^2}(-2a_{111}+a_{131}+a_{122})$$

$$i_{\lambda\eta} = \frac{1}{\eta}\left[a_{11} - \lambda^2 a_{31} - \lambda a_{22}\right] + \sqrt{\frac{2}{\pi}}\frac{\alpha-1}{\eta}\Big[a_{21}(\lambda) + \frac{1}{1+\lambda^2}\mathbb{E}(z w_1 w(\lambda))\Big], \quad i_{\alpha\eta} = \frac{1}{\eta}a_{111}$$

$$i_{\lambda\lambda} = \lambda a_{31} + a_{22} - \sqrt{\frac{2}{\pi}}\frac{2\lambda(\alpha-1)}{(1+\lambda^2)^2}a_{01}(\lambda) + \sqrt{\frac{2}{\pi}}(\alpha-1)\Big[-\frac{\lambda}{1+\lambda^2}a_{21}(\lambda) + \sqrt{\frac{2}{\pi}}\frac{1}{(1+\lambda^2)^2}a_{02}(\lambda)\Big]$$

$$i_{\alpha\lambda} = \sqrt{\frac{2}{\pi}}\frac{1}{1+\lambda^2}a_{01}(\lambda), \qquad i_{\alpha\alpha} = \frac{1}{\alpha^2}$$

For $\lambda = 0$ and $\alpha = 1$ use the approximation

$$\frac{1}{\pi}\frac{\phi(z)}{\sqrt{\Phi(z)[1-\Phi(z)]}} \approx \frac{1}{\sqrt{2\pi}(\pi/2)}\exp\left(-\frac{z^2}{2(\pi^2/4)}\right)$$

given in Chaibub-Neto & Branco (2003). The expected information matrix is

$$I_F(\theta) = \begin{pmatrix} \frac{1}{\eta^2} & 0 & \sqrt{\frac{2}{\pi}}\frac{1}{\eta} & \frac{\sqrt{\pi}}{2}\frac{1}{\eta} \\ 0 & \frac{2}{\eta^2} & 0 & \frac{1}{4\eta}\frac{\pi^2}{\sqrt{8+\pi^2}} \\ \sqrt{\frac{2}{\pi}}\frac{1}{\eta} & 0 & \frac{2}{\pi} & \sqrt{\frac{1}{2}} \\ \frac{\sqrt{\pi}}{2}\frac{1}{\eta} & \frac{1}{4\eta}\frac{\pi^2}{\sqrt{8+\pi^2}} & \sqrt{\frac{1}{2}} & 1 \end{pmatrix} \tag{9}$$

whose determinant $|I_F(\theta)| = 0$.

Therefore, we conclude that the expected information matrix of the model is singular for the special case of a LN distribution. The upper 3×3 submatrix is the expected information matrix from the log-skew-normal distribution.

As in (9) the third column (respectively, row) is equal to first column (respectively, row) multiply by $\eta\sqrt{\frac{2}{\pi}}$, $I_F(\theta)$ is singular. Using results from Rotnitzky, Cox, Bottai & Robins (2000) we find the asymptotic distribution of the maximum likelihood estimator of θ. DiCiccio & Monti (2004) explains: "(Rotnitzky et al. 2000) *derived the asymptotic distribution of the MLE* $\widehat{\theta} = (\widehat{\theta}^1, \widehat{\theta}^2, \ldots, \widehat{\theta}^q)$ *under two conditions: a single component of the score function, say* S_{θ^1}*, vanishes at some point* $\theta = \theta^*$*, and some higher-order derivatives of* S_{θ^1} *taken with respect to* θ^1 *are possibly 0 at that point but the first nonzero derivative is not a linear combination of the other score function components* $S_{\theta^2}, \ldots, S_{\theta^q}$".

Using an iterative process suggested by Rotnitzky et al. (2000), we find a new parameterization to PSN model that fulfill the two conditions in the same way that Chiogna (1998) and DiCiccio & Monti (2004) for the skew-normal distribution and the skew exponential power distribution, respectively. Let $\theta* = (\xi*, \eta*, 0, 1)$ denote the vector parameter of interest. For $\theta = \theta*$, let $S_\theta(\theta*, Y) = \partial\ell/\partial\theta^* = (S^*_\xi, S^*_\eta, S^*_\lambda, S^*_\alpha)$ denote the score vector, so

$$S_\theta(\theta*, Y) = \left(\frac{Z^*}{\eta^*}, \frac{Z^{*2}-1}{\eta^*}, \sqrt{\frac{2}{\pi}}Z^*, 1 + \log(\Phi(Z^*))\right)$$

whit $Z^* = \frac{Y-\xi^*}{\eta^*}$. After some calculations we take the new parameterization $\tilde{\theta} = \tilde{\theta}(\theta) = (\tilde{\xi}, \tilde{\eta}, \lambda, \alpha)$ with $\tilde{\xi} = \xi + \sqrt{\frac{2}{\pi}}\eta^*\lambda$ and $\tilde{\eta} = \eta - \eta^*\frac{\lambda^2}{\pi}$.

Making use of Theorem 3 in Rotnitzky et al. (2000) with the new parameterization we can conclude that:

1. The MLE of θ is unique with probability tending to 1, and it is consistent.
2. The likelihood ratio statistic for testing the simple null hypothesis $H_0 : \theta = \theta^*$ converges in distribution to the χ^2 distribution with four degrees of freedom.
3. The random vector
   $\left(n^{1/2}(\tilde{\xi} - \xi + \sqrt{\frac{2}{\pi}}\eta^*\lambda), n^{1/2}(\tilde{\eta} - \eta - \eta^*\frac{\lambda^2}{\pi}), n^{1/6}\widehat{\lambda}, n^{1/2}(\widehat{\alpha} - 1)\right)$
   converges to $(Y_1, Y_2, Y_3^{1/3}, Y_4)$, where (Y_1, Y_2, Y_3, Y_4) is a normal random vector with mean zero and covariance matrix equals to the inverse of the covariance matrix

$$\begin{pmatrix} \frac{1}{\eta^2} & 0 & \frac{2-\pi}{\sqrt{2\pi^3}}\frac{1}{\eta} & \frac{\sqrt{\pi}}{2}\frac{1}{\eta} \\ 0 & \frac{2}{\eta^2} & 0 & \frac{1}{4\eta}\frac{\pi^2}{\sqrt{8+\pi^2}} \\ \frac{2-\pi}{\sqrt{2\pi^3}}\frac{1}{\eta} & 0 & \frac{5\pi^2-28\pi+44}{6\pi^3} & \sqrt{\frac{1}{2}} \\ \frac{\sqrt{\pi}}{2}\frac{1}{\eta} & \frac{1}{4\eta}\frac{\pi^2}{\sqrt{8+\pi^2}} & \sqrt{\frac{1}{2}} & 1 \end{pmatrix}^{-1}$$

2.4. Illustration

Precipitation data (measured in inches) were collected from the Colombian Institute of Hydrology, Meteorology and Environmental Studies in Córdoba, Colombia (IDEAM 2006). Descriptive statistics for the variable under study are provided in Table 2. The quantities $\sqrt{\widehat{\beta}_1} = \sqrt{b_1}$ and $\widehat{\beta}_2 = b_2$, where β_1 and β_2 defined in (7), indicate the asymmetry and kurtosis coefficients respectively.

Table 2: Descriptive statistics of the precipitation variable

Variables	n	Mean	Variance	$\sqrt{b_1}$	b_2
Y	273	4.8360	9.7871	0.4632	2.6035
$log(Y)$	273	1.2219	1.1155	-1.5608	5.5276

The asymmetry and kurtosis coefficients are different from the corresponding values expected for LN model and normal model. Precipitation data are fitted using the LPSN model.

The LPSN model is compared to the LN model as well as the LSN model to the $LPSN_{\lambda=0}$ model. The maximum likelihood method for estimating the parameters is used and the Akaike information criterion (AIC), (Akaike 1974), is applied for

contrast. Firstly, the LN model is compared to the LPSN model by the hypothesis tests

$$H_0 : (\lambda, \alpha) = (0, 1) \quad \text{versus} \quad H_1 : (\lambda, \alpha) \neq (0, 1)$$

Using the likelihood ratio statistic,

$$\Lambda = \frac{\ell_{LN}(\widehat{\theta})}{\ell_{LPSN}(\widehat{\theta})}$$

we obtain

$$-2\log(\Lambda) = -2(-735.4023 + 670.2293) = 130.346$$

which is greater than the value of the $\chi^2_{2,95\%} = 5.99$. Then the LPSN model is a good alternative for fitting the precipitation data. The LPSN model is also compared to the $LPSN_{\lambda=0}$ model and the LSN models by the hypothesis tests

$$H_{01} : \lambda = 0 \quad \text{versus} \quad H_{11} : \lambda \neq 0, \qquad \text{and} \qquad H_{02} : \alpha = 1 \quad \text{versus} \quad H_{12} : \alpha \neq 1$$

respectively, using the likelihood ratio statistics

$$\Lambda_1 = \frac{\ell_{LPSN_{\lambda=0}(\theta)}}{\ell_{LPSN}(\theta)} \qquad \text{and} \qquad \Lambda_2 = \frac{\ell_{LSN}(\theta)}{\ell_{LPSN}(\theta)}$$

After numerical evaluations, we obtain

$$-2\log(\Lambda_1) = 61.5960 \qquad \text{and} \qquad -2\log(\Lambda_2) = 15.5056$$

which is greater than the value of the $\chi^2_{1,95\%} = 3.84$. The best fit, with respect to the other models, is shown by the LPSN model. Table 3 presents the MLEs and the estimated standard errors (in parentheses) for LN, LSN, LPSN and models. Figure 3 shows the histogram of precipitacion data and fitted curves for the proposed models in which the LPSN model presents the better fit of asymmetry and kurtosis with respect to the other models.

Table 3: Parameters and estimated standard errors of the log-normal (LN), log-skew-normal (LSN), log-skew-normal alpha-power $\lambda = 0$ ($LPSN_{\lambda=0}$), and the log-skew-normal alpha-power (LPSN) distributions.

Parameter	**Log-normal**	**LSN**	**LPSN$_{\lambda=0}$**	**LPSN**
Loglik	-735.4023	-677.9821	-701.0273	-670.2292
AIC	1474.8050	1361.964	1408.0550	1348.5490
ξ	1.2219(0.0638)	2.4217(0.0392)	2.8280(0.0817)	2.2647(0.0529)
η	1.0542(0.0451)	1.5971(0.0763)	0.1668(0.0507)	4.8760(0.3363)
λ	–	-10.0515(2.2917)	–	-19.2702(2.4450)
α	–	–	0.0144(0.0008)	4.8579(0.5925)

The Figure 4 shows the qqplots for LN, LSN and LPSN models. The LPSN model shows better fit with respect to the LN and LSN models.

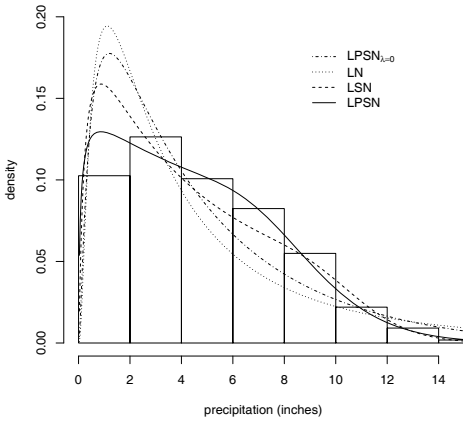
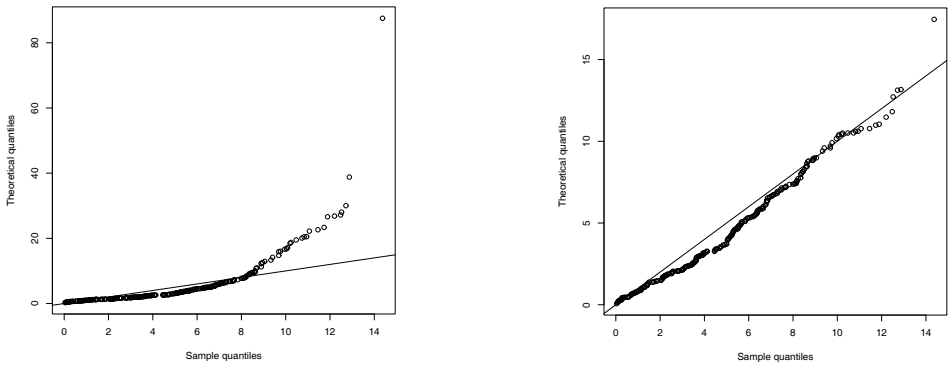
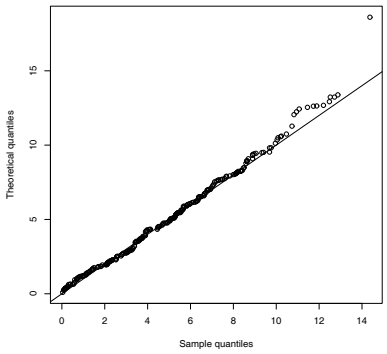


FIGURE 3: Histogram of the precipitation data. Densities are estimated by maximum likelihood.



(a) (b)



(c)

FIGURE 4: Q-Qplot: (a) log-normal model, (b) log-skew-normal model, and (c) log-skew-normal alpha-power model.

3. Conclusion

In this paper we propose a more flexible model than LN and LSN models fit data with greater asymmetry and more platikurtic or leptokurtic than Azzalini (1985) and Durrans (1992) models. General expressions for the moments are found, maximum likelihood estimators are studied, observed and expected information matrix are found, and also an asymptotic distribution of a MLEs vector is found. Finally, an illustration is presented (see Figure 4). We contrast the LN, LSN, and LPSN models through some precipitation data. According to AIC selection criterion, the LPSN model makes the better fit with respect to the other models considered.

[Recibido: junio de 2012 — Aceptado: abril de 2013]

References

Akaike, H. (1974), 'A new look at statistical model identification', *IEEE Transaction on Automatic Control* (AU-19), 716–722.

Arnold, B. C. & Beaver, R. (2002), 'Skewed multivariate models related to hidden truncation and/or selective reporting'.

Azzalini, A. (1985), 'A class of distributions which includes the normal ones', *Scandinavian Journal of Statistics* (12), 171–178.

Chaibub-Neto, E. & Branco, M. (2003), *Bayesian Reference Analysis for Binomial Calibration Problem*, IME-USP.

Chiogna, M. (1998), 'Some results on the scalar skew-normal distribution', *Journal Italian Statistical Society* **1**, 1–13.

DiCiccio, T. J. & Monti, A. C. (2004), 'Inferential aspects of the skew exponential power distribution', *Journal of the American Statistical Association* **99**, 439–450.

Durrans, S. R. (1992), 'Distributions of fractional order statistics in hydrology', *Water Resources Research* pp. 1649–1655.

Gupta, D. & Gupta, R. C. (2008), 'Analyzing skewed data by power normal model', *Test* **17**(1), 197–210.

Gupta, R. S. & Gupta, R. D. (2004), 'Generalized skew normal model', *Test* **13**(2), 501–524.

IDEAM (2006), *Estudio Agroclimático del Departamento de Córdoba*, Fondo Editorial Universidad de Córdoba.

Lin, G. D. & Stoyanov, J. (2009), 'The logarithmic skew-normal distributions are moment-indeterminate', *Journal of Applied Probability* **46**(3), 909–916.

Martínez-Flórez, G. (2011), Extensões do modelo α-potêncial, Tese de doutorado, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

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

Rotnitzky, A., Cox, D. R., Bottai, M. & Robins, J. (2000), 'Likelihood-based inference with singular information matrix', *Bernoulli* **6**(2), 243–284.

Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 59 a 77

On the Moment Characteristics for the Univariate Compound Poisson and Bivariate Compound Poisson Processes with Applications

Sobre las características de los momentos de los procesos de Poisson compuestos univariados y bivariados con aplicaciones

Gamze Özel[1,a]

[1]Department of Statistics, The Faculty of Science, Hacettepe University, Ankara, Turkey

Abstract

The univariate and bivariate compound Poisson process (CPP and BCPP, respectively) ensure a better description than the homogeneous Poisson process for clustering of events. In this paper, new explicit representations of the moment characteristics (general, central, factorial, binomial and ordinary moments, factorial cumulants) and some covariance structures are derived for the CPP and BCPP. Then, the skewness and kurtosis of the univariate CPP are obtained for the first time and special cases of the CPP are studied in detail. Applications to two real data sets are given to illustrate the usage of these processes.

***Key words*:** Bivariate distribution, Compound Poisson process, Cumulant, Factorial moments, Moment.

Resumen

Los procesos univariados y bivariados compuestos de Poisson (CPP y BCCPP, por sus siglas en inglés respectivamente) permiten una mejor descripción que los procesos homogéneos de Poisson para agrupamiento de eventos. En este artículo, se muestran específicamente las representaciones de las características de momentos (general, central, factorial, momentos binomiales y ordinarios, acumuladas factoriales) y algunas estructuras de covarianza para los CPP y BCPP. Adicionalmente, el sesgo y la curtosis de los procesos univariados CPP son presentados y casos especiales son estudiados en detalle. La aplicación a dos conjuntos de datos reales es usada con el fin de ilustrar el uso de estos procesos.

***Palabras clave*:** acumuladas factoriales, conjuntas, distribución bivariada, distribución compuesta de Poisson, momento.

[a]Lecturer. E-mail: gamzeozl@hacettepe.edu.tr

1. Introduction

Let $\{N_t, t \geq 0\}$ be a homogeneous Poisson process with parameter $\lambda > 0$ and let X_i, $i = 1, 2, \ldots$, be identically and independent distributed (i.i.d.) non-negative, integer-valued random variables, independent of $\{N_t, t \geq 0\}$. Then, $\{S_t, t \geq 0\}$ has a univariate CPP if it is defined as

$$S_t = \sum_{i}^{N_t} X_i \tag{1}$$

The univariate CPP has many applications in various areas such as transport, ecology, radiobiology, quality control, telecommunications (see Ata & Özel 2012, Chen, Randolph & Tian-Shy 2005, Gudowska-Nowak, Lee, Nasonova, Ritter & Scholz 2007, Özel & Inal 2008, Robin 2002, Rosychuk, Huston & Prasad 2006). However, the investigation of the properties of the univariate CPP mixtures is much more complicated than the homogeneous Poisson process. The applications of the univariate CPP often run into the obstacle of numerical evaluation of the corresponding probability functions. Hence, moment characteristics of the univariate CPP play a very important role in the probability theory.

Bivariate stochastic processes have also received considerable attention in the literature, in an effort to explain phenomena in various areas of application (see Kocherlakota & Kocherlakota 1997, Özel 2011*a*, Wienke, Ripatti, Palmgren & Yashin 2010, Wienke 2011). Paired count data in time 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 (the 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) and econometrics (number of voluntary and involuntary job changes). In this study we consider the following BCPP. Let $\{N_t, t \geq 0\}$ be a homogeneous Poisson process and let X_i, Y_i, $i = 1, 2, \ldots$, be independent of the process $\{N_t, t \geq 0\}$ Then the BCPP is defined as

$$\left(S_t^{(1)} = \sum_{i}^{N_t} X_i, S_t^{(2)} = \sum_{i}^{N_t} Y_i \right) \tag{2}$$

where $X_i, Y_i,\ i = 1, 2, \ldots$, are mutually independent random variables.

The CPP is studied in Özel & Inal (2008) but mainly from the evaluation of its probability function. The recursive formulas for the joint probability functions of the BCPP in (2) are derived by Hesselager (1996) and Sundt (1992). Özel & Inal (2008) defined a different kind of BCPP and obtained the joint probability function, moments and cumulants. On the other hand, non-existence of moment characteristics obstacles usage of them in probability theory itself and its applications in seismology, actuarial science, survival analysis, etc. Consequently, since relative results are sparse and case oriented, the aim of this study is to obtain the moment characteristics and covariance structures of the univariate CPP and BCPP.

The paper is organised as follows. In Section 2, moments, cumulants and some relationships are derived for the first time and special cases are obtained for the univariate CPP. In Section 3, new explicit expressions for the moments, cumulants, covariances, and correlation coefficients of the BCPP are derived. In Section 4, the results are illustrated on two real data sets. The conclusion is given in Section 5.

2. The Univariate Compound Poisson Process

2.1. Moments of the Univariate CPP

The moment generating function (mgf) makes it possible to compute general (raw) moments of $\{S_t, t \geq 0\}$. Let X_i, $i = 1, 2, \ldots$, be i.i.d. discrete random variables in (1) with the probabilities $P(X_i = j) = p_j$, $j = 0, 1, \ldots$ The common mgf of X_i, $i = 1, 2, \ldots$, is given by $M_x(u) = \sum_{j=0}^{\infty} p_j u^j = p_0 + p_1 u + p_2 u^2 + \cdots$ and the mgf of $\{S_t, t \geq 0\}$ is given by

$$\begin{aligned} M_{S_t}(u) &= \sum_{n=0}^{\infty} \exp(-\lambda t) \frac{(\lambda t)^n}{n!} [M_x(u)]^n \\ &= \exp(-\lambda t) \left[1 + \frac{\lambda t M_x(u)}{1!} + \frac{[\lambda t M_x(u)]^2}{2!} + \cdots \right] \\ &= \exp(\lambda t \left[M_x(u) - 1 \right]) \end{aligned} \tag{3}$$

Let us assume that the random variable X takes finite values $j = 0, 1, \ldots, m$. Define the parameters $\lambda_j = \lambda p_j$, $j = 0, 1, \ldots, m$, then we have

$$M_{S_t}(u) = \exp\left[-\lambda t(1 - p_0)\right] \exp\left[\lambda_1 t \exp(u) + \ldots + \lambda_m t \exp(u^m)\right] \tag{4}$$

Thus, the rth general moment of the univariate CPP can be obtained by differentiating (4) with respect to u and substituting in $\mu'_r = E\left(S_t^r\right) = \left. \frac{d^r}{du^r} M_{S_t}(u) \right|_{u=0}$, $r = 1, 2, \ldots, n$, after some algebraic manipulations, the general moments of $\{S_t, t \geq 0\}$ are obtained as follows:

$$\begin{aligned} \mu'_1 &= (\lambda t \xi_1) \\ \mu'_2 &= (\lambda t \xi_1)^2 + (\lambda t \xi_2) \\ \mu'_3 &= (\lambda t \xi_1)^3 + 3\,(\lambda t \xi_1)\,(\lambda t \xi_2) + (\lambda t \xi_3) \\ \mu'_4 &= (\lambda t \xi_1)^4 + 6(\lambda t \xi_1)^2\,(\lambda t \xi_2) + 4\,(\lambda t \xi_3)\,(\lambda t \xi_1) + 3(\lambda t \xi_2)^2 + (\lambda t \xi_4) \\ \mu'_5 &= (\lambda t \xi_1)^5 + 10(\lambda t \xi_1)^3\,(\lambda t \xi_2) + 10\,(\lambda t \xi_3)\,(\lambda t \xi_1)^2 + 15\,(\lambda t \xi_1)\,(\lambda t \xi_2)^2 \\ &\quad + 5\,(\lambda t \xi_4)\,(\lambda t \xi_1) + 10\,(\lambda t \xi_2)\,(\lambda t \xi_3) + (\lambda t \xi_5) \end{aligned} \tag{5}$$

where $\xi_r = E(x^r)$, $r = 1, 2, \ldots, n$, is the rth general moment of X_i, $i = 1, 2, \ldots$, and $\{N_t, t \geq 0\}$ is a homogeneous Poisson process with parameter $\lambda > 0$ in (1).

A recursive formula for the factorial moments of $\{S_t, t \geq 0\}$ is derived from (3). For this aim, we observe that

$$\frac{dM_{S_t}(u)}{du} = \lambda t M_{S_t}(u) \frac{dM_X(u)}{du}$$

so that applying the Leibniz differentiation rule for $r \geq 1$ we obtain

$$\begin{aligned}\mu'_r &= \lambda t \, \frac{d^{r-1}}{du^{r-1}} \left[\exp\left[\lambda t (M_x(u) - 1)\right] \frac{dM_x(u)}{du} \right] \Bigg|_{u=0} \\ &= \lambda t \sum_{k=0}^{r-1} \binom{r-1}{k} \frac{d^k M_{S_t}(u)}{du^k} \frac{d^{r-k} M_X(u)}{du^{r-k}} \Bigg|_{u=0}\end{aligned}$$

Then, the following recursive formula for the general moments of $\{S_t, t \geq 0\}$ is given by

$$\mu'_r = \lambda t \sum_{k=0}^{r-1} \binom{r-1}{k} \mu'_k \, \xi_{r-k}$$

where ξ_r, $r = 1, 2, \ldots, n$, is the rth general moment of X_i, $i = 1, 2, \ldots$, in (1).

Now consider the central moments μ_r of $\{S_t, t \geq 0\}$. The generating function $G_{S_t}(u)$ of μ_r, if the rth central moment exists, is defined by the relation

$$G_{S_t}(u) = E\left[\exp(u(S_t - \mu))\right] = \exp(-u\mu) M_{S_t}(u) \tag{6}$$

where $\mu'_r = \mu = E(S_t) = \lambda t \xi_1$. Then, rth central moment of $\{S_t, t \geq 0\}$ can be obtained by

$$\mu_r = E(S_t - \mu)^r = \frac{d^r}{du^r} G_{S_t}(0) = \frac{d^r}{du^r} \exp(-u\mu) M_{S_t}(u) \Bigg|_{u=0} \tag{7}$$

From (4) and (7), we have

$$\begin{aligned}
\mu_1 &= (\mu + \lambda t \xi_1) \\
\mu_2 &= (\mu + \lambda t \xi_1)^2 + (\lambda t \xi_2) \\
\mu_3 &= (\mu + \lambda t \xi_1)^3 + 3(\mu + \lambda t \xi_1)(\lambda t \xi_2) + (\lambda t \xi_3) \\
\mu_4 &= (\mu + \lambda t \xi_1)^4 + 6(\mu + \lambda t \xi_1)^2 (\lambda t \xi_2) + 4(\mu + \lambda t \xi_1)(\lambda t \xi_3) \\
&\qquad + 3(\lambda t \xi_2)^2 + (\lambda t \xi_4) \\
\mu_5 &= (\mu + \lambda t \xi_1)^5 + 10(\mu + \lambda t \xi_1)^3 (\lambda t \xi_2) + 10(\mu + \lambda t \xi_1)^2 (\lambda t \xi_3) \\
&\quad + 15(\mu + \lambda t \xi_1)(\lambda t \xi_2)^2 + 5(\mu + \lambda t \xi_1)(\lambda t \xi_4) + 10(\lambda t \xi_2)(\lambda t \xi_3) + (\lambda t \xi_5)
\end{aligned} \tag{8}$$

where ξ_r, $r = 1, 2, \ldots, n$, is the rth general moment of X_i, $i = 1, 2, \ldots$

Commonly used indices of the shape of a distribution are the moment ratios such as skewness and kurtosis. Since $\{S_t, t \geq 0\}$ has finite moments of orders up to the third, then the skewness of S_t is defined as

$$\sqrt{\beta_1} = E\left(\frac{S_t - \mu}{\sigma}\right)^3 = \frac{\mu_3}{\mu_2^{3/2}} \tag{9}$$

where σ is the standard deviation of S_t. From (9), the skewness of $\{S_t, t \geq 0\}$ is obtained using the central moments in (8) as follows:

$$\sqrt{\beta_1} = \frac{(\mu + \lambda t\xi_1)^3 + 3(\mu + \lambda t\xi_1)(\lambda t\xi_2) + (\lambda t\xi_3)}{\left[(\mu + \lambda t\xi_1)^2 + (\lambda t\xi_2)\right]^{3/2}} \tag{10}$$

Similarly, the kurtosis of $\{S_t, t \geq 0\}$ is obtained from (8) as

$$\begin{aligned} \beta_2 &= E\left[\frac{S_t - \mu}{\sigma}\right]^4 - 3 = \frac{\mu_4}{\mu_2^2} - 3 \\ &= \frac{4(\mu + \lambda t\xi_1)(\lambda t\xi_3) - 2(\mu + \lambda t\xi_1)^4 + (\lambda t\xi_4)}{\left[(\mu + \lambda t\xi_1)^2 + (\lambda t\xi_2)\right]^2} \end{aligned} \tag{11}$$

Since $M_{S_t}(u)$ is exponential form in (4), it is useful to consider the cumulants (semi invariants) κ_r, defined formally as the coefficients of the Taylor expansion of the logarithm of the characteristic function $\varphi_{S_t}(u)$ and having the cumulant generating function

$$C_{S_t}(u) = \ln \varphi_{S_t}(u) = \sum_{r=1}^{\infty} \kappa_r \frac{(iu)^r}{r!} \tag{12}$$

where i denotes the imaginary number ($i^2 = -1$) and the characteristic function of $\{S_t, t \geq 0\}$ is given by $\varphi_{S_t}(u) = \exp[\lambda t(\varphi_X(u) - 1)]$. Here, $\varphi_X(u)$ the common characteristic function of X_i, $i = 1, 2, \ldots$ Then, if X takes finitely many values $j = 0, 1, \ldots, m$, we get

$$\begin{aligned} C_{S_t}(u) &= \lambda t\left[\varphi_X(u) - 1\right] \\ &= \lambda t\left[p_0 + p_1 \exp(iu) + p_2 \exp(2iu) + \cdots + p_m \exp(miu)\right] - \lambda t \\ &= \lambda t\left[(p_0 - 1) + p_1 \exp(iu) + p_2 \exp(2iu) + \cdots + p_m \exp(miu))\right] \end{aligned} \tag{13}$$

Using Taylor series expansion, we obtain a cumulant generating function from (13) as

$$\begin{aligned} C_{S_t}(u) = \lambda t\Bigg[& p_1\left(\frac{(iu)}{1!} + \cdots\right) + p_2\left(\frac{(2iu)}{1!} + \cdots\right) + \cdots \\ & + p_m\left(\frac{(miu)}{1!} + \cdots\right)\Bigg] \end{aligned}$$

$$
\begin{aligned}
&= \left[\frac{(iu)}{1!}\{\lambda t(p_1 + \cdots)\} + \frac{(iu)^2}{2!}\{\lambda t(p_1 + \cdots)\} + \cdots \right. \\
&\qquad \left. + \frac{(iu)^m}{m!}\{\lambda t(p_1 + \cdots)\} \right] \\
&= \sum_{r=1}^{\infty} \left(\lambda t \sum_{j=1}^{\infty} r^j p_j \right) \frac{(iu)^r}{r!} \\
&= \sum_{r=1}^{\infty} \lambda t E(X^r) \frac{(iu)^r}{r!}
\end{aligned} \tag{14}
$$

Then, for every $r = 1, 2, \ldots, n$ we have

$$\kappa_r = \lambda t \xi_r \tag{15}$$

Here, $\xi_r = E(X^r)$, $r = 1, 2, \ldots n$, is the rth general moment of X_i, $i = 1, 2, \ldots$ We also obtain a relationship between the general moments and the ordinary cumulants of $\{S_t, t \geq 0\}$ as follows:

$$
\begin{aligned}
\mu_1' &= \kappa_1 \\
\mu_2' &= \kappa_1^2 + \kappa_2 \\
\mu_3' &= \kappa_1^3 + 3\kappa_1\kappa_2 + \kappa_3 \\
\mu_4' &= \kappa_1^4 + 6\kappa_1^2\kappa_2 + 4\kappa_1\kappa_3 + 3\kappa_2^2 + \kappa_4 \\
\mu_5' &= \kappa_1^5 + 10\kappa_1^3\kappa_2 + 10\kappa_1^2\kappa_3 + 15\kappa_1\kappa_2^2 + 5\kappa_1\kappa_4 + 10\kappa_2\kappa_3 + \kappa_5
\end{aligned} \tag{16}
$$

In problems with discrete random variables one often uses the factorial moments. Let $\mu_{[r]}$ be the rth factorial moment of $\{S_t, t \geq 0\}$ in (1). $\mu_{[r]}$ can be obtained by inverting the factorial moment generating function (fmgf) of $\{S_t, t \geq 0\}$

$$\mu_{[r]} = \left. \frac{d^r}{du^r} P_{S_t}(1 + u) \right|_{u=0} \tag{17}$$

where fmgf of $\{S_t, t \geq 0\}$ is

$$
\begin{aligned}
P_{S_t}(1 + u) = \exp[-\lambda t(1 - p_0)] \exp[\lambda_1 t \exp(1 + u) + \lambda_2 t \exp((1 + u)^2) \\
+ \cdots + \lambda_m t \exp((1 + u)^m)]
\end{aligned} \tag{18}
$$

Here, the random variable X has finite values $j = 0, 1, \ldots m$. Differentiating (18) and substituting in (17), after some manipulations, we obtain the factorial moments as follows:

$$\begin{aligned}
\mu_{[1]} &= (\lambda t\xi_{[1]}) \\
\mu_{[2]} &= (\lambda t\xi_{[1]})^2 + (\lambda t\xi_{[2]}) \\
\mu_{[3]} &= (\lambda t\xi_{[1]})^3 + 3(\lambda t\xi_{[1]})(\lambda t\xi_{[2]}) + (\lambda t\xi_{[3]}) \\
\mu_{[4]} &= (\lambda t\xi_{[1]})^4 + 6(\lambda t\xi_{[1]})^2(\lambda t\xi_{[2]}) + 4(\lambda t\xi_{[3]})(\lambda t\xi_{[1]}) + 3(\lambda t\xi_{[2]})^2 + (\lambda t\xi_{[4]}) \\
\mu_{[5]} &= (\lambda t\xi_{[1]})^5 + 10(\lambda t\xi_{[1]})^3(\lambda t\xi_{[2]}) + 10(\lambda t\xi_{[3]})(\lambda t\xi_{[1]})^2 + 15(\lambda t\xi_{[1]})(\lambda t\xi_{[2]})^2 \\
&\quad + 5(\lambda t\xi_{[4]})(\lambda t\xi_{[1]}) + 10(\lambda t\xi_{[2]})(\lambda t\xi_{[3]}) + (\lambda t\xi_{[5]})
\end{aligned} \tag{19}$$

where $\xi_{[r]} = E[(X)(X-1)\cdots(X-(r-1))]$, $r = 1, 2, \ldots, n$, is the rth factorial moment of X_i, $i = 1, 2, \ldots$ If $E(X^r) < \infty$, factorial moments of $\{S_t, t \geq 0\}$ can also be calculated recursively. We observe that

$$\frac{dg_{S_t}(u)}{du} = \lambda t \exp[\lambda t(g_X(u) - 1)]\frac{dg_X(u)}{du} = \lambda t g_{S_t}(u)\frac{dg_X(u)}{du}$$

Now using the Leibniz formula for the derivatives of higher orders, we get

$$\frac{d^r g_{S_t}(u)}{du^r} = \lambda t \sum_{k=0}^{r-1} \binom{r-1}{k} \frac{d^{r-k-1} g_{S_t}(u)}{du^{r-k-1}} \frac{d^{k+1} g_X(u)}{du^{k+1}} \tag{20}$$

From (21) and the relations $\mu_{[r]} = \left.\dfrac{d^r g_{S_t}(u)}{du^r}\right|_{u=1}$, $\xi_r = \left.\dfrac{d^r g_X(u)}{du^r}\right|_{u=1}$, we have

$$\mu_{[r]} = \lambda t \sum_{k=0}^{r-1} \binom{r-1}{k} \mu_{[r-k-1]}\xi_{[k+1]} \tag{21}$$

The logarithm of the fmgf is called factorial cumulant generating function (fcgf). The coefficient of $u^r/r!$ in the Taylor expansion of this function is the rth factorial cumulant $\kappa_{[r]}$. The fcgf is given by $\ln P(1+u) = \sum_{r=1}^{\infty} \dfrac{\kappa_{[r]} u^r}{r!}$ where $\kappa_{[r]}$ denotes the rth factorial cumulant. Then, the factorial cumulants of $\{S_t, t \geq 0\}$ are given by

$$\kappa_{[r]} = \lambda t \xi_{[r]} \tag{22}$$

where $\xi_{[r]} = E[(X)(X-1)\ldots(X-(r-1))]$, $r = 1, 2, \ldots, n$, is the rth factorial moment of X_i, $i = 1, 2, \ldots$.

Let us point out that the factorial cumulants are related to the ordinary cumulants in the same way as the factorial moments are related to the general moments for $\{S_t, t \geq 0\}$. A relationship of the factorial cumulants with the ordinary cumulants is also obtained for $\{S_t, t \geq 0\}$ as follows:

$$\begin{aligned}
\kappa_{[1]} &= \kappa_1 \\
\kappa_{[2]} &= \kappa_2 - \kappa_1 \\
\kappa_{[3]} &= \kappa_3 - 3\kappa_2 + 2\kappa_1 \\
\kappa_{[4]} &= \kappa_4 - 6\kappa_3 + 11\kappa_2 - 6\kappa_1 \\
\kappa_{[5]} &= \kappa_5 - 10\kappa_4 + 35\kappa_3 - 50\kappa_2 + 24\kappa_1
\end{aligned} \tag{23}$$

The binomial moments, closely connected with $\mu_{[r]}$, are defined as $B_r = E\binom{S_t}{r} = \frac{1}{r!}\mu_{[r]}$. The binomial moment generating function is $B_{S_t}(u) = \sum_{j=0} B_j u^j = \sum_{j=0} \mu_{[r]} \frac{u^j}{j!} = P_{S_t}(1+u)$, so that, if $M_{S_t}(u)$ exists, then

$$B_r = \frac{1}{r!}\frac{d^r}{du^r} P_{S_t}(1+u)\Big|_{u=0}$$

Hence, rth binomial moments of $\{S_t, t \geq 0\}$ is obtained as follows:

$$\begin{aligned}
B_1 &= \frac{(\lambda t \xi_{[1]})}{1!}, \\
B_2 &= \frac{\left[(\lambda t \xi_{[1]})^2 + (\lambda t \xi_{[2]})\right]}{2!} \\
B_3 &= \frac{\left[(\lambda t \xi_{[1]})^3 + 3(\lambda t \xi_{[1]})(\lambda t \xi_{[2]}) + (\lambda t \xi_{[3]})\right]}{3!} \\
B_4 &= \frac{\left[(\lambda t \xi_{[1]})^4 + 6(\lambda t \xi_{[1]})^2(\lambda t \xi_{[2]}) + 4(\lambda t \xi_{[3]})(\lambda t \xi_{[1]}) + 3(\lambda t \xi_{[2]})^2 + (\lambda t \xi_{[4]})\right]}{4!} \\
B_5 &= \frac{\left[(\lambda t \xi_{[1]})^5 + 10(\lambda t \xi_{[1]})^3(\lambda t \xi_{[2]}) + 10(\lambda t \xi_{[3]})(\lambda t \xi_{[1]})^2 + 15(\lambda t \xi_{[1]})(\lambda t \xi_{[2]})^2\right]}{5!} \\
&\quad + \frac{\left[5(\lambda t \xi_{[4]})(\lambda t \xi_{[1]}) + 10(\lambda t \xi_{[2]})(\lambda t \xi_{[3]}) + (\lambda t \xi_{[5]})\right]}{5!}
\end{aligned} \tag{24}$$

where $\xi_{[r]} = E[(X)(X-1)\ldots(X-(r-1))]$, $r = 1, 2, \ldots, n$, is the rth factorial moment of X_i, $i = 1, 2, \ldots$.

2.2. The Covariance Structure of CPP

In this section, we derived the covariance between $\{N_t, t \geq 0\}$ and $\{S_t, t \geq 0\}$ (1) for the case that $\{N_t, t \geq 0\}$ is a homogeneous Poisson process with parameter $\lambda > 0$ and X_i, $i = 1, 2, \ldots$. are discrete random variables with finite values $j = 0, 1, \ldots$ The characteristic function of the random vector N_t, S_t is defined as

$$\varphi_{N_t,S_t}(u,v) = E\left[\exp(iuN_t + ivS_t)\right] = E\left[E(\exp(iuN_t + ivS_t)|\, N_t)\right] \tag{25}$$

where i is the imaginary number. (25) can be written as

$$\begin{aligned}
\varphi_{N_t,S_t}(u,v) &= E\left[E(\exp(iuN_t + iv\sum_{i=1}^{N_t} X_i\ )|\, N_t = n)\right] \\
&= E\left[\exp(iuN_t)\prod_{i=1}^{n}\exp(ivX_i)\right] \\
&= E\left[\exp(iuN_t)\left(\prod_{i=1}^{n}\varphi_X(v)\right)\right] \\
&= E\left[\exp(iuN_t)\varphi_X^n(v)\right]
\end{aligned} \tag{26}$$

where $\varphi_X(v)$ is the characteristic function of X_i, $i = 1, 2, \ldots$ Since $\{N_t, t \geq 0\}$ has a homogeneous Poisson process with parameter λ, we get

$$\begin{aligned} \varphi_{N_t,S_t}(u,v) &= \exp(-\lambda t) \sum_{k=0}^{\infty} \frac{[\lambda t \varphi_X(v) \exp(iu)]^k}{k!} \\ &= \exp[\lambda t(\varphi_X(v) \exp(iu) - 1)] \end{aligned} \tag{27}$$

To derive the covariance, we have $E(N_t) = \lambda t$ and $E(S_t) = \lambda t E(X)$. In order to complete the derivation of the covariance of N_t and S_t, we need to evaluate $E(N_t S_t) = \left.\dfrac{\partial^2 \varphi_{N_t,S_t}(u,v)}{\partial u \partial v}\right|_{u=v=0}$. The derivative of $\varphi_{N_t,S_t}(u,v)$ with respect to u is $\dfrac{\partial \varphi_{N_t,S_t}(u,v)}{\partial u} = i\lambda t \varphi_X(u) \varphi_{N_t,S_t}(u,v)$ and the derivative of the latter with respect to v is

$$\begin{aligned} \frac{\partial^2 \varphi_{N_t,S_t}(u,v)}{\partial u \partial v} &= i\lambda t \varphi_X(v) \lambda t \exp(iu) \varphi'_X(v) \varphi_{N_t,S_t}(u,v) + i\lambda t \varphi'_X(v) \varphi_{N_t,S_t}(u,v) \\ &= i\lambda t \varphi_{N_t,S_t}(u,v) \varphi'_X(v) \left[\varphi_X(v) \lambda t \exp(iu) + 1\right] \end{aligned}$$

Since $\varphi_{N_t,S_t}(0,0) = \varphi_X(0) = 1$ and $\varphi'_X(0) = iE(X)$, it follows that

$$\left.\frac{\partial^2 \varphi_{N_t,S_t}(u,v)}{\partial u \partial v}\right|_{u=v=0} = i^2 \lambda t(\lambda t + 1) E(X)$$

Therefore, $E(N_t, S_t) = \lambda t(\lambda t + 1) E(X)$ and the covariance of $\{N_t, t \geq 0\}$ and $\{S_t, t \geq 0\}$ is given by

$$Cov(N_t, S_t) = \lambda t E(X) = \lambda t \xi_1 \tag{28}$$

Hence, the coefficient of correlation is

$$\rho = Corr(N_t, S_t) = \frac{Cov(N_t, S_t)}{\sqrt{Var(N_t) Var(S_t)}} = \frac{\xi_1}{\sqrt{\xi_2}} \tag{29}$$

where $Var(N_t) = \lambda t$ and $Var(S_t) = \lambda t E(X^2) = \lambda t \xi_2$.

2.3. Special Cases of the Univariate CPP

In this section we study some special cases of the univariate CPP. Expressions for various moments and cumulants are presented. The Neyman type A, B and Pólya-Aeppli are four major CPPs. The Neyman type A and B processes are defined by Neyman (1939) as 'contagious'. This definition implies that each favourable event enhances the probability of each succeding event. The Pólya-Aeppli process is derived by Getis (1974) to model the clustered point process. Note that some examples of such processes with their corresponding probability functions are discussed in Özel & Inal (2012).

Example 1. The Neyman Type A Process: Let $\{N_t, t \geq 0\}$ be a homogeneous Poisson process with parameter $\lambda > 0$ and let X_i, $i = 1, 2, \ldots$ be Poisson distributed with parameter ν in (1), then $\{S_t, t \geq 0\}$ is called a Neyman type A or Poisson-Poisson process. First four moments and cumulants of the Neyman type A process are given in Table 1 .

Table 1: First four moments and cumulants of the Neyman type A process.

μ'_1	(λtv)
μ'_2	$(\lambda tv)^2 + [\lambda t(v + v^2)]$
μ'_3	$(\lambda tv)^3 + 3(\lambda tv)[\lambda t(v + v^2)] + [\lambda t(v + 3v^2 + v^3)]$
μ'_4	$(\lambda tv)^4 + 6(\lambda tv)^2[\lambda t(v + v^2)] + 4(\lambda tv)[\lambda t(v + 3v^2 + v^3)] + [\lambda t(v + 7v^2 + 6v^3 + v^4)]$
μ_1	$(2\lambda tv)$
μ_2	$(2\lambda tv)^2 + [\lambda t(v + v^2)]$
μ_3	$(2\lambda tv)^3 + 6(\lambda tv)[\lambda t(v + v^2)] + [\lambda t(v + 3v^2 + v^3)]$
μ_4	$(2\lambda tv)^4 + 6(2\lambda tv)^2[\lambda t(v + v^2)] + 4(2\lambda tv)[\lambda t(v + 3v^2 + v^3)] + [\lambda t(v + 7v^2 + 6v^3 + v^4)]$
κ_1	(λtv)
κ_2	$[\lambda t(v + v^2)]$
κ_3	$[\lambda t(v + 3v^2 + v^3)]$
κ_4	$[\lambda t(v + 7v^2 + 6v^3 + v^4)]$
$\mu_{[1]}$	(λtv)
$\mu_{[2]}$	$(\lambda tv)^2 + (\lambda tv)$
$\mu_{[3]}$	$(\lambda tv)^3 + 3(\lambda tv)^2 + (\lambda tv)$
$\mu_{[4]}$	$(\lambda tv)^4 + 6(\lambda tv)^3 + 7(\lambda tv)^2 + (\lambda tv)$
$\kappa_{[1]}$	(λtv)
$\kappa_{[2]}$	(λtv)
$\kappa_{[3]}$	(λtv)
$\kappa_{[4]}$	(λtv)
B_1	(λtv)
B_2	$[(\lambda tv)^2 + (\lambda tv)]/2!$
B_3	$[(\lambda tv)^3 + 3(\lambda tv)^2 + (\lambda tv)]/3!$
B_4	$[(\lambda tv)^4 + 6(\lambda tv)^3 + 7(\lambda tv)^2 + (\lambda tv)]/4!$

Example 2. The Neyman Type B Process: Let $\{N_t, t \geq 0\}$ be a homogeneous Poisson process with parameter $\lambda > 0$ and let X_i, $i = 1, 2, \ldots$ be binomial distributed with parameters m and p in (1), then $\{S_t, t \geq 0\}$ has a Neyman type B or Poisson-binomial process. First four moments and cumulants of the Neyman type B process are presented in Table 2.

Example 3. The Pólya-Aeppli Process: Let $\{N_t, t \geq 0\}$ be a homogeneous Poisson process with parameter $\lambda > 0$ and let X_i, $i = 1, 2, \ldots$ be geometric distributed random variables with parameter θ. Then, $\{S_t, t \geq 0\}$ has a Pólya-Aeppli or geometric Poisson process. First four moments and cumulants of the Pólya-Aeppli process are given in Table 3.

TABLE 2: First four moments and cumulants of the Neyman type B process.

μ'_1	(λtmp)
μ'_2	$(\lambda tmp)^2 + \lambda t[mp + m(m-1)p^2]$
μ'_3	$(\lambda tmp)^3 + 3(\lambda tmp)^2 + (\lambda tmp) + 6[\lambda tmp(m(m-1)p^2)] + [\lambda t(m(m-1)(m-2)p^3)]$
	$(\lambda tmp)^4 + 6(\lambda tmp)^2[\lambda t(mp + (m-1)p^2)] + 4(\lambda tmp)[\lambda t(mp + 3m(m-1)p^2$
	$+m(m-1)(m-2)p^3)]$
μ'_4	$(\lambda tv)^4 + 6(\lambda tv)^2[\lambda t(v+v^2)] + 4(\lambda tv)[\lambda t(v+3v^2+v^3)] + [\lambda t(v+7v^2+6v^3+v^4)]$
μ_1	$(2\lambda tmp)$
μ_2	$(2\lambda tmp)^2 + [\lambda t(mp + m(m-1)p^2)]$
μ_3	$(2\lambda tmp)^3 + 3(\lambda tmp)[\lambda t(mp + m(m-1)p^2)] + [\lambda t(mp + 3m(m-1)p^2$
	$+m(m-1)(m-2)p^3)]$
μ_4	$(2\lambda tmp)^4 + 6(2\lambda tmp)^2[\lambda t(mp + m(m-1)p^2)] + 4(2\lambda tmp)[\lambda t(mp + 3m(m-1)p^2$
	$+m(m-1)(m-2)p^3)] + 3[\lambda t(mp + m(m-1)p^2)]^2 + [\lambda t(mp + 7m(m-1)p^2$
	$+6m(m-1)(m-2)p^3 + m(m-1)(m-2)(m-3)p^4)]$
κ_1	(λtmp)
κ_2	$[\lambda t(mp + m(m-1)p^2)]$
κ_3	$[\lambda t(mp + 3m(m-1)p^2 + m(m-1)(m-2)p^3)]$
κ_4	$[\lambda t(mp + 7m(m-1)p^2 + 6m(m-1)(m-2)p^3 + m(m-1)(m-2)(m-3)(m-4)p^4)]$
$\mu_{[1]}$	(λtmp)
$\mu_{[2]}$	$(\lambda tmp)^2 + (\lambda tm(m-1)p^2)$
$\mu_{[3]}$	$(\lambda tmp)^3 + 3(\lambda tmp(mp + m(m-1)p^2))$
	$+(\lambda t(mp + 3m(m-1)p^2 + m(m-1)(m-2)p^3)$
$\mu_{[4]}$	$(\lambda tmp)^4 + 6(\lambda tmp)^3[\lambda t(mp + m(m-1)p^2)] + 4(\lambda tmp)[\lambda t(mp + 3m(m-1)p^2$
	$+m(m-1)(m-2)p^3)] + 3[\lambda t(mp + m(m-1)p^2)]^2 + [\lambda t(mp + 7m(m-1)p^2$
	$+6m(m-1)(m-2)p^3 + m(m-1)(m-2)(m-3)p^4)]$
$\kappa_{[1]}$	(λtmp)
$\kappa_{[2]}$	$[\lambda tm(m-1)p^2]$
$\kappa_{[3]}$	$[\lambda tm(m-1)(m-2)p^3]$
$\kappa_{[4]}$	$[\lambda tm(m-1)(m-2)(m-3)p^4]$
B_1	(λtmp)
B_2	$[(\lambda tmp)^2 + (\lambda tm(m-1)p^2)]/2!$
B_3	$[(\lambda tmp)^3 + 3(\lambda tmp)[\lambda tm(m-1)p^2] + [\lambda tm(m-1)(m-2)p^3]]/3!$
B_4	$[(\lambda tmp)^4 + 6(\lambda tmp)^2[\lambda tm(m-1)p^2] + 4(\lambda tmp)[\lambda tm(m-1)(m-2)p^3]$
	$+3[\lambda tm(m-1)p^2]^2 + [\lambda tm(m-1)(m-2)(m-3)p^4]]/4!$

Note that the random variable X has infinite values both the Neyman type A and the Pólya-Aeppli process. However, the moments and cumulants of these processes can be obtained using (13), (18) and (31). This is due to the probability $P(X_i = j)$ and $\lambda_j = \lambda p_j$ approach zero for $j \to \infty$.

Table 3: First four moments and cumulants of the Pólya-Aeppli process.

μ_1'	$[\lambda t(1-\theta)/\theta]$
μ_2'	$[\lambda t(1-\theta)/\theta]^2 + [\lambda t(1-\theta)(2-\theta)/\theta^2]$
μ_3'	$[\lambda t(1-\theta)/\theta]^3 + 3[\lambda t(1-\theta)/\theta][\lambda t(1-\theta)(2-\theta)/\theta^2] + [\lambda t(1-\theta)(6+\theta(\theta-6))/\theta^3]$
μ_4'	$[\lambda t(1-\theta)/\theta]^4 + 6[\lambda t(1-\theta)/\theta]^2[\lambda t(1-\theta)(2-\theta)/\theta^2] + 4[\lambda t(1-\theta)/\theta][\lambda t(1-\theta)$ $(6+\theta(\theta-6))/\theta^3] + 3[\lambda t(1-\theta)(2-\theta)/\theta^2]^2 + [\lambda t(2-\theta)(1-\theta)(12+(\theta-12)\theta)/\theta^4]$
μ_1	$[2\lambda t(1-\theta)/\theta]$
μ_2	$[2\lambda t(1-\theta)/\theta]^2 + [\lambda t(1-\theta)(2-\theta)/\theta^2)]$
μ_3	$[2\lambda t(1-\theta)/\theta]^3 + 3[2\lambda t(1-\theta)/\theta][\lambda t(1-\theta)(2-\theta)/\theta^2] + [\lambda t(1-\theta)(6+\theta(\theta-6))/\theta^3]$
μ_4	$[2\lambda t(1-\theta)/\theta]^4 + 6[2\lambda t(1-\theta)/\theta]^2[\lambda t(1-\theta)(2-\theta)/\theta^2] + 4[2\lambda t(1-\theta)/\theta][\lambda t(1-\theta)(6+$ $\theta(\theta-6))/\theta^3] + 3[\lambda t(1-\theta)(2-\theta)/\theta^2]^2 + [\lambda t(2-\theta)(1-\theta)(12+(\theta-12)\theta)/\theta^4]$
κ_1	$[\lambda t(1-\theta)/\theta]$
κ_2	$[\lambda t(2-\theta)(1-\theta)/\theta^2]$
κ_3	$[\lambda t(1-\theta)(6+\theta(\theta-6))/\theta^3]$
κ_4	$[\lambda t(1-\theta)(2-\theta)(12+\theta(\theta-12))/\theta^4]$
$\mu_{[1]}$	$[\lambda t(1-\theta)/\theta]$
$\mu_{[2]}$	$[\lambda t(1-\theta)/\theta]^2 + [\lambda t(2-\theta)(1-\theta)/\theta^2]$
$\mu_{[3]}$	$[\lambda t(1-\theta)/\theta]^3 + 3[\lambda t(2-\theta)(1-\theta)/\theta^2][\lambda t(1-\theta)/\theta] + [\lambda t(1-\theta)(6+(\theta-6)\theta)/\theta^3]$
$\mu_{[4]}$	$[\lambda t(1-\theta)/\theta]^4 + 6[\lambda t(1-\theta)/\theta]^2[\lambda t(2-\theta)(1-\theta)/\theta^2] + 4[\lambda t(1-\theta)/\theta)(\lambda t(1-\theta)$ $(6+(\theta-6)\theta)/\theta^3] + 3[\lambda t(2-\theta)(1-\theta)/\theta^2]^2 + [\lambda t(2-\theta)(1-\theta)(12+(\theta-12)\theta)/\theta^4]$
$\kappa_{[1]}$	$[\lambda t(1-\theta)/\theta]$
$\kappa_{[2]}$	$2[\lambda t(1-\theta)/\theta]^2$
$\kappa_{[3]}$	$6[\lambda t(1-\theta)/\theta]^3$
$\kappa_{[4]}$	$24[\lambda t(1-\theta)/\theta]^4$
B_1	$[\lambda t(1-\theta)/\theta]$
B_2	$[(\lambda t(1-\theta)/\theta)^2 + 2(\lambda t(1-\theta)/\theta)^2]/2!$
B_3	$[(\lambda t(1-\theta)/\theta)^3 + 6(\lambda t(1-\theta)/\theta)(\lambda t(1-\theta)/\theta)^2 + 6(\lambda t(1-\theta)/\theta)^3]/3!$
B_4	$[(\lambda t(1-\theta)/\theta)^4 + 12(\lambda t(1-\theta)/\theta)^2(\lambda t(1-\theta)/\theta)^2 + 24(\lambda t(1-\theta)/\theta)(\lambda t(1-\theta)/\theta)^3$ $+12(\lambda t(1-\theta)/\theta)^2 + 24(\lambda t(1-\theta)/\theta)^4]/4!$

3. The Bivariate Compound Poisson Process

In this section, we turn now to the consideration of factorial moments, cumulants, and the coefficient of correlation for the BCPP. Let $\{N_t, t \geq 0\}$ be a homogeneous Poisson process with parameter $\lambda > 0$ and let X_i, Y_i, $i = 1, 2, \ldots$, be mutually i.i.d. discrete random variables taking finite values with the probabilities $P(X_i = j) = p_j$, $j = 0, 1, \ldots m$ and $P(Y_i = k) = q_k$, $k = 0, 1, \ldots \ell$ in (2). We start by finding factorial moments $\mu_{[r,s]}$ for $r = 1, 2, \ldots$, $s = 1, 2, \ldots$ For this purpose, we first compute the joint probability generating function (pgf) $S_t^{(1)}$ and $S_t^{(2)}$ as follows

$$\begin{aligned} g_{S_t^{(1)},S_t^{(2)}}(u_1,u_2) &= \sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty} P\left(\sum_i^{N_t} X_i = s_1, \sum_i^{N_t} Y_i = s_2\right) u_1^{s_1}u_2^{s_2} \\ &= \sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty}\sum_{n=0}^{\infty} P\left(\sum_i^{n} X_i = s_1, \sum_i^{n} Y_i = s_2\right) P_{N_t}(n) u_1^{s_1}u_2^{s_2} \end{aligned}$$

where $P_{N_t}(n) = P(N_t = n)$, $n = 0, 1, \ldots$, is the probability function of the homogeneous Poisson process. Since X_i, Y_i, $i = 1, 2, \ldots$ are i.i.d. random variables, we get

$$\begin{aligned} g_{S_t^{(1)},S_t^{(2)}}(u_1,u_2) &= p_{N_t}(0)+p_{N_t}(1) \\ &\qquad \sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty} P(X_1=s_1)P(Y_1=s_2)u_1^{s_1}u_2^{s_2}+\cdots \\ &= p_{N_t}(0)+p_{N_t}(1)g_{X_1}(u_1)g_{Y_1}(u_2) \\ &\qquad +p_{N_t}(2)g_{X_1+X_2}(u_1)g_{Y_1+Y_2}(u_2)+\cdots \\ &= p_{N_t}(0)+p_{N_t}(1)g_X(u_1)g_Y(u_2) \\ &\qquad +p_{N_t}(2)[g_X(u_1)]^2[g_Y(u_2)]^2+\cdots \end{aligned} \tag{30}$$

where $g_X(u_1)$, $g_Y(u_2)$ are the common pgfs of X_i, Y_i, $i=1,2,\ldots$, respectively. Using (30), it is more convenient to deal with

$$\begin{aligned} g_{S_t^{(1)},S_t^{(2)}}(u_1,u_2) &= g_{N_t}[g_X(u_1)g_Y(u_2)] \\ &= \exp[\lambda t[g_X(u_1)g_Y(u_2)-1]] \\ &= \exp(-\lambda t)\exp[\lambda t(p_0q_0+p_0q_1u_2+\cdots+p_0q_lu_2^l+p_1q_0u_1 \\ &+p_1q_1u_1u_2+\cdots+p_1q_lu_1u_2^l+p_mq_0u_1^m+\cdots+p_mq_lu_1^mu_2^l)] \end{aligned} \tag{31}$$

The joint pgf in (31) can be differentiated any number of times with respect to r and s and evaluated at $(0,0)$ yielding

$$\mu_{[r,s]} = \left.\frac{\partial^{r+s} g_{S_t^{(1)},S_t^{(2)}}(u_1,u_2)}{\partial u_1^r \partial u_2^s}\right|_{u_1=u_2=1} \tag{32}$$

Differentiating (31) and substituting in (32), after some algebraic manipulations, the factorial moments of $S_t^{(1)}$ and $S_t^{(2)}$ are given by

$$\begin{aligned} \mu_{[1,1]} &= (\lambda t\xi_{[1]})(\lambda t\varsigma_{[1]})+(\lambda t\xi_{[1]\varsigma[1]}) \\ \mu_{[2,1]} &= (\lambda t\xi_{[1]})^2(\lambda t\varsigma_{[1]})+(\lambda t\xi_{[1]})(\lambda t\xi_{[1]\varsigma[1]})+(\lambda t\xi_{[2]})(\lambda t\varsigma_{[1]})+(\lambda t\xi_{[2]\varsigma[1]}) \\ \mu_{[2,2]} &= (\lambda t\xi_{[1]})^2(\lambda t\varsigma_{[1]})^2+(\lambda t\xi_{[1]})(\lambda t\varsigma_{[1]})(\lambda t\xi_{[1]\varsigma[1]})+(\lambda t\xi_{[2]})(\lambda t\varsigma_{[1]})^2 \\ &\quad +(\lambda t\varsigma_{[1]})(\lambda t\xi_{[2]\varsigma[1]})+(\lambda t\xi_{[1]})^2(\lambda t\varsigma_{[2]})+(\lambda t\xi_{[1]})(\lambda t\xi_{[1]\varsigma[2]}) \\ &\quad +(\lambda t\xi_{[1]\varsigma[1]})^2+(\lambda t\xi_{[2]})(\lambda t\varsigma_{[2]})+(\lambda t\xi_{[2]\varsigma[2]}) \\ \mu_{[2,3]} &= (\lambda t\xi_{[1]})^2(\lambda t\varsigma_{[1]})^3+(\lambda t\xi_{[1]})(\lambda t\varsigma_{[1]})^2(\lambda t\xi_{[1]\varsigma[1]})+(\lambda t\xi_{[2]})(\lambda t\varsigma_{[1]})^3 \\ &\quad +(\lambda t\varsigma_{[1]})^2(\lambda t\xi_{[2]\varsigma[1]})+(\lambda t\varsigma_{[1]})(\lambda t\xi_{[1]\varsigma[1]})^2+(\lambda t\xi_{[1]})(\lambda t\varsigma_{[1]})(\lambda t\xi_{[1]\varsigma[2]}) \\ &\quad +(\lambda t\xi_{[1]})^2(\lambda t\varsigma_{[2]})(\lambda t\varsigma_{[1]})+(\lambda t\xi_{[2]})(\lambda t\varsigma_{[2]})(\lambda t\varsigma_{[1]})+(\lambda t\varsigma_{[1]})(\lambda t\xi_{[2]\varsigma[2]}) \\ &\quad +(\lambda t\xi_{[2]})(\lambda t\xi_{[1]\varsigma[1]})(\lambda t\varsigma_{[1]})+(\lambda t\varsigma_{[2]})(\lambda t\xi_{[2]\varsigma[1]})+(\lambda t\xi_{[1]\varsigma[1]})(\lambda t\xi_{[1]\varsigma[2]}) \\ &\quad +(\lambda t\xi_{[1]\varsigma[3]})(\lambda t\varsigma_{[1]})+(\lambda t\varsigma_{[3]})(\lambda t\xi_{[1]})^2+(\lambda t\xi_{[2]})(\lambda t\varsigma_{[3]})+(\lambda t\xi_{[2]\varsigma[3]}) \end{aligned} \tag{33}$$

where $\xi_{[r]} = E[X(X-1)\ldots(X-(r-1))]$, $r=1,2,\ldots$, is the rth factorial moment of X_i, $i=1,2,\ldots$ and where $\varsigma_{[s]} = E[Y(Y-1)\ldots(Y-(s-1))]$, $s=1,2,\ldots$, is the sth factorial moment of Y_i, $i=1,2,\ldots$ in (2). Note that $\mu_{[r,s]}=\mu_{[s,r]}$ for $r=1,2,\ldots$, $s=1,2,\ldots$

Similar to univariate CPP, let X_i, Y_i, $i = 1, 2, \ldots$, have finite values with the probabilities $P(X_i = j) = p_j$, $j = 0, 1, \ldots m$ and $P(Y_i = k) = q_k$, $k = 0, 1, \ldots \ell$. (31) and (34) can be used when $P(X_i = j) = p_j$ and $P(Y_i = k) = q_k$ approach to zero for $j, k \to \infty$.

The joint cumulant generating function of $S_t^{(1)}$ and $S_t^{(2)}$ is given by

$$\begin{aligned}\kappa(u_1, u_2) = -\lambda t + \lambda t[(p_0 q_0 + \cdots + p_0 q_r \exp(u_2^r)) + (p_1 q_0 \exp(u_1) + \cdots \\ + p_1 q_r \exp(u_1) \exp(u_2^r)) + (p_m q_0 \exp(u_1^m) + \cdots \\ + p_m q_r \exp(u_1^m) \exp(u_2^r))] \quad (34)\end{aligned}$$

From (35) we have

$$\kappa_{r,s} = \lambda t(\xi_r \varsigma_s), \ r = 1, 2, \ldots, s = 1, 2, \ldots \quad (35)$$

where $\xi_r = E(X^r)$, $r = 1, 2, \ldots$, and $\varsigma_s = E(Y^s)$, $s = 1, 2, \ldots$, are expected values of X_i and Y_i, $i = 1, 2, \ldots$, respectively.

The covariance of $S_t^{(1)}$ and $S_t^{(2)}$ is obtained using (34)

$$\begin{aligned}Cov\left(S_t^{(1)}, S_t^{(2)}\right) &= E\left(S_t^{(1)} S_t^{(2)}\right) - E\left(S_t^{(1)}\right) E\left(S_t^{(2)}\right) \\ &= \lambda t(\lambda t + 1)\xi_1 \varsigma_1 - (\lambda t \xi_1)(\lambda t \varsigma_1) \\ &= \lambda t \xi_1 \varsigma_1\end{aligned} \quad (36)$$

Then, the coefficient of correlation for $S_t^{(1)}$ and $S_t^{(2)}$ is given by

$$\begin{aligned}\rho = Corr\left(S_t^{(1)}, S_t^{(2)}\right) &= \frac{Cov\left(S_t^{(1)}, S_t^{(2)}\right)}{\sqrt{Var\left(S_t^{(1)}\right) Var\left(S_t^{(2)}\right)}} = \frac{\lambda t \xi_1 \varsigma_1}{\sqrt{\left[\lambda t E\left(X^2\right)\right]\left[\lambda t E\left(Y^2\right)\right]}} \\ &= \frac{\xi_1 \varsigma_1}{\sqrt{\xi_2 \varsigma_2}}\end{aligned} \quad (37)$$

4. Numerical Examples

To illustrate the usage of the univariate CPP and BCPP, we present two data sets. The first data is taken from Meintanis (1997) and Özel & Inal (2010). It corresponds to the number of traffic accidents and fatalities recorded on Sundays of each month over the period 1997-2004 in the region of Groningen. In this study the same data is used to show applicability of the univariate CPP the with following random variables: N_t is the number of Sunday accidents which occurs in Groningen between years 1997-2004; X_i, $i = 1, 2, \ldots$, are the number of fatalities the ith type of accident; $S_t = \sum_i^{N_t} X_i$ is the total number of fatalities in the time interval $(0, t]$.

The homogeneous Poisson process provide an adequate fit to the number of Sunday accidents ($p - value < 0.01$, $\chi^2 = 2.94$) for $\lambda = 9.84$ (in month). The independency of X_i, $i = 1, 2, \ldots$, and $\{N_t, t \geq 0\}$ is shown using the $Spearman's\ \rho$ test ($Spearman's\ \rho = 0.084$; $p = 0.432$). Then, we have to decide the best distribution of X_i, $i = 1, 2, \ldots$ among the Poisson, binomial and geometric distributions for the number of fatalities. For this aim, a goodness of fit test can be performed to choose the correct distribution (Agresti 2002). However, one can take into consideration the number of values of X_i, $i = 1, 2, \ldots$ If X_i, $i = 1, 2, \ldots$ have finite values, the binomial distribution can be used. Similarly, geometric or Poisson distribution can be more suitable when X_i, $i = 1, 2, \ldots$ have infinite values. The goodness of fit test is applied to decide the best distribution. It is found seen that the Poisson distribution with parameter $\nu = 0.53$ ($p - value < 0.001$, $\chi^2 = 0.20$), the binomial distribution with parameters $m = 4$, $p = 0.12$ ($p - value < 0.01$, $\chi^2 = 1.52$) and the geometric distribution with parameter $\theta = 0.62$ ($p - value < 0.001$, $\chi^2 = 0.06$) fit the data. Then it can be said that $\{S_t, t \geq 0\}$ has a Pólya-Aeppli process. Note that the goodness-of-fit are applied sequentially without taking into account the dependence amongst these tests, which of course influences the overall size of the test, i.e., when we test all hypothesis each at level α, the computation of the overall level becomes more complicated.

The moments and cumulants for the Pólya-Aeppli process are computed from Table 3 for the parameters $\lambda = 9.84$; $\theta = 0.62$ and several values of t. The results are presented in Table 4. Then, the values of the skewness, kurtosis, $Cov(N_t, S_t)$ and $Corr(N_t, S_t)$ are computed for the Pólya-Aeppli process and the results are given in Table 5.

Table 4: The moments and cumulants of the Pólya-Aeppli process for the traffic accidents in Groningen.

t	μ'_1	μ'_2	μ'_3	μ'_4	$\mu_{[1]}$	$\mu_{[2]}$	$\mu_{[3]}$	$\mu_{[4]}$
0.5	3.02	15.81	109.04	922.47	3.02	12.79	67.66	396.66
1	6.03	49.80	613.16	7885.24	6.03	43.77	366.71	3352.23
2	12.06	172.34	3246.47	64054.05	12.06	160.28	2317.11	35671.34
3	18.09	367.62	9216.11	241607.41	18.09	349.53	7167.36	154264.24
4	24.12	635.66	19838.25	645397.44	24.12	611.53	16233.64	448189.00
	μ_1	μ_2	μ_3	μ_4	$\kappa_{[1]}$	$\kappa_{[2]}$	$\kappa_{[3]}$	$\kappa_{[4]}$
0.5	6.03	43.08	340.80	3513.44	3.02	18.19	164.52	1984.45
1	12.06	164.95	2458.88	41475.28	6.03	72.75	1316.17	31751.14
2	24.12	620.87	16855.10	487496.27	12.06	290.98	10529.37	508018.17
3	36.19	1367.78	53718.02	2201987.32	18.09	654.71	35536.61	2571841.99
4	48.25	2405.66	123577.01	6556890.97	24.12	1163.92	84234.93	8128290.74
	κ_1	κ_2	κ_3	κ_4	B_1	B_2	B_3	B_4
0.5	3.02	6.71	20.90	86.33	3.02	6.39	11.28	16.53
1	6.03	19.45	41.80	172.67	6.03	21.88	61.12	139.68
2	12.06	38.91	83.61	345.33	12.06	80.14	386.18	1486.31
3	18.09	58.36	125.41	518.00	18.09	174.77	1194.56	6427.68
4	24.12	77.82	167.21	690.66	24.12	305.77	2705.61	18674.54

Table 5: The skewness, kurtosis, covariance, and the coefficient of correlation of the Pólya-Aeppli process for the traffic accidents in Groningen.

t	$\sqrt{\beta_1}$	β_2	$Cov(N_t, S_t)$	$Corr(N_t, S_t)$
0.5	0.002131	-1.107261	3.015484	0.52475
1	0.000274	-1.475558	6.030968	
2	0.000035	-1.735354	12.061935	
3	0.000010	-1.822979	18.092903	
4	0.000004	-1.867004	24.123871	

A second data set comes from earthquakes in Turkey which is given by Özel (2011*a*) and Özel (2011*b*). The mainshocks with surface wave magnitudes $M_s \geq 5.0$ that occurred in Turkey between 1900 and 2009, their foreshock and aftershock sequences are considered. For the construction of a model to explain the total number of foreshocks and aftershocks with the BCPP in (2), the following random variables are defined: N_t is the number of mainshocks that occurred in Turkey between 1903 and 2009; X_i, $i = 1, 2, \ldots$ are the number of foreshocks of ith mainshock; Y_i, $i = 1, 2, \ldots$ are the number of aftershocks of the ith mainshock; and $\left(S_t^{(1)} = \sum_i^{N_t} X_i, S_t^{(2)} = \sum_i^{N_t} Y_i\right)$ is the total number of foreshocks and aftershocks for the mainshocks. The goodness of fit test is performed to compare the observed frequency distribution to the theoretical Poisson distribution. Chi-square value ($\chi^2 = 0.051$ with $df = 9$, $p - value = 0.525$) indicates that $\{N_t, t \geq 0\}$ fits the Poisson process with parameter $\lambda = 1.037$ (in years) at the level of 0.05. Spearman's ρ test verifies the absence of correlation between N_t and X_i, $i = 1, 2, \ldots$ (Spearman's $\rho = 0.071$; $p = 0.412$). No correlation is also found between N_t and Y_i, $i = 1, 2, \ldots$ (Spearman's $\rho = 0.034$; $p = 0.589$). Similarly, it is shown that there is no statistically significant dependence between X_i and Y_i, $i = 1, 2, \ldots$ (Spearman's $\rho = 0.048$; $p = 0.493$). As discussed by Özel (2011*b*), if the occurrence of foreshock sequences is assumed to be independent of the occurrence of mainshocks, then the distribution of foreshocks can be treated as a binomial distribution. The goodness-of-fit test for the binomial distribution provided an adequate fit with a p-value of 0.999 and chi-squared value $\chi^2 = 0.003$ with 34 degrees of freedom. This means that the binomial distribution with parameters ($m = 35$, $p = 0.15$) fits the probability function of X_i, $i = 1, 2, \ldots$ It is pointed out that the number of aftershocks of a mainshock has a geometric distribution (Christophersen & Smith 2000). After obtaining the frequency distribution of aftershocks and the goodness-of-fit test ($\chi^2 = 1.587$ with $df = 35$), it is seen that Y_i, $i = 1, 2, \ldots$ have a geometric distribution with parameter $\theta = 0.175$. Then, we can write $\left(S_t^{(1)} = \sum_i^{N_t} X_i, S_t^{(2)} = \sum_i^{N_t} Y_i\right)$ and suggest that $(S_t^{(1)}, S_t^{(2)})$ has a BCPP. So, the joint factorial moments and cumulants are calculated from (33) and (35) for the parameters $\lambda = 1.037$; $\theta = 0.175$; ($m = 35$, $p = 0.15$) and several values of t. Then, $Cov\left(S_t^{(1)}, S_t^{(2)}\right)$ and $Corr\left(S_t^{(1)}, S_t^{(2)}\right)$ are computed from (34) and (36) and the results are presented in Table 6.

TABLE 6: The moments, cumulants, covariance, and coefficient of correlation of the BCPP for the earthquakes in Turkey.

t	$\mu_{[1,1]}$	$\mu_{[2,1]}$	$\mu_{[2,2]}$	$\mu_{[2,3]}$	$\kappa_{1,1}$	$\kappa_{2,1}$	$\kappa_{2,2}$	$\kappa_{2,3}$	$Cov(N_t, S_t)$	$Corr(N_t, S_t)$
0.5	19.49	152.43	546.09	26543.35	12.83	78.28	816.35	3422.87	12.83	0.6237
1	52.28	551.27	713.82	128434.33	25.67	156.56	1632.71	6845.74	25.67	
2	157.79	2522.87	11344.95	1118387.43	51.33	313.12	3265.42	13691.49	51.33	
2.5	230.51	4313.00	27684.65	2578433.75	64.16	391.40	4081.77	17114.36	64.16	
3	316.54	6784.23	57395.00	5330414.38	77.00	469.68	4898.13	20537.23	77.00	

5. Conclusion

In this paper, the moments, cumulants, skewness, kurtosis and covariance of the univariate CPP are derived. Some special cases of the univariate CPP are provided and a numerical example based on the traffic accidents in Groningen is given. Then, BCPP is defined and some important probabilistic characteristics such as moments, cumulants, covariances, and the coefficient of correlation for the BCPP are obtained.

Earthquake is an unavoidable natural disaster for Turkey. Application to the earthquake data in Turkey is presented to illustrate the usage of the BCPP and its properties. Earthquakes could be regarded as discrete events, representing some real but not well-known tectonic process. Following that scheme and keeping in mind the highly random characteristics of all earthquake parameters, it is quite natural to consider a sequence of earthquakes as a stochastic process. The stochastic modeling of the earthquake occurrence has proved very useful in earthquake prediction studies, in understanding the nature of the earthquake phenomena, and in assessing seismicity and seismic hazard. Existing approaches in the research of seismic hazard assessment are generally based on the homogeneous Poisson process. However, new studies have been done using CPP and BCPP and give more information than homogeneous Poisson process. For this reason, the factorial moments and cumulants of BCPP, which are obtained in this study, can be a good tool to understand earthquake behaviour.

[Recibido: marzo de 2012 — Aceptado: marzo de 2013]

References

Agresti, A. (2002), *Categorical Data Analysis*, John Wiley & Sons, New Jersey.

Ata, N. & Özel, G. (2012), 'Survival functions for the frailty models based on the discrete compound Poisson process', *Journal of Statistical Computation and Simulation (Online Published)* . DOI: 10.1080/00949655.2012.679943.

Chen, C. W., Randolph, P. & Tian-Shy, L. (2005), 'Using CUSUM control schemes for monitoring quality levels in compound Poisson production environments: the geometric Poisson process', *Quality Engineering* **17**, 207–217.

Christophersen, A. & Smith, E. G. C. (2000), A global model for aftershock behaviour, Proceedings of the 12th World Conference on Earthquake Engineering. Paper 0379, Auckland, New Zealand.

Getis, A. (1974), Representation of spatial point processes by Pólya methods, Proceedings of the 1972 meeting of the IGU Commission on Quantitative Geography. Montreal, Canada.

Gudowska-Nowak, E., Lee, R., Nasonova, E., Ritter, S. & Scholz, M. (2007), 'Effect of let and track structure on the statistical distribution of chromosome aberrations', *Advances in Space Research* **39**, 1070–ĂŞ1075.

Hesselager, O. (1996), 'Recursions for certain bivariate counting distributions and their compound distributions', *ASTIN Bulletin* **26**, 35–52.

Kocherlakota, S. & Kocherlakota, K. (1997), *Bivariate Discrete Distributions*, Wiley, NewYork.

Meintanis, S. G. (1997), 'A new goodness of fit test for certain bivariate distributions applicable to traffic accidents', *Statistical Methodology* **4**, 22–34.

Neyman, J. (1939), 'On a new class of contagious distributions applicable in entomology and bacteriology', *Annals of Mathematical Statistics* **10**, 35–57.

Özel, G. (2011*a*), 'A bivariate compound Poisson model for the occurrence of foreshock and aftershock sequences in Turkey', *Environmetrics* **22**(7), 847–856.

Özel, G. (2011*b*), 'On certain properties of a class of bivariate compound Poisson distributions and an application to earthquake data', *Revista Colombiana de Estadistica* **34**(3), 545–566.

Özel, G. & Inal, C. (2008), 'The probability function of the compound Poisson process and an application to aftershock sequences', *Environmetrics* **19**, 79–85.

Özel, G. & Inal, C. (2010), 'The probability function of a geometric Poisson distribution', *Journal of Statistical Computation and Simulation* **80**, 479–487.

Özel, G. & Inal, C. (2012), 'On the probability function of the first exit time for generalized Poisson processes', *Pakistan Journal of Statistics* **28**(1), 27–40.

Robin, S. (2002), 'A compound Poisson model for word occurrences in DNA sequences', *Applied Statistics* **51**, 437–451.

Rosychuk, R. J., Huston, C. & Prasad, N. G. N. (2006), 'Spatial event cluster detection using a compound Poisson distribution', *Biometrics* **62**, 465–470.

Sundt, B. (1992), 'On some extensions of Panjer's class of counting distributions', *ASTIN Bulletin* **22**, 61–80.

Wienke, A. (2011), *Frailty Model in Survival Analysis*, Chapman and Hall.

Wienke, A., Ripatti, S., Palmgren, J. & Yashin, A. (2010), 'A bivariate survival model with compound Poisson frailty', *Statistics in Medicine* **29**(2), 275–283.

Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 79 a 93

Comparing TL-Moments, L-Moments and Conventional Moments of Dagum Distribution by Simulated data

Comparación de momentos TL, momentos L y momentos convencionales de la distribución Dagum mediante datos simulados

Mirza Naveed Shahzad[1,a], Zahid Asghar[2,b]

[1]Department of Statistics, University of Gujrat, Gujrat, Pakistan
[2]Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan

Abstract

Modeling income, wage, wealth, expenditure and various other social variables have always been an issue of great concern. The Dagum distribution is considered quite handy to model such type of variables. Our focus in this study is to derive the L-moments and TL-moments of this distribution in closed form. Using L & TL-moments estimators we estimate the scale parameter which represents the inequality of the income distribution from the mean income. Comparing L-moments, TL-moments and conventional moments, we observe that the TL-moment estimator has lessbias and root mean square errors than those of L and conventional estimators considered in this study. We also find that the TL-moments have smaller root mean square errors for the coefficients of variation, skewness and kurtosis. These results hold for all sample sizes we have considered in our Monte Carlo simulation study.

***Key words*:** Dagum distribution, L-moments, Method of moments, Parameter estimation, TL-moments.

Resumen

La modelación de ingresos, salarios, riqueza, gastos y muchas otras variables de tipo social han sido siempre un tema de gran interés. La distribución Dagum es considerada para modelar este tipo de variables. Nos centraremos en este artículo en la derivación de los momentos L y los momentos TL de esta distribución de manera cerrada. Mediante el uso de los estimadores de momentos L y TL, estimamos el parámetro de escala que representa la desigualdad de la distribución de ingresos a partir de la media. Comparando los

[a]Professor. E-mail: nvd.shzd@uog.edu.pk
[b]Doctor. E-mail: g.zahid@gmail.com

momentos L, los momentos TL y los momentos convencionales, concluimos que los momentos TL tienen menor sesgo y errores cuadráticos medios. También concluimos que los momentos TL tiene la menor error cuadrático medio para los coeficientes de variación, sesgo y curtosis. Estas conclusiones son igualmente aplicables para todos los tamaños de muestras considerados en nuestro estudio de simulación de Monte Carlo.

***Palabras clave*:** distribución Dagum, estimacón de parámetros, momentos TL, momentos L, método de momentos.

1. Introduction

Dagum (1977*a*, 1977*b*) studied the income, wage and wealth distribution using the Dagum Distributions. Dagum Distribution (DD) belongs to the family of Beta distributions. Kleiber (1996) showed that this family models income distribution at the univariate level. Dagum (1690) considered DD to model income data of several countries and found that it provides superior fit over the whole range of data. Perez & Alaiz (2011) studied personal income data of Spain using DD and found this model to be adequate. Quintano & Dagostino (2006) analyzed the single-person household income distribution for four European countries, and concluded that DD provide a better fit for all four countries. Bandourian, McDonald & Turley (2003) showed that DD provide the best fit in the case of two or three parameter distributions for data from 23 countries. Various other studies also support the use of DD as model for income data.

Identifying the pattern of income distribution is very important because the trend provides a guide for the assessment of living standards and level of income inequality in the population of a country. Recently, there has been an increasing interest in the exploration of parametric models for income distribution and DD has proved to be quite useful in modeling such data. But this distribution has yet not been studied and estimated assuming the L-moment and TL moment. It is has been demonstrated that L & TL- moments provide accurate fit and more exact parameter estimation compared to the other techniques. The method of L-moments and TL-moments were introduced Hosking (1990) and Elamir & Seheult (2003), respectively. TL-moments have some merits over L-moments because the former can be calculated, even if mean data does not exist.

This paper seeks to derive the first four L & TL-moments of DD and coefficient of variation (CV), coefficient of skewness (CS) and coefficient of kurtosis (CK) estimators. To our knowledge, these moments for DD has not been derived and evaluated. We estimate the scale parameter of DD assuming L & TL-moments estimators and compare these with conventional moments. To achieve this objective, we measure the biasedness and RMSEs to recommend an efficient method of estimation. We also estimate the CV, CS & CK with the central, L & TL-moments estimators. e set up a Monte Carlo simulation study assuming different sample sizes and parametric values.

TL-moment estimators (TLMEs), L-moments estimators (LMEs) are derived and compared with the conventional method of moment estimators (MMEs) for

DD. The rest of the study is organized as follows: Section two is about the introduction of the population and sample TL-moments and L-moments. In Section 3, probability density function (pdf), distribution function, conventional moments and some other details of DD are presented. The derivations of the first four L & TL-moments is given in Section 4 and the coefficients are also presented. In Section 5, we setup the Monte Carlo simulation study to compare the properties of the TLMEs, LMEs and MMEs of DD. Finally we conclude our study in the final section.

2. L-Moment and TL-Moments

Hosking (1990) introduced L-moments and showed that these moments provide superior fit, parameter estimation, hypothesis testing and empirical description of data. Bílková (2012) used the L-moment of lognormal distribution to model the income distribution data of the Czech Republic in 1992–2007 and obtained consistent results as compared to the other methods of estimation. Due to the advantages of L-moments over the convention moments, many distributions are analyzed by these moments. Linear combinations of the ordered data values are used to compute L-moments. Furthermore, these moments are less sensitive in the case of outlier (Vogel & Fennessey 1993). Hosking (1990) defined the rth population L-moments (λ_r)as the linear combinations of probability weighted moments of an ordered sample data($\mathrm{Y_{1:n}} \leq \mathrm{Y_{2:n}} \leq ... \leq \mathrm{Y_{n:n}}$), that is

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} E\left(Y_{r-k:r}\right), \qquad r = 1, 2, 3, \ldots \tag{1}$$

For the real-numbered random variable $Y$with cumulative distribution function (cdf) $F(y)$; let $y(F)$ denote the quantile function of the ordered statistics of the sample of size n; then $E(Y_{j:r})$is given by

$$\begin{aligned} E(Y_{j:r}) &= \frac{r!}{(1-r)!(r-j)!}\int_0^1 y(F)^{j-1}(1-F)^{r-j}dF \\ &= \frac{r!}{(1-r)!(r-j)!}\int_{-\infty}^{\infty} yf(y)\left[F(y)\right]^{j-1}\left[1-F(y)\right]^{r-j}dy;\ j = 1, 2, \ldots, n \end{aligned} \tag{2}$$

The first and second L-moments (λ_1, λ_2)are equal to the measure of location and dispersion respectively. The ratio of the third L-moment (λ_3) to the second L-moment and ratio of the fourth L-moments (λ_4)to the second L-moment are the measure of skewness $\tau_{cs}^L = \lambda_3/\lambda_2$ and kurtosis, $\tau_{ck}^L = \lambda_4/\lambda_2$ respectively. The sample L-moments (l_r)are $l_1 = d_0$, $l_2 = 2d_1 - d_0$, $l_3 = 6d_2 - 6d_1 + d_0$and$l_4 = 20d_3 - 30d_2 + 12d_1 - d_0$and the sample L-skewness and L-kurtosis are $t_{cs}^L = l_3/l_2$, $t_{ck}^L = l_4/l_2$respectively. These ratios are less biased than for the conventional moments in estimation. The above mentioned d_r $(r = 1, 2, 3, 4)$ are given by

$$d_r = \frac{1}{n}\sum_{j=r+1}^{n}\frac{(j-1)(j-2)\cdots(j-r)}{(n-1)(n-2)\cdots(n-r)}y_{j:n} \tag{3}$$

where the size of data is n.

Elamir & Seheult (2003) introduced the TL-moments. TL-moments does not have the assumption of the existence of the mean. The TL-moments for the Cauchy distribution are derived by Shabri, Ahmad & Zakaria (2011), even though the mean of this distribution does not exist. According to Elamir & Seheult (2003), the r^{th} TL-moments and sample TL-moment are given by

$$\lambda_r^{(t)} = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \begin{pmatrix} r-1 \\ k \end{pmatrix} E\left(Y_{r+t-k:r+2t}\right), \qquad \begin{matrix} r = 1, 2, 3, \ldots \\ t = 1, 2, 3, \ldots \end{matrix} \tag{4}$$

and

$$l_r^{(t)} = \frac{1}{r}\sum_{j=t+1}^{n-t}\left[\sum_{k=0}^{r-1}(-1)^k C\right] Y_{j:n} \tag{5}$$

where $C = \begin{pmatrix} r-1 \\ k \end{pmatrix}\begin{pmatrix} j-1 \\ r+t-1-k \end{pmatrix}\begin{pmatrix} n-j \\ t+k \end{pmatrix}\Big/\begin{pmatrix} n \\ r+2t \end{pmatrix}$ respectively. The sample TL-skewness and TL-kurtosis are defined as $t_{cs}^{(t)} = l_3^{(t)}\Big/l_2^{(t)}$ and $t_{ck}^{(t)} = l_4^{(t)}\Big/l_2^{(t)}$, respectively.

3. Dagum Distribution

The Dagum distribution is a special case of the Generalized Beta type-II distribution ($DD(a,b,p) = GB2(a,b,p,1)$)as mentioned by Kleiber (1996). It is often used to model wage, wealth and income data. It was introduced by Dagum (1977*b*). The pdf of the distribution is given by

$$f(y) = \frac{ap\,(y)^{ap-1}}{b^{ap}\left[1+(y/b)^a\right]^{p+1}}, \tag{6}$$

where $p > 0$ and $a > 0$ are the shape parameters and $b > 0$ is the scale parameter. The cdf and r^{th} moment about zero are given by $F(y) = \left[(y/b)^{-a}+1\right]^{-p}$ and $E(Y^r) = b^r\Gamma(p+r/a)\Gamma(1-r/a)/\Gamma p$ respectively. The three-parameter DD provides a flexible distribution (Dagum & Lemmi 1988), and has better performace than other commonly used models (Kleiber 1996).

To evaluate the best method of estimation among the considered methods, we used the criteria of bias and RMSE. Bias is the expected difference between estimated and true value of the parameter. According to Daud, Kassim, Desa & Nguyen (2002) the $RMSE = \left[\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)}{n-m}\right]^{1/2}$, where y_i is actual observations, $\hat{y}_i$ is the estimated value obtained from the fitted distribution, $n-m$ is the difference between the number of observations in the sample and the number of parameters being estimated.

4. L-Moments and TL-Moments for the DD

As mentioned earlier, to best of our knowledge, there is no derivation of the L & TL-moments for DD in the literature. In this section, we derive L & TL-moments of DD using a general rule. The derivation of the L & TL-moments is given in the following Subsections 4.1-4.3.

4.1. L-moments of DD

Let $Y_{1:n} \leq Y_{2:n} \leq Y_{3:n} \leq \cdots \leq Y_{n:n}$ denote the order statistics from DD. The expected value of the r^{th} order statistics $Y_{r:n}$ is

$$E(Y_{r:n}) = \frac{n!ap}{(r-1)!(n-r)} \int (y_{r:n}/b)^{ab} \left[(y_{r:n}/b)^{-a} + 1\right]^{-pr-1} \times \left[1 - \left((y_{r:n}/b)^{-a} + 1\right)^{-p}\right]^{n-r} dy_{r:n} \tag{7}$$

where $y(F) = b\left(F^{-1/p} - 1\right)^{-1/a}$ is the quantile function of the DD. Now using the general form of L-moments, we have the first four L-moments for DD as follows

$$\lambda_1 = E(Y) = b\Gamma(1-\alpha)G_1 \tag{8}$$

$$\lambda_2 = b\Gamma(1-\alpha)\left(-G_1 + G_2\right) \tag{9}$$

$$\lambda_3 = b\Gamma(1-\alpha)\left(G_1 - 3G_2 + 2G_3\right) \tag{10}$$

$$\lambda_4 = b\Gamma(1-\alpha)\left(-G_1 + 6G_2 - 10G_3 + 5G_4\right) \tag{11}$$

where $G_i = \Gamma(ip+\alpha)/\Gamma ip$; $i = 1, 2, 3, 4, 5$.

Equating the population L-moments with sample L-moments and after simplification, we get the following results that could be used for the parameter estimation of DD

$$l_1 = bp \times Beta\left(1-\alpha, p+\alpha\right) \tag{12}$$

$$l_2 = -l_1 + 2bp \times Beta\left(1-\alpha, 2p+\alpha\right) \tag{13}$$

$$l_3 = -2l_1 - 3l_2 + 6bp \times Beta\left(1-\alpha, 3p+\alpha\right) \tag{14}$$

$$l_4 = -15l_1 - 24l_2 - 10l_3 + 20bp \times Beta\left(1-\alpha, 4p+\alpha\right) \tag{15}$$

where 'Beta' is the beta function ($Beta(\theta_1, \theta_2) = \Gamma\theta_1\Gamma\theta_2/\Gamma(\theta_1+\theta_2)$).

4.2. TL-moments of DD

L-moments are the foundation of TL-moments. TL-moments are more robust than L-moments (Elamir and Seheult, 2003) because they trim the extreme values on the data. The close from of the first four TL-moments are

$$\lambda_1^{(t)} = b\Gamma(1-\alpha)\left(3G_2 - 2G_3\right) \tag{16}$$

$$\lambda_2^{(t)} = b\Gamma(1-\alpha)\left(3G_2 + 6G_3 - 3G_4\right) \tag{17}$$

$$\lambda_3^{(t)} = (10b\Gamma(1-\alpha)/3)\left(G_2 - 4G_3 + 5G_4 - 2G_5\right) \tag{18}$$

$$\lambda_4^{(t)} = 15b\Gamma(1-\alpha)\left(-G_2/4 + 5G_3/3 - 15G_4/4 + 7G_5/2 - 7G_6/6\right) \tag{19}$$

4.3. L & TL Coefficient of Variation, Skewness and Kurtosis

The population coefficient of variation $\left(\tau_{cv}^L\right)$lies between 0 and 1, τ_{cs}^L also has the range 0 and 1, andτ_{ck}^Lmeasure the peakness of any distribution, lies within the range of $\left(5\left(\tau_{cs}^L\right)^2 - 1\right)/4 \leq \tau_{ck}^L < 1$ according to Hosking (1990) The $\tau_{cv}^L, \tau_{cs}^L$and $\tau_{ck}^L$of DD are expressed as follows:

$$\tau_{cv}^L = \frac{G_2}{G_1} - 1 \tag{20}$$

$$\tau_{cs}^L = \frac{G_1 - 3G_2 + 2G_3}{-G_1 + G_2} \tag{21}$$

$$\tau_{ck}^L = \frac{-G_1 + 6G_2 - 10G_3 + 5G_4}{-G_1 + G_2} \tag{22}$$

The population TL-moments CV, CS and CK are represented with the notation $\tau_{cv}^{(t)}, \tau_{cs}^{(t)}$and $\tau_{ck}^{(t)}$of DD and expressed as follows:

$$\tau_{cv}^{(t)} = \frac{3G_2 + 6G_3 - 3G_4}{3G_2 - 2G_3} \tag{23}$$

$$\tau_{cs}^{(t)} = \frac{10\left(G_2 - 4G_3 + 5G_4 - 2G_5\right)}{3\left(3G_2 + 6G_3 - 3G_4\right)} \tag{24}$$

$$\tau_{ck}^{(t)} = \frac{5\left(-G_2 + 5G_3 - 15G_4 + 7G_5 - 7G_6\right)}{\left(G_2 + 2G_3 - G_4\right)} \tag{25}$$

5. Monte Carlo Simulation Study

In this section, we use Monte Carlo simulated experiments to compare the three methods of moment estimators, conventional, L & TL-moments estimators of DD. This comparison is based on a measure of biasedness, root mean square estimators (RMSEs), sample CV, sample CS and sample CK. We use MATLAB-7 software to conduct our experiment. We perform our experiments for various sample sizes (15, 30, 50, 100, 500 and 1,000) as well as for different values of parameters. We have repeated each of our experiment 10,000 times. We use same parametric values for DD as were used by Ye, Oluyede & Pararai (2012).

In each case, for the estimation of b (scale parameter), we equate the sample moments to the corresponding population moments, and finally get the biasness and RMSEs of the b assuming the MMEs, LMEs & TLMEs of DD. Graphical shapes of the distribution on the bases of these parameters are given in Figure 1.

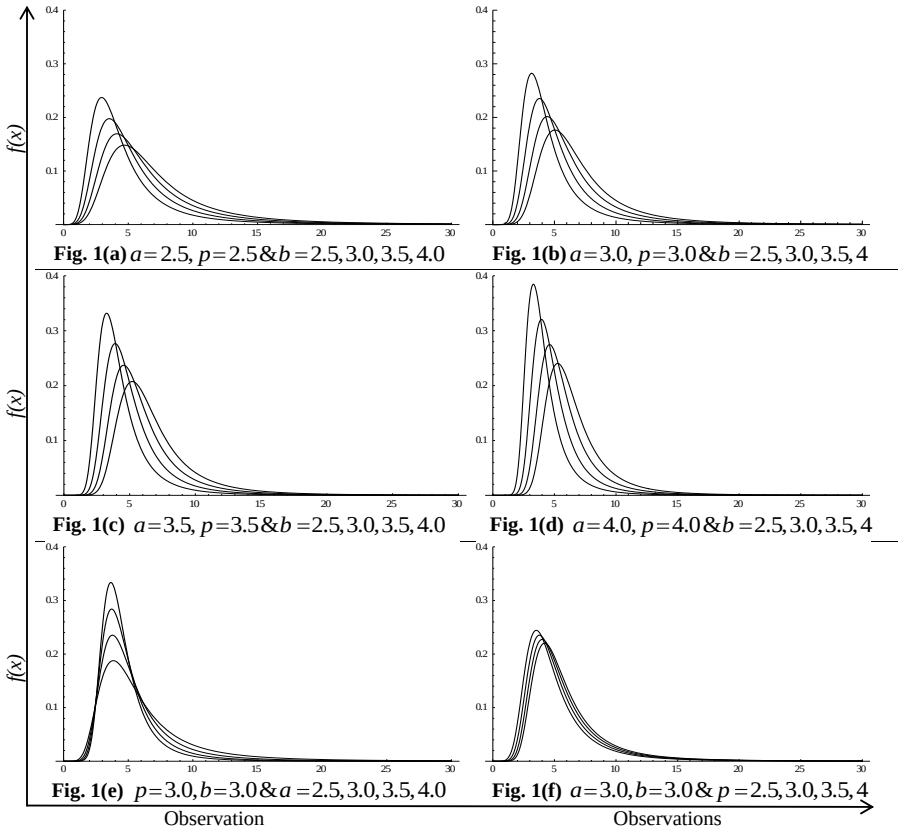


FIGURE 1: Dagum Distribution trend with different values of the parameters.

The results are presented in the Table 1 - 5. We find that the method of moments estimator (MME) gave biased results of the scale parameter with higher RMSEs. The L-moments estimator (LME) gave biased results with lower RMSEs than MME, and the TLMEs has a smaller bias with respect to the scale parameter and with lowest RMSEs. TLME results are very close to the true parametric values. So the TL-moments provide an unbiased estimator. According to the RMSEs, we can define the relation of these three moments estimators asTLME $<$ LME $<$ MME. These results hold for all the sample sizes we have considered. Therefore, the TL-moments provide precise and accurate estimates of the scale parameters of DD. If we do not want to trim the extreme values then L-moments provide better results.

The mean, L-moment standard deviation (LSD), $\tau_{\text{cv}}^{\text{L}}$, $\tau_{\text{cs}}^{\text{L}}$and $\tau_{\text{ck}}^{\text{L}}$ are computed using equations (21), (22) and (23) respectively. TL-moment standard deviation (TLSD),$\tau_{\text{cv}}^{\text{TL}}$, $\tau_{\text{cs}}^{\text{TL}}$and $\tau_{\text{ck}}^{\text{TL}}$ are also computed using equations (24), (25) and (26) respectively. These results are presented in Table A.6, assuming the same parametric values as those used for the scale parameter estimation. We observe that L & TL-moments coefficients are in the defined range and TL-moments coefficients

have a relatively smaller value than the conventional and L-moments ones. We also observe that the shape parameters (a and p)make some effect on the coefficients value but for different values of scale parameters (b) coefficients remain constant. Finally, we sum up all the above description in the favour of TL-moments for DD.

6. Conclusion

We have derived L and TL-moments for DD, compared parameter estimates and descriptive statistics with the conventional methods of moment estimates, assuming different parametric values for small to large samples. For parameter estimation, we found TL-moments provide unbiased and efficient results compared to the remaining moments because it is more robust against outliers. L-moments also provide more or less unbiased results and is more efficient than conventional moments. In distribution fitting, according to the location, scale, RMSE, skewness and kurtosis, TL-moments are better for DD parameter estimation. We find that TL moments estimators are the best, and L-moments are better than conventional moments for untrimmed data. These results hold for all sample sizes and parametric values which we have considered in our study.

TABLE 1: Biases and RMSEs of the parameter estimations for different types of estimators assuming DD for b when $b = 2.5$

Parameters			**n = 50**			**n = 100**		
a	***p***		**MME**	**LME**	**TLME**	**MME**	**LME**	**TLME**
2.5	2.5	Bias	-0.0645	-0.0002	-0.0016	-0.0370	-0.0004	0.0003
		RMSEs	0.4567	0.3497	0.3442	0.3268	0.2487	0.2430
	3.5	Bias	-0.0494	0.0016	0.0032	-0.0318	-0.0043	-0.0045
		RMSEs	0.3571	0.3069	0.3264	0.2543	0.2141	0.2255
	5.0	Bias	-0.0034	-0.0222	0.0034	-0.0222	0.0004	0.001
		RMSEs	0.3182	0.2893	0.3191	0.2222	0.1020	0.2204
3.5	2.5	Bias	-0.0535	-0.0050	-0.0035	-0.0226	0.0021	0.0017
		RMSEs	0.3422	0.2991	0.3215	0.2460	0.2123	0.2123
	3.5	Bias	-0.0427	-0.0014	-0.0036	-0.0215	-0.0002	-0.0010
		RMSEs	0.2933	0.2741	0.3066	0.2086	0.1943	0.2164
	5.0	Bias	-0.0417	-0.0011	0.0011	-0.0196	0.0002	0.0012
		RMSEs	0.2685	0.2634	0.3041	0.1899	0.1846	0.2118
5.0	2.5	Bias	-0.0444	-0.0018	-0.0021	-0.0209	0.0002	-0.0002
		RMSEs	0.2971	0.2815	0.3109	0.2067	0.1944	0.2145
	3.5	Bias	-0.0422	-0.0033	-0.0045	-0.0208	-0.0010	-0.0006
		RMSEs	0.2680	0.2654	0.3074	0.1885	0.1860	0.2128
	5.0	Bias	-0.0325	0.0052	0.0033	-0.0213	-0.0022	-0.0020
		RMSEs	0.2566	0.2594	0.3029	0.1810	0.1816	0.2103
Parameters			**n = 500**			**n = 1,000**		
a	***p***		**MME**	**LME**	**TLME**	**MME**	**LME**	**TLME**
2.5	2.5	Bias	-0.0051	0.0009	0.00003	-0.0039	-0.0002	-0.0004
		RMSEs	0.1612	0.1107	0.1069	0.0754	0.0783	0.0754
	3.5	Bias	-0.0058	-0.0003	0.0001	-0.0040	-0.0011	-0.0010
		RMSEs	0.1202	0.0983	0.1025	0.0843	0.0687	0.0717
	5.0	Bias	-0.0045	-.00001	-.00005	-0.0027	-0.0003	-0.0002
		RMSEs	0.1023	0.0908	0.0978	0.0712	0.0631	0.0684
3.5	2.5	Bias	-0.0056	-0.0004	0.00001	-0.0027	-0.0002	0.0001
		RMSEs	0.1138	0.0953	0.1004	0.0815	0.0670	0.0699
	3.5	Bias	-0.0043	0.0002	0.0009	-0.0024	-0.0005	-0.0010
		RMSEs	0.0958	0.0876	0.0958	0.0668	0.0612	0.0673
	5.0	Bias	-0.0044	-0.0006	-0.0009	-0.0026	-0.0007	-0.0008
		RMSEs	0.0836	0.0813	0.0939	0.0602	0.0582	0.0665
5.0	2.5	Bias	-0.0050	-0.0006	-0.0001	-0.0038	-0.0017	-0.0018
		RMSEs	0.0945	0.088	0.0974	0.0665	0.0620	0.0679
	3.5	Bias	-0.0043	-0.0003	-0.0002	-0.0019	0.0001	0.0007
		RMSEs	0.0844	0.0830	0.0943	0.0599	0.0588	0.0667
	5.0	Bias	-0.0040	-0.0002	0.0002	-0.0029	-0.0011	-0.0014
		RMSEs	0.0801	0.0807	0.0934	0.0568	0.0573	0.0665

TABLE 2: Biases and RMSEs of the parameter estimations for different types of estimators assuming DD for b when $b = 3.5$

Parameters			**n = 50**			**n = 100**		
a	***p***		**MME**	**LME**	**TLME**	**MME**	**LME**	**TLME**
2.5	2.5	Bias	-0.0896	0.0022	0.0046	-0.0479	0.0012	0.0024
		RMSEs	0.6254	0.4854	0.4810	0.4653	0.3447	0.3347
	3.5	Bias	-0.0701	0.0016	0.0017	-0.0409	-0.0048	-0.0057
		RMSEs	0.5065	0.4349	0.4584	0.3680	0.3074	0.3200
	5.0	Bias	-0.0620	-0.0011	-0.0042	-0.0382	-0.0059	-0.0052
		RMSEs	0.4364	0.3964	0.4377	0.3103	0.2805	0.3099
3.5	2.5	Bias	-0.0686	-0.0010	-0.0019	-0.0319	0.0014	-0.0019
		RMSEs	0.4849	0.4228	0.4483	0.3482	0.2971	0.3125
	3.5	Bias	-0.0689	-0.0094	-0.0102	-0.0349	-0.0041	-0.0032
		RMSEs	0.4114	0.3858	0.4313	0.2870	0.2689	0.3014
	5.0	Bias	-0.0614	-0.0056	-0.0036	-0.0290	-0.0011	0.0003
		RMSEs	0.3760	0.3678	0.4258	0.2654	0.2586	0.2967
5.0	2.5	Bias	-0.0596	-0.0010	-0.0031	-0.0276	0.0016	0.0015
		RMSEs	0.4127	0.3923	0.4346	0.2915	0.2739	0.3046
	3.5	Bias	-0.0504	0.0027	-0.0022	-0.0265	0.0003	-0.0016
		RMSEs	0.3723	0.3686	0.4228	0.2635	0.2596	0.2963
	5.0	Bias	-0.0540	-0.0005	-0.0001	-0.0232	0.0031	0.0007
		RMSEs	0.3608	0.3641	0.4256	0.2497	0.2522	0.2928
Parameters			**n = 500**			**n = 1,000**		
a	***p***		**MME**	**LME**	**TLME**	**MME**	**LME**	**TLME**
2.5	2.5	Bias	-0.0085	-.00003	-0.0012	-0.0052	0.0003	0.0004
		RMSEs	0.2297	0.1563	0.1494	0.1649	0.1113	0.1064
	3.5	Bias	-0.0045	0.0023	0.0020	-0.0047	-0.0002	-0.0001
		RMSEs	0.1693	0.1379	0.1431	0.1186	0.0973	0.1019
	5.0	Bias	-0.0061	-.00004	-0.0002	-0.0035	-0.0004	-0.0001
		RMSEs	0.1403	0.1248	0.1362	0.0999	0.0885	0.0964
3.5	2.5	Bias	-0.0062	0.0004	-.00004	-0.0038	-0.0007	-0.0009
		RMSEs	0.1578	0.1316	0.1391	0.1126	0.0926	0.0971
	3.5	Bias	-0.0066	-0.0010	-0.0020	-0.0039	-0.0007	-0.0002
		RMSEs	0.1327	0.1220	0.1343	0.0937	0.0861	0.0948
	5.0	Bias	-0.0053	0.0001	0.0003	-0.0045	-0.0018	-0.0020
		RMSEs	0.1182	0.1147	0.1313	0.0843	0.0816	0.0927
5.0	2.5	Bias	-0.0069	-0.0005	.00006	-0.0027	0.0003	0.0006
		RMSEs	0.1316	0.1221	0.1333	0.0927	0.0862	0.0948
	3.5	Bias	-0.0034	0.0017	0.0015	-0.0022	0.0005	0.0005
		RMSEs	0.1152	0.1151	0.1312	0.0847	0.0830	0.0937
	5.0	Bias	-0.0060	-0.0006	-0.0003	-0.0031	-0.0004	-0.0001
		RMSEs	0.1116	0.1122	0.1303	0.0792	0.0797	0.0925

TABLE 3: Biases and RMSEs of the parameter estimations for different types of estimators assuming DD for b when $b = 5$

Parameters			**n = 50**			**n = 100**		
a	***p***		**MME**	**LME**	**TLME**	**MME**	**LME**	**TLME**
2.5	2.5	Bias	-0.1247	0.0074	0.0054	-0.0795	-0.0035	-0.0001
		RMSEs	0.9049	0.7027	0.6920	0.6469	0.4887	0.4805
	3.5	Bias	-0.0996	-0.0019	-0.0076	-0.0528	-0.0013	-0.0007
		RMSEs	0.7260	0.6181	0.6502	0.5265	0.4393	0.4588
	5.0	Bias	-0.0962	-0.0962	-0.0124	-0.0507	-0.0058	-0.0061
		RMSEs	0.6166	0.5616	0.6176	0.4481	0.4047	0.4439
3.5	2.5	Bias	-0.1018	-0.0061	-0.0101	-0.0451	0.0016	-0.0026
		RMSEs	0.6879	0.6015	0.6371	0.5018	0.4248	0.4466
	3.5	Bias	-0.0937	-0.0105	-0.0159	-0.0378	0.0045	0.0053
		RMSEs	0.5907	0.5554	0.6179	0.4201	0.3901	0.4309
	5.0	Bias	-0.0808	-0.0017	-0.0020	-0.0391	-0.0004	-0.0014
		RMSEs	0.5389	0.5281	0.6078	0.3799	0.3698	0.4258
5	2.5	Bias	-0.0875	-0.0045	-0.0051	-0.0464	-0.0037	-0.0036
		RMSEs	0.5914	0.5611	0.6207	0.4152	0.3917	0.4342
	3.5	Bias	-0.0752	0.0024	0.0012	-0.0432	-0.0036	-0.0027
		RMSEs	0.5409	0.5344	0.6110	0.3821	0.3771	0.4301
	5.0	Bias	-0.0750	0.0016	0.0047	-0.0386	-0.0003	0.00058
		RMSEs	0.5133	0.5199	0.6120	0.3583	0.3613	0.4231
Parameters			**n = 500**			**n = 1,000**		
a	***p***		**MME**	**LME**	**TLME**	**MME**	**LME**	**TLME**
2.5	2.5	Bias	-0.0162	0.0015	0.0045	-0.0078	-0.0017	-0.0026
		RMSEs	0.3183	0.2176	0.2121	0.2338	0.1558	0.1515
	3.5	Bias	-0.0111	-0.0013	-0.0020	-0.0060	-0.0003	0.0005
		RMSEs	0.2371	0.1926	0.2023	0.1694	0.1366	0.1423
	5.0	Bias	-0.0091	-.00003	-0.0001	-0.0049	-0.0006	-0.0009
		RMSEs	0.2046	0.1816	0.1957	0.1431	0.1269	0.1382
3.5	2.5	Bias	-0.0117	-0.0014	-0.0012	-0.0078	-0.0023	-0.0021
		RMSEs	0.2266	0.1883	0.1972	0.1598	0.1333	0.1404
	3.5	Bias	-0.0085	0.0003	0.0007	-0.0041	0.0002	0.0003
		RMSEs	0.1869	0.1730	0.1925	0.1354	0.1238	0.1360
	5.0	Bias	-0.0074	0.0005	0.0016	-0.0042	-0.0005	-0.0011
		RMSEs	0.1704	0.1654	0.1895	0.1201	0.1159	0.1319
5	2.5	Bias	-0.0122	-0.0035	-0.0042	-0.0048	-.00028	0.00003
		RMSEs	0.1869	0.1742	0.1901	0.1323	0.1231	0.1348
	3.5	Bias	-0.0076	0.0004	0.00155	-0.0041	-0.0004	-0.0010
		RMSEs	0.1705	0.1676	0.1909	0.1200	0.1174	0.1328
	5.0	Bias	-0.0092	-0.0014	-0.0005	-0.0036	0.0002	0.00034
		RMSEs	0.1623	0.1635	0.1896	0.1146	0.1150	0.1325

Table 4: Biases and RMSEs of the parameter estimations for different types of estimators assuming DD for b when $b = 10$

Parameters			**n = 50**			**n = 100**		
a	***p***		**MME**	**LME**	**TLME**	**MME**	**LME**	**TLME**
2.5	2.5	Bias	-0.2562	0.0065	0.0132	-0.1483	-0.0074	-0.0070
		RMSEs	1.7869	1.3870	1.3742	1.3197	0.9804	0.9584
	3.5	Bias	-0.1635	-0.0045	-0.0097	-0.1153	-0.0129	-0.0188
		RMSEs	1.4408	1.2335	1.2969	1.0372	0.8624	0.8994
	5.0	Bias	-0.1772	-0.0032	-0.0122	-0.0883	-0.0020	-0.0058
		RMSEs	1.2469	1.1327	1.2507	0.8985	0.8096	0.8840
3.5	2.5	Bias	-0.1964	-0.0037	-0.0100	-0.0901	0.0065	0.0053
		RMSEs	1.3818	1.2082	1.2771	0.9798	0.8342	0.8827
	3.5	Bias	-0.1818	-0.0133	-0.0187	-0.0862	-0.0038	-0.0091
		RMSEs	1.1858	1.1119	1.2317	0.8413	0.7762	0.8647
	5.0	Bias	-0.1586	-0.0010	0.0004	-0.0839	-0.0022	0.0060
		RMSEs	1.0750	1.0537	1.2175	0.7585	0.7381	0.8469
5	2.5	Bias	-0.1621	0.0052	-.00006	-0.0879	0.0010	0.0071
		RMSEs	1.1603	1.1058	1.2293	0.8236	0.7823	0.8742
	3.5	Bias	-0.1635	-0.0045	-0.0097	-0.0826	-0.0043	-0.0060
		RMSEs	1.0567	1.0512	1.2127	0.7544	0.7424	0.8457
	5.0	Bias	-0.1418	0.0104	0.01018	-0.0852	-0.0090	-0.0071
		RMSEs	10.214	10.333	12.134	0.7252	0.7291	0.8485
Parameters			**n = 500**			**n = 1,000**		
a	***p***		**MME**	**LME**	**TLME**	**MME**	**LME**	**TLME**
2.5	2.5	Bias	-0.0310	-0.0004	0.0015	-0.0154	-0.0011	-0.0019
		RMSEs	0.6482	0.4411	0.4260	0.4611	0.3135	0.3018
	3.5	Bias	-0.0247	-0.0026	-0.0004	-0.0068	0.0037	0.0056
		RMSEs	0.4792	0.3910	0.4080	0.3467	0.2768	0.2851
	5.0	Bias	-0.0152	0.0021	0.00211	-0.0138	-0.0043	-0.0041
		RMSEs	0.3983	0.3538	0.3866	0.2834	0.2529	0.2767
3.5	2.5	Bias	-0.0236	-0.0027	-0.0035	-0.0108	-0.0007	0.0003
		RMSEs	0.4556	0.3812	0.4035	0.3263	0.2682	0.2798
	3.5	Bias	-0.0163	0.0006	-0.0004	-0.0124	-0.0047	-0.0062
		RMSEs	0.3777	0.3469	0.3830	0.2678	0.2459	0.2709
	5.0	Bias	-0.0159	0.0001	0.0002	-0.0093	-0.0021	-0.0042
		RMSEs	0.3378	0.3290	0.3780	0.2397	0.2316	0.2638
5	2.5	Bias	-0.0083	-0.0011	-0.0017	-0.0087	-0.0004	-0.0018
		RMSEs	0.2654	0.2470	0.2733	0.2669	0.2483	0.2718
	3.5	Bias	-0.0165	-0.0005	-0.0008	-0.0094	-0.0021	-0.0037
		RMSEs	0.3389	0.3332	0.3789	0.2414	0.2364	0.2671
	5.0	Bias	-0.0060	0.0017	0.0027	-0.0070	.00005	-0.0008
		RMSEs	0.2279	0.2294	0.2668	0.2280	0.2283	0.2616

Table 5: Mean, S.D, CV, CS and CK different parametric values assuming MMEs, LMEs and TLMEs

Parameters			**Mean**	**S.D**	**CV**	**CS**	**CK**
a	***b***	***p***	**Method of Moment Estimates**				
2.5	2.5	2.5	1.74618	0.96423	0.55219	1.59264	9.96322
3.5			1.87901	0.73405	0.39066	0.87000	5.18644
5.0			2.01460	0.55391	0.27495	0.41096	3.83527
10			2.22241	0.31206	0.14041	-0.1101	3.47769
2.5	2.5	2.5	1.74618	0.96423	0.55219	1.59264	9.96322
	3.5		2.44465	1.34992	0.55219	1.59264	9.96322
	5.0		3.49236	1.92847	0.55219	1.59264	9.96322
	10		6.98473	3.85693	0.55219	1.59264	9.96322
2.5	2.5	2.5	1.74618	0.96423	0.55219	1.59264	9.96322
		3.5	1.46679	0.74442	0.50751	1.08950	5.67080
		5.0	1.23674	0.59315	0.47961	0.81551	4.25391
		10	0.90892	0.41060	0.45174	0.56301	3.35399
a	***b***	***p***	**L-Moment Estimates**				
2.5	2.5	2.5	1.74618	0.50944	0.29174	0.18854	0.16014
3.5			1.87901	0.40268	0.21430	0.11113	0.14526
5.0			2.01460	0.30854	0.15315	0.05080	0.14071
10			2.22241	0.17429	0.07842	-0.0220	0.14334
2.5	2.5	2.5	1.74618	0.50944	0.29174	0.18854	0.16014
	3.5		2.44465	0.71321	0.29174	0.18854	0.16014
	5.0		3.49236	1.01888	0.29174	0.18854	0.16014
	10		6.98473	2.03776	0.29174	0.18854	0.16014
2.5	2.5	2.5	1.74618	0.50944	0.29174	0.18854	0.16014
		3.5	1.46679	0.40484	0.27600	0.15103	0.14100
		5.0	1.23674	0.32781	0.26506	0.12369	0.12821
		10	0.90892	0.23011	0.25316	0.09272	0.11494
a	***b***	***p***	**TL-Moment Estimates**				
2.5	2.5	2.5	1.65012	0.25671	0.15557	0.10855	0.07708
3.5			1.83426	0.20651	0.11258	0.06144	0.07183
5.0			1.99892	0.15907	0.07958	0.02539	0.07064
10			2.22625	0.08958	0.04024	-0.01743	0.07239
2.5	2.5	2.5	1.65012	0.25671	0.15557	0.10855	0.07708
	3.5		2.31018	0.35939	0.15557	0.10855	0.07708
	5.0		3.30025	0.51342	0.15557	0.10855	0.07708
	10		6.60051	1.02685	0.15557	0.10855	0.07708
2.5	2.5	2.5	1.65012	0.25671	0.15557	0.10855	0.07708
		3.5	1.40564	0.20865	0.14844	0.08752	0.06943
		5.0	1.19619	0.17147	0.14334	0.07193	0.06422
		10	0.88759	0.12219	0.13767	0.05396	0.05870

[Recibido: septiembre de 2012 — Aceptado: mayo de 2013]

References

Bandourian, R., McDonald, J. & Turley, R. S. (2003), 'A comparison of parametric models of income distribution across countries and over time', *Estadistica* (55), 135–152.

Bílková, D. & Mala, I. (2012), 'Application of the L-moment method when modelling the income distribution in the Czech Republic', *Austrian Journal of Statistics* **41**(2), 125–132.

Dagum, C. (1690), *A model of Net Wealth Distribution Specified for Negative, Null and Positive Wealth. A Case of Study: Italy*, Springer Verlag Berlin, New York.

Dagum, C. (1977*a*), 'The generation and distribution of income, the Lorenz curve and the Gini ratio', *Economie Appliquee* **33**(2), 327–367.

Dagum, C. (1977*b*), 'A new model of personal income distribution: Specification and estimation', *Economie Appliquee* **30**(3), 413–437.

Dagum, C. & Lemmi, A. (1988), *A Contribution to the Analysis of Income Distribution and Income Inequality and a Case Study: Italy*, JAI Press, Greenwich.

Daud, M. Z., Kassim, A. H. M., Desa, M. N. M. & Nguyen, V. T. V. (2002), Statistical analysis of at-site extreme rainfall processes in peninsular Malaysia, *in* H. A. J. van Laanen & S. Demuth, eds, 'FRIEND 2002-Regional Hydrology: Bridging the Gap between Research and Practice', number 274, Proceedings of International Conferences, IAHS Publications, Cape Town, South Africa, pp. 61–68.

Elamir, E. A. & Seheult, A. H. (2003), 'Trimmed L-moments', *Computational Statistics and Data Analysis* (43), 299–314.

Hosking, J. R. M. (1990), 'L-moments: Analysis and estimation of distributions using linear combinations of order statistics', *Journal of the Royal Statistical Society. Series B. Statistical Methodological* (52), 105–124.

Kleiber, C. (1996), 'Dagum vs. Singh-Maddala income distributions', *Economics Letters* (53), 265–268.

Perez, C. G. & Alaiz, M. P. (2011), 'Using the Dagum model to explain changes in personal income distribution', *Applied Economics* (43), 4377–4386.

Quintano, C. & Dagostino, A. (2006), 'Studying inequality in income distribution of single person households in four developed countries', *Review of Income and Wealth* (52), 525–546.

Shabri, A., Ahmad, N. U. & Zakaria, A. Z. (2011), 'TL-moments and L-moments estimation of the generalized logistic distribution', *Journal of Mathematics Research* (10), 97–106.

Vogel, R. M. & Fennessey, N. M. (1993), 'L-moment diagrams should replace product moment diagrams', *Water Resources Research* (29), 1745–1752.

Ye, Y., Oluyede, B. O. & Pararai, M. (2012), 'Weighted generalized Beta distribution of the second kind and related distributions', *Journal of Statistical and Econometric Methods* (1), 13–31.

Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 95 a 114

Properties and Inference for Proportional Hazard Models

Propiedades e inferencia para modelos de Hazard proporcional

Guillermo Martínez-Florez[1,a], Germán Moreno-Arenas[2,b], Sandra Vergara-Cardozo[3,c]

[1]Departamento de Matemáticas y Estadística, Facultad de Ciencias, Universidad de Córdoba, Montería, Colombia

[2]Escuela de Matemáticas, Facultad de Ciencias, Universidad Industrial de Santander, Bucaramanga, Colombia

[3]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

We consider an arbitrary continuous cumulative distribution function $F(x)$ with a probability density function $f(x) = dF(x)/dx$ and hazard function $h_f(x) = f(x)/[1 - F(x)]$. We propose a new family of distributions, the so-called proportional hazard distribution-function, whose hazard function is proportional to $h_f(x)$. The new model can fit data with high asymmetry or kurtosis outside the range covered by the normal, t-student and logistic distributions, among others. We estimate the parameters by maximum likelihood, profile likelihood and the elemental percentile method. The observed and expected information matrices are determined and likelihood tests for some hypotheses of interest are also considered in the proportional hazard normal distribution. We show an application to real data, which illustrates the adequacy of the proposed model.

***Key words*:** Hazard function, Kurtosis, Method of moments, Profile likelihood, Proportional hazard model, Skewness, Skew-normal distribution.

Resumen

Consideramos una función de distribución continua arbitraria $F(x)$ con función de densidad de probabilidad $f(x) = dF(x)/dx$ y función de riesgo $h_f(x) = f(x)/[1 - F(x)]$. En este artículo proponemos una nueva familia de distribuciones cuya función de riesgo es proporcional a la función de riesgo $h_f(x)$. El modelo propuesto puede ajustar datos con alta asimetría o curtosis

[a]Professor. E-mail: gmartinez@correo.unicordoba.edu.co
[b]Associate professor. E-mail: gmorenoa@uis.edu.co
[c]Assistant Professor. E-mail: svergarac@unal.edu.co

fuera del rango de cobertura permitido por la distribución normal, t-Student, logística, entre otras. Estimamos los parámetros del modelo usando máxima verosimilitud, verosimilitud perfilada y el método elemental de percentiles. Calculamos las matrices de información esperada y observada. Consideramos test de verosimilitudes para algunas hipótesis de interés en el modelo con función de riesgo proporcional a la distribución normal. Presentamos una aplicación con datos reales que ilustra que el modelo propuesto es adecuado.

***Palabras clave*:** asimetría, curtosis, distribución skew-normal, función de riesgo, método de los momentos, modelo de riesgo proporcional, verosimilitud perfilada.

1. Introduction

When data originates from heavy-tailed or asymmetrical distributions, the normality-based inferential processes are inadequate. In these situations many authors choose to transform the variables in order to attain symmetry or normality. These transformations produce unsatisfactory results because the interpretation of the results becomes cumbersome. Although the class of elliptic distributions is a good alternative for situations with heavy-tailed behavior, this is not appropriate when the distribution is asymmetric. These circumstances prompted the search for new distributions, better suited to fit data with high asymmetry or kurtosis. The literature on families of flexible distributions has experienced great increase in the last three or four decades. Some early results include Lehmann (1953), Roberts (1966) and O'Hagan & Leonard (1976), among others. Azzalini (1985), Durrans (1992), Fernandez & Steel (1998), Mudholkar & Hutson (2000), Gupta, Chang & Huang (2002), Arellano-Valle, Gómez & Quintana (2004, 2005), Gómez, Venegas & Bolfarine (2007), Arnold, Gómez & Salinas (2009), Pewsey, Gómez & Bolfarine (2012) represent some of the important contributions.

Azzalini (1985) defines a probability density function of a random variable Z with skew-normal distribution and parameter λ, given by

$$f_{SN}(z;\lambda) = 2\phi(z)\Phi(\lambda z),\ z \in \mathbb{R} \tag{1}$$

where ϕ and Φ denote the standard normal density and the cumulative distribution functions, respectively. The skewness is controlled by the parameter λ. We denote this by $Z \sim SN(\lambda)$. The asymmetry and kurtosis coefficients for this distribution are in the intervals $(-0.9953, 0.9953)$ and $[3, 3.8692)$, respectively. The skew-normal distribution was first introduced by O'Hagan & Leonard (1976) as a prior distribution for estimating a normal location parameter. The density (1) has also been studied widely by Henze (1986), Chiogna (1998), Pewsey (2000) and Gómez et al. (2007).

Durrans (1992), in a hydrological context, introduced the fractional order statistics distribution with density function given by

$$g_F(z;\alpha) = \alpha f(z)\{F(z)\}^{\alpha-1},\ z \in \mathbb{R},\ \alpha \in \mathbb{R}^+ \tag{2}$$

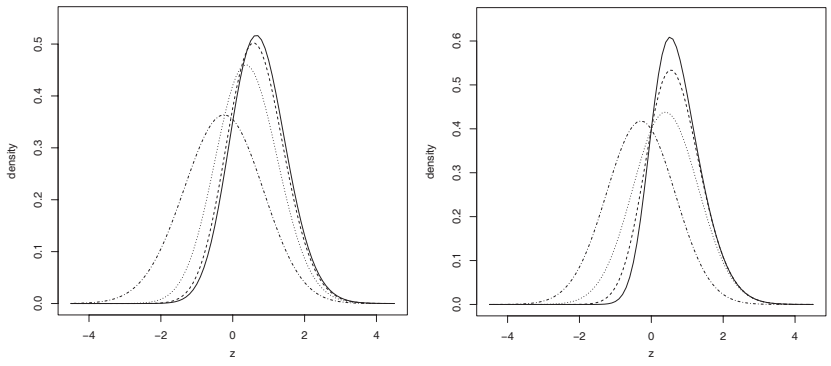
where F is an absolutely continuous distribution function, f is a corresponding density function and α is a shape parameter that controls the amount of asymmetry in the distribution. We refer to this model as the power distribution. We use the notation $Z \sim AP(\alpha)$.

Following the idea of Durrans, Gupta & Gupta (2008) we define the power-normal distribution whose distribution function is given by

$$g_\Phi(z;\alpha) = \alpha\phi(z)\{\Phi(z)\}^{\alpha-1},\ z \in \mathbb{R},\ \alpha \in \mathbb{R}^+ \tag{3}$$

We use the notation $Z \sim PN(\alpha)$. Pewsey, Gómez and Bolfarine (2012) showed that the expected information matrix is nonsingular for the neighborhood of $\alpha = 1$, contrary to the skew-normal distribution where the information matrix is singular under the symmetry hypothesis ($\lambda = 0$). They also found that the asymmetry and kurtosis coefficients for this distribution are in the intervals $[-0.6115, 0.9007]$ and $[1.7170, 4.3556]$, respectively.

Figure 1 shows how the parameters α and λ control the asymmetry and kurtosis of the (1) and (3) models.



(a) (b)

FIGURE 1: Probability density function (a) $PN(\alpha)$ for $\alpha = 0.746$ (dashed-dotted line), 1.626 (dotted line), 2.254 (dashed line) and 2.516 (solid line). (b) $SN(\lambda)$ for λ =-0.40 (dashed-dotted line), 0.60 (dotted line), 1.40 (dashed line) and 2.20 (solid line).

In this paper we present a new family of distributions so-called proportional hazard distribution-functions. The paper is presented as follows. In Section 2 we define the proportional hazard distribution-function, study some of its properties and discuss maximum likelihood estimation. The location-scale extension for proportional hazard distribution-function is presented in Section 3. In Section 4, we define the *location-scale proportional hazard normal* model and different methods for parameter estimation; we derive the information matrix and discuss likelihood ratio tests for some hypotheses of interest. Further, the asymptotic distribution of maximum likelihood estimators is obtained. The usefulness of the proposed model is illustrated in an application to real data in Section 5. Finally, some concluding remarks are found in Section 6.

2. Proportional Hazard Distribution-Function

Let $F(x)$ be a continuous cumulative distribution function with probability density function $f(x)$ and hazard function $h_f(x) = f(x)/(1 - F(x))$. We will say that Z has proportional hazard distribution-function associated with F and f and parameter $\alpha > 0$ if its probability density function is

$$\varphi_F(z;\alpha) = \alpha f(z)\{1 - F(z)\}^{\alpha-1}, \quad z \in \mathbb{R} \tag{4}$$

where α is a positive real number and F is a continuous distribution function with continuous density function f. We use the notation $Z \sim PHF(\alpha)$. The distribution function of the PHF model is given by

$$\mathbb{F}(z) = 1 - \{1 - F(z)\}^{\alpha}, \ z \in \mathbb{R} \tag{5}$$

We observe that the name "proportional hazard distribution-function" is appropriate because its hazard function with respect to the density φ_F is

$$h_{\varphi_F}(x,\alpha) = \alpha h_f(x)$$

The inversion method can be used to generate a $PHF(\alpha)$ distribution. Thus, if U is a uniform random variable on $(0, 1)$,

$$Z = F^{-1}(1 - (1 - u)^{1/\alpha})$$

obeys a $PHF(\alpha)$ distribution, whose median, $Z_{0.5}$, can be found from the inverse of F through

$$Z_{0.5} = F^{-1}\left(\frac{2^{1/\alpha} - 1}{2^{1/\alpha}}\right)$$

where F^{-1} is the inverse of the distribution F. In general, the *p-th* percentile can be computed by

$$Z_p = F^{-1}\left(1 - (1 - p)^{1/\alpha}\right)$$

The distribution mode is the solution to the non-linear equation

$$[1 - F(z)]\, f'(z) - (\alpha - 1) f^2(z) = 0$$

where f' is the derivative of F.

In the next section we present some particular cases of the PHF distribution.

2.1. Proportional Hazard Normal Distribution

When $F = \Phi$, the standard normal distribution function, we obtain the *proportional hazard normal* distribution, which we denote by $PHN(\alpha)$. Its density function is given by

$$\varphi_\Phi(z;\alpha) = \alpha\phi(z)\{1 - \Phi(z)\}^{\alpha-1}, \ z \in \mathbb{R} \tag{6}$$

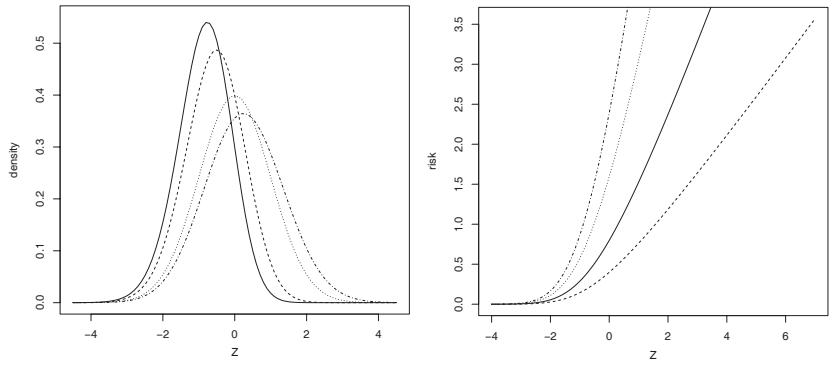
This model is also an alternative to accommodate data with asymmetry and kurtosis that are outside the ranges allowed by the normal distribution. The PHN is a special case of Eugene, Lee & Famoye (2002)'s *beta-normal* distribution. A simple comparison makes clear that $PHN(1) = SN(0) = PN(1) = N(0, 1)$.

The survival function and the hazard function are given, respectively, by

$$S(t) = \{1 - \Phi(t)\}^{\alpha} \quad \text{and} \quad h_{\varphi_\Phi}(t) = \alpha h_{\phi}(t)$$

That is to say, the PHN model's hazard function is directly proportional to the normal model's hazard function. It can then be said that the hazard function is a non decreasing (and unimodal) function of T, but an increasing function of parameter α. It can also be said that for $\alpha > 1$, the PHN's model hazard is greater than the normal's model, while for $\alpha > 1$ the opposite occurs.

In Figure 2-(a) we can see the behavior of the $PHN(\alpha)$ density and Figure 2-(b) shows the model's hazard function for a few values of the parameter α.



(a) (b)

FIGURE 2: (a) $PHN(\alpha)$ for α =0.75 (solid line), 1.0 (dashed line), 2.0 (dotted line), and 3.0 (dashed-dotted line) (b) $h_{\varphi_\Phi}(z)$ for α =0.25 (dashed line), 1.0 (solid line), 2.0 (dotted line), 3.0 (dashed-dotted line)

2.2. Proportional Hazard Logistic Distribution

The *proportional hazard logistic* distribution is defined by the probability density function

$$\varphi_L(z; \alpha) = \alpha \exp(x) \left\{ \frac{1}{1 + \exp(x)} \right\}^{\alpha+1} \tag{7}$$

We denote it by $PHL(\alpha)$. Figure 3 shows the behaviour of the this distribution for diferents values of the α.

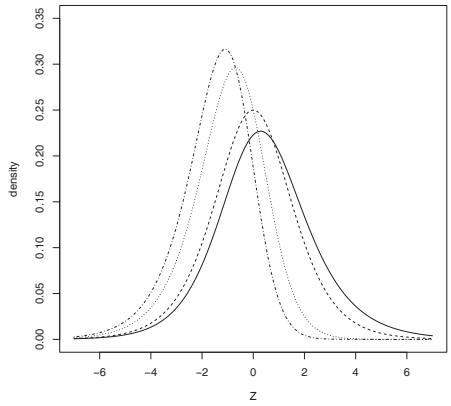


FIGURE 3: $PHL(\alpha)$ distribution for α =0.75 (solid line), 1.0 (dashed line), 2.0 (dotted line) and 3.0 (dashed-dotted line)

2.3. Proportional Hazard t-Student Distribution

The *proportional hazard t-student* distribution is defined by the probability density function

$$\varphi_T(z;\alpha,v) = \frac{\alpha\Gamma(\frac{v+1}{2})}{(v\pi)^{1/2}\Gamma(\frac{v}{2})}\left[1+\frac{z^2}{v}\right]^{-(v+1)/2}\{1-T(z)\}^{\alpha-1} \tag{8}$$

where T is the cumulative distribution function of the t-student distribution and v is the number of degrees of freedom. The notation we use is $PHt(v,\alpha)$. Figure 4 shows the behavior of this distribution.

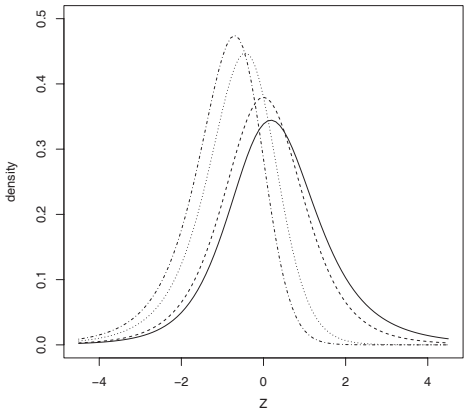


FIGURE 4: $PHt(v,\alpha)$ distribution for α =0.75 (solid line), 1.0 (dashed line), 2.0 (dotted line) and 3.0 (dashed-dotted line)

2.4. Proportional Hazard Cauchy Distribution

When $v = 1$ in $PHt(v, \alpha)$ gives the *proportional hazard Cauchy* distribution, whose probability density function is

$$\varphi_C(z;\alpha) \quad = \frac{\alpha}{\pi[1+z^2]} \left\{ \frac{1}{2} - \frac{1}{\pi} \arctan(z) \right\}^{\alpha-1} \tag{9}$$

We denote it by $PHC(\alpha)$. Figure 5 shows the behavior of this distribution for different values of the α parameter.

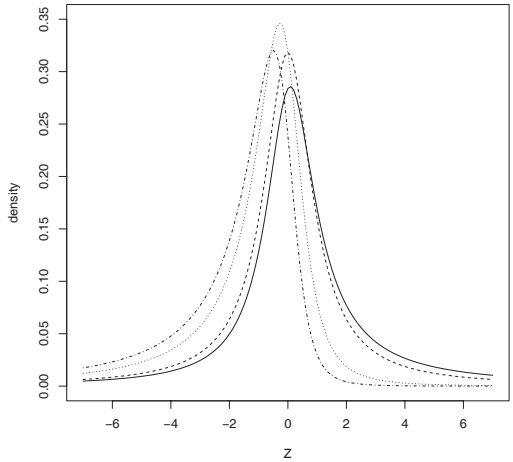


FIGURE 5: $PHC(\alpha)$ distribution for α =0.75 (solid line), 1.0 (dashed line), 2.0 (dotted line) and 3.0 (dashed-dotted line)

2.5. Moments of the PHF

The moment generating function for the PHF distribution is given by

$$M(t) = \alpha \int_0^1 \exp\left\{tF^{-1}(y)\right\} (1-y)^{\alpha-1} dy \tag{10}$$

There is no closed form for the moments of a random variable Z with distribution $PHF(\alpha)$; these are computed numerically.

The r-th Z moment for the random variable $Y \sim PHF$ can be obtained with the expression

$$\mu_r = \alpha \int_0^1 \left\{F^{-1}(y)\right\}^r (1-y)^{\alpha-1} dy, \ r = 0, 1, 2, \ldots \tag{11}$$

This expectation agrees with the expected value of the function $\left\{F^{-1}(y)\right\}^r$ where Y is a random variable with a Beta distribution with parameters α and 1. The central moments $\acute{\mu_r} = E(Z - E(Z))^r$ for $r = 2, 3, 4$ can be found from the expressions

$\acute{\mu}_2 = \mu_2 - \mu_1^2$, $\acute{\mu}_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3$ and $\acute{\mu}_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4$. Consequently, the variance asymmetry and kurtosis coefficients are $\sigma^2 = Var(Z) = \acute{\mu}_2$, $\sqrt{\beta_1} = \acute{\mu}_3/[\acute{\mu}_2]^{3/2}$ and $\beta_2 = \acute{\mu}_4/[\acute{\mu}_2]^2$, respectively.

For $F = \Phi$, that is, the case of the $PHN(\alpha)$ distribution, the r-th Z moment is given by

$$\mu_r = \alpha \int_0^1 \left\{\Phi^{-1}(y)\right\}^r (1-y)^{\alpha-1} dy,\ r = 0, 1, 2, \ldots \tag{12}$$

Thus, for α values between 0.0005 and 9,000 the asymmetry and kurtosis coefficients, $\sqrt{\beta_1}$ and β_2 of the variable $Z \sim PHN(\alpha)$ belong to the intervals (-1.1578, 0.9918) and (1.1513,4.3023), respectively. Therefore the PHN distribution clearly fits data with less negative asymmetry and more platykurtic than the SN and PN distributions do. It also fits distributions with a higher positive asymmetry than PN and more leptokurtic than SN. It is evident that the PHN distribution fits data with as much positive asymmetry as SN distribution does and as much kurtosis as PN distribution does.

3. Location-Scale PHF

Let $Z \sim PHF(\alpha)$ with $\alpha \in \mathbb{R}^+$. The family of PHF distributions with location-scale parameters is defined as the distribution of $X = \xi + \eta Z$ for $\xi \in \mathbb{R}$ and $\eta > 0$. The corresponding density function is given by

$$\varphi_F(x; \xi, \eta, \alpha) = \frac{\alpha}{\eta} f\left(\frac{x-\xi}{\eta}\right) \left\{1 - F\left(\frac{x-\xi}{\eta}\right)\right\}^{\alpha-1},\ x \in \mathbb{R} \tag{13}$$

where ξ is the location parameter and η is the scale parameter. We use the notation $PHF(\xi, \eta, \alpha)$.

3.1. Estimation and Inference for the Location-Scale PHF

We now deduce the maximum likelihood estimators (MLE) for the parameters of the $PHF(\xi, \eta, \alpha)$ distribution and the respective observed and expected information matrices.

For n observations, $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ from the $PHF(\xi, \eta, \alpha)$ distribution, the log-likelihood function of $\boldsymbol{\theta} = (\xi, \eta, \alpha)'$, given $\mathbf{x}$, is

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = n\log(\alpha) - n\log(\eta) + \sum_{i=1}^{n} \log(f(z_i)) + (\alpha - 1)\sum_{i=1}^{n} \log(1 - F(z_i))$$

where $z_i = \frac{x_i - \xi}{\eta}$. Thus, under the assumption that the derivative of f exists, the score function is given by:

$$\begin{aligned} U(\xi) &= -\frac{1}{\eta}\sum_{i=1}^{n}\frac{f'(z_i)}{f(z_i)} + \frac{\alpha-1}{\eta}\sum_{i=1}^{n}\frac{f(z_i)}{1-F(z_i)}, \\ U(\eta) &= -\frac{n}{\eta} - \frac{1}{\eta}\sum_{i=1}^{n} z_i\frac{f'(z_i)}{f(z_i)} + \frac{\alpha-1}{\eta}\sum_{i=1}^{n} z_i\frac{f(z_i)}{1-F(z_i)}, \\ U(\alpha) &= \frac{n}{\alpha} + \sum_{i=1}^{n}\log[1-F(z_i)] \end{aligned}$$

MLE estimators are the solutions to this system of equations usually solved by iterative numerical methods. It is usual to use a software algorithm implemented in R (R Development Core Team 2013).

3.1.1. Observed Information Matrix for the Location-Scale PHF

Assuming the existence of the second derivative of f and putting $w_i = \frac{f(z_i)}{1-F(z_i)}$, $s_i = \frac{f'(z_i)}{1-F(z_i)}$, $t_i = \frac{f''(z_i)}{f(z_i)}$ and $v_i = \frac{f'(z_i)}{f(z_i)}$, the observed information matrix entries, $j_{\xi\xi}, j_{\eta\xi}, \ldots, j_{\alpha\alpha}$, are obtained:

$$j_{\xi\xi} = -\frac{n}{\eta^2}\left\{(\overline{v^2} - \overline{t}) + (\alpha-1)\left[\overline{w^2} + \overline{s}\right]\right\}$$

$$j_{\eta\xi} = -\frac{n}{\eta^2}(\overline{v} + \overline{t} - \overline{v^2}) + n\frac{\alpha-1}{\eta^2}\left[\overline{zw^2} + \overline{zs} + \overline{w}\right]$$

$$j_{\eta\eta} = -\frac{n}{\eta^2} + \frac{n}{\eta^2}\left[-2\overline{zv} - \overline{z^2 t} + \overline{z^2 v^2}\right] + n\frac{\alpha-1}{\eta^2}\left[2\overline{zw} + \overline{z^2 s} + \overline{z^2 w^2}\right]$$

$$j_{\alpha\xi} = -\frac{n}{\eta}\overline{w} \qquad j_{\alpha\eta} = -\frac{n}{\eta}\overline{zw} \qquad j_{\alpha\alpha} = \frac{n}{\alpha^2}$$

where $\overline{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$, $\overline{v^2} = \frac{1}{n}\sum_{i=1}^{n} v_i^2$,...,$\overline{z^2 w^2} = \frac{1}{n}\sum_{i=1}^{n} z_i^2 w_i^2$.

3.1.2. Expected Information Matrix for the Location-Scale PHF

The expected information matrix entries are n^{-1} times the expected value of the observed information matrix elements, that is,

$$I_{\theta_r\theta_p} = n^{-1}E\left\{-\frac{\partial^2\ell(\boldsymbol{\theta};\mathbf{x})}{\partial\theta_r\partial\theta_p}\right\},\ r,p = 1,2,3,\ \text{with } \theta_1 = \xi,\ \theta_2 = \eta \text{ and } \theta_3 = \alpha.$$

Considering the notation below (Pewsey et al. 2012):

$$\begin{aligned} a_{kj} &= E\{z^k\,(f(z)/[1-F(z)])^j\} \\ b_k &= E\{z^k f'(z)/[1-F(z)]\} \\ c_{kj} &= E\{z^k(f'(z)/f(z))^j\} \\ d_k &= E\{z^k f''(z)/f(z)\} \text{ for } k = 0,1,2 \text{ and } j = 1,2 \end{aligned}$$

the observed information matrix elements these are given by

$$\begin{aligned} I_{\xi\xi} &= \{(c_{02}-d_0)+(\alpha-1)(a_{02}+b_0)\}/\eta^2, \\ I_{\xi\eta} &= \{(c_{12}-c_{01}-d_1)+(\alpha-1)(a_{12}+b_1+a_{01})\}/\eta^2 \\ I_{\xi\alpha} &= E(w)/\eta = a_{01}/\eta \\ I_{\eta\eta} &= \{(c_{22}-d_2-2c_{11}-1)+(\alpha-1)(a_{22}+b_2+2a_{11})\}/\eta^2 \\ I_{\eta\alpha} &= E(zw)/\eta = a_{11}/\eta, \text{ and } I_{\alpha\alpha} = 1/\alpha^2 \end{aligned}$$

In general, these expected values are computed using numerical integration. When $\alpha = 1$, we have $\varphi(x;\xi,\eta,1) = \frac{1}{\eta} f\left(\frac{x-\xi}{\eta}\right)$, the location-scale f function model, thus the matrix information is reduced to

$$\begin{pmatrix} (c_{02}-d_0)/\eta^2 & (c_{12}-c_{01}-d_1)/\eta^2 & a_{01}/\eta \\ (c_{12}-c_{01}-d_1)/\eta^2 & (c_{22}-d_2-2c_{11}-1)/\eta^2 & a_{11}/\eta \\ a_{01}/\eta & a_{11}/\eta & 1 \end{pmatrix}$$

The properties of this matrix depend on the function f.

4. Location-Scale Proportional Hazard Normal

A very special particular case of the $PHF(\xi,\eta,\alpha)$ model occurs when $F = \Phi$, the standard normal distribution function. In this case the probability density function is

$$\varphi_\Phi(x;\xi,\eta,\alpha) = \frac{\alpha}{\eta}\phi\left(\frac{x-\xi}{\eta}\right)\left\{1-\Phi\left(\frac{x-\xi}{\eta}\right)\right\}^{\alpha-1}, \; x \in \mathbb{R} \tag{14}$$

which we call *location-scale proportional hazard normal*. Note that when $\alpha = 1$ we are in the case of the location-scale normal distribution.

In what follows we discuss estimation by moments, maximum likelihood, profiled likelihood and elemental percentile method for the $PHN(\xi,\eta,\alpha)$ model and show the respective observed and information matrices for a PHN random variable.

4.1. Estimation by the Method of Moments for the Location-Scale PHN

The mean (μ), variance (σ^2) and asymmetry coefficient ($\sqrt{\beta_1}$) in the location-scale case are:

$$\mu = \xi + \eta\Phi_1(\alpha), \quad \sigma^2 = \eta^2\Phi_2(\alpha) \;\text{ and }\; \sqrt{\beta_1} = \frac{\mu_3'}{\sigma^3} = \Phi_3(\alpha)$$

Thus, the estimators for α, ξ and η can be obtained by substituting, in the above expressions, μ, σ^2 and $\sqrt{\beta_1}$ for their respective sample moments $\bar{y}$, s^2 and

$\sqrt{b_1}$. First the α estimator is obtained as in the standard case and its value can be used to estimate $\Phi_1(\alpha)$ and $\Phi_2(\alpha)$, leaving a simple 2×2 system of linear equations to solve, whose solution gives the ξ and η estimators. The asymptotic distribution of moment estimators is widely studied in Sen & Singer (1993) and Sen, Singer & Pedroso de Lima (2010).

4.2. Maximum Likelihood Estimation for the Location-Scale PHN

For n observations, $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ from the $PHN(\xi, \eta, \alpha)$ distribution, the log-likelihood function of $\theta = (\xi, \eta, \alpha)^\top$ given $\mathbf{x}$ is

$$\ell(\theta; \mathbf{x}) = n\log(\alpha) - n\log(\eta) + \sum_{i=1}^{n}\log(\phi(z_i)) + (\alpha-1)\sum_{i=1}^{n}\log(1-\Phi(z_i))$$

where $z_i = \frac{x_i-\xi}{\eta}$. Thus, the score function, defined as the derivative of the log-likelihood function with respect to each of the parameters, is:

$$U(\alpha) = \frac{n}{\alpha} + \sum_{i=1}^{n}\log[1-\Phi(z_i)]$$

$$U(\xi) = \frac{1}{\eta}\sum_{i=1}^{n} z_i + \frac{\alpha-1}{\eta}\sum_{i=1}^{n}\frac{\phi(z_i)}{1-\Phi(z_i)}$$

$$U(\eta) = -\frac{n}{\eta} + \frac{1}{\eta}\sum_{i=1}^{n} z_i^2 + \frac{\alpha-1}{\eta}\sum_{i=1}^{n} z_i\frac{\phi(z_i)}{1-\Phi(z_i)}$$

Setting these expressions to zero, we get the corresponding score equations whose numerical solution leads to the MLE estimators.

4.2.1. Observed information matrix for location-scale PHN

The observed information matrix follows from minus the second derivatives of the log-likelihood function, which are denoted by $j_{\xi\xi}, j_{\xi\eta}, \ldots, j_{\alpha\alpha}$, and are given by

$$j_{\xi\xi} = \frac{n}{\eta^2} + n\frac{\alpha-1}{\eta^2}\left[\overline{w^2} - \overline{zw}\right]$$

$$j_{\xi\eta} = \frac{2n}{\eta^2}\overline{z} + n\frac{\alpha-1}{\eta^2}\left[-\overline{zw^2} - \overline{z^2w} + \overline{w}\right]$$

$$j_{\eta\eta} = -\frac{n}{\eta^2} + \frac{3n}{\eta^2}\overline{z^2} + n\frac{\alpha-1}{\eta^2}\left[2\overline{zw} + \overline{z^2w^2} - \overline{z^3w}\right]$$

$$j_{\xi\alpha} = -\frac{n}{\eta}\overline{w} \qquad j_{\eta\alpha} = -\frac{n}{\eta}\overline{zw} \qquad j_{\alpha\alpha} = \frac{n}{\alpha}, \text{ where } w_i = \frac{\phi(z_i)}{1-\Phi(z_i)}$$

$\overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$, $\overline{w^2} = \frac{1}{n}\sum_{i=1}^{n} w_i^2$, $\overline{zw} = \frac{1}{n}\sum_{i=1}^{n} z_i w_i$ $\ldots$, $\overline{z^2w^2} = \frac{1}{n}\sum_{i=1}^{n} z_i^3 w_i^2$

4.2.2. Expected Information Matrix for the Location-Scale PHN

Considering $a_{kj} = \mathbb{E}\{z^k w^j\}$, the expected information matrix entries are:

$$I_{\xi\xi} = \frac{1}{\eta^2}\left[1 + (\alpha - 1)(a_{02} - a_{11})\right] \qquad I_{\eta\xi} = \frac{2}{\eta^2}a_{10} + \frac{\alpha - 1}{\eta^2}\left[a_{01} - a_{02} + a_{12}\right]$$

$$I_{\eta\eta} = -\frac{1}{\eta^2} + \frac{3}{\eta^2}a_{20} + \frac{\alpha - 1}{\eta^2}\left[a_{22} + 2a_{11} - a_{31}\right]$$

$$I_{\alpha\xi} = -\frac{1}{\eta}a_{01} \qquad I_{\alpha\eta} = -\frac{1}{\eta}a_{11} \qquad I_{\alpha\alpha} = \frac{1}{\alpha^2}$$

The expected values of the above variables are generally calculated using numerical integration. When $\alpha = 1$, $\varphi(x; \xi, \eta, 1) = \frac{1}{\eta}\phi\left(\frac{x-\xi}{\eta}\right)$, the location-scale normal density function. Thus, the information matrix becomes

$$I(\theta) = \begin{pmatrix} 1/\eta^2 & 0 & -a_{01}/\eta \\ 0 & 2/\eta^2 & -a_{11}/\eta \\ -a_{01}/\eta & -a_{11}/\eta & 1 \end{pmatrix}$$

Numerical integration shows that the determinant is

$$|I(\theta)| = \frac{1}{\eta^4}[2 - a_{11}^2 - 2a_{01}^2] = \frac{0.013687}{\eta^4} \neq 0$$

so in the case of a normal distribution the information matrix of the model is non-singular. The upper left 2×2 submatrix is the normal distribution's information matrix for the normal distribution.

For large n and under regularity conditions we have

$$\hat{\theta} \xrightarrow{A} N_3(\theta, I(\theta)^{-1})$$

and the conclusion follows that $\hat{\theta}$ is consistent and asymptotically approaches the normal distribution with $I(\theta)^{-1}$ as the covariance matrix, for large samples.

4.3. Profile Likelihood Estimation for the Location-Scale PHN

Maximum likelihood estimators of the $PHN(\xi, \eta, \alpha)$ distribution parameters usually display high levels of bias in the estimation of the shape parameter α when the sample size is small. Other estimation techniques can be used that result in a more consistent estimation of α. Among these are the profile likelihood and the modified profile likelihood (see Barndorff-Nielsen 1983, Severini 1998). Thus, calling $\boldsymbol{\tau} = (\xi, \eta)^\top$ the vector of parameters of interest and $\phi = \alpha$ the nuisance parameter, the profile likelihood is

$$L_p(\boldsymbol{\tau}) = L(\boldsymbol{\tau}, \hat{\phi}_{\boldsymbol{\tau}})$$

where $\hat{\phi}_{\boldsymbol{\tau}} = \hat{\alpha}(\xi,\eta) = -n\left\{\sum_{i=1}^{n}\log\left[1-\Phi\left(\frac{x_i-\xi}{\eta}\right)\right]\right\}^{-1}$. Substituting $\hat{\alpha}(\xi,\eta)$ in the original likelihood we obtain the profile log-likelihood, defined as the logarithm of the profile likelihood:

$$\ell_p(\xi,\eta) = n\left[\log(n) - \log\left(-\sum_{i=1}^{n}\log\left[1-\Phi\left(z_i\right)\right]\right) - \log(\eta) - \frac{1}{2}\log(2\pi) - 1\right] - \frac{1}{2}\sum_{i=1}^{n} z_i^2 - \sum_{i=1}^{n}\log\left[1-\Phi\left(z_i\right)\right] \tag{15}$$

where $z_i = (x_i - \xi)/\eta$.

Consequently, the profiled maximum likelihood estimators for ξ and η, that is, $\hat{\xi}_p$ and $\hat{\eta}_p$, are the solutions to the nonlinear equations $u_p(\xi) = \frac{\partial \ell_p(\xi,\eta)}{\partial \xi} = 0$ and $u_p(\eta) = \frac{\partial \ell_p(\xi,\eta)}{\partial \eta} = 0$, which are obtained with iterative numerical methods.

Since sometimes the estimation of parameters by maximum likelihood can be inconsistent or inefficient, Barndorff-Nielsen (1983) proposes a modified profiled likelihood. Severini (2000) presents an alternative that is easier to apply in certain models like $PHN(\xi,\eta,\alpha)$. The profiled likelihood is not an actual likelihood, because some of the likelihood properties are not verified. for instance, the score function may have a nonzero mean and the observed information can have a bias. Nevertheless this function has go some interesting properties that make it look like an actual likelihood. For more examples, properties and uses of estimation by modified or unmodified profiled likelihood see Farias, Moreno & Patriota (2009).

4.4. Estimation by the Elemental Percentile Method for the Location-Scale PHN

The elemental percentile method can also be used in the estimation of the $PHN(\xi,\eta,\alpha)$ parameters applying the theory developed in Castillo & Hadi (1995).

Estimation of ξ and η when α is known. If the shape parameter α is known, the elemental percentile method for two order statistics $x_{(i)}$ and $x_{(j)}$, with $i < j$, leads to the equations

$$\hat{\eta}(i,j) = \frac{x_{(j)} - x_{(i)}}{\Phi^{-1}\left(1-\left(\frac{(n-j)+1}{n+1}\right)^{1/\alpha}\right) - \Phi^{-1}\left(1-\left(\frac{(n-i)+1}{n+1}\right)^{1/\alpha}\right)}$$

and

$$\hat{\xi}(i,j) = x_{(j)} - \hat{\eta}(i,j)\Phi^{-1}\left(1-\left(\frac{(n-j)+1}{n+1}\right)^{1/\alpha}\right)$$

Then, proceeding like in the previous case (for α), we select m samples of two order statistics and estimate the parameters ξ and η and again using robust statistics we finally get the estimators for these parameters.

A second estimation method, in two steps, using percentiles is illustrated next. It is motivated in the fact that the maximum likelihood method gives fairly good estimations of the location and scale (ξ and η) parameters.

Initially it is assumed that the location and scale parameters are known and their actual values are the MLE estimators, and we estimate α like in the standard case. Once the α estimator is known, the second step is to suppose that this is the actual value of the parameter and then we estimate the ξ and η values under the assumption that α is known. The standard errors for the parameter estimations can be computed using resampling techniques such as Jackknife or Bootstrap (see Efron (1982, 1979)). In both cases above we took $p_i = i/(n+1)$, given that we know $\mathbb{E}(\mathbb{F}) = i/(n+1)$.

4.5. Simulation Study

To study the MLE estimator properties of the $PHN(\xi, \eta, \alpha)$ distribution, a simulation was carried out for $\alpha = 0.75$, 1.5 and 3.0. Without loss of generality the location and scale parameters were set at $\xi = 0$ and $\eta = 1$.

The sample sizes in the simulation were $n = 50$, 100, 200 and 500 with 2,000 replications in each case. The random variable X with distribution $PHN(\xi, \eta, \alpha)$ was obtained with the algorithm

$$X = \xi + \eta\Phi^{-1}(1 - (1-u)^{1/\alpha}),$$

where u is a uniform random variable $U(0,1)$. In all cases, the bias and root mean square errors of the MLE estimators were calculated.

The results shown in Tables (1) and (2) demonstrate that when the sample size increases, the bias and root mean square error decrease, that is, the estimators are asymptotically consistent. Still, a high bias in the shape parameter α for small sample sizes is evident. In conclusion, this estimation process would be recommended for very large sample sizes. Using the profiled likelihood estimation method for α we found biases 0.2511 and 0.7241 for values $\alpha = 0.75$ and 1.5 respectively with a sample size 100.

Table 1: Bias of the MLE from PHN model parameters.

	$\alpha = 0.75$			$\alpha = 1.5$			$\alpha = 3.0$		
n	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$
50	0.1546	-0.0635	1.5529	0.0947	-0.0700	2.0300	-0.1128	-0.0915	1.9252
100	0.1523	-0.0061	0.4897	0.0899	-0.0163	1.4511	-0.0722	-0.0547	1.8106
200	0.0725	-0.0020	0.1831	0.0636	-0.0054	0.5113	0.0665	-0.0148	1.2823
500	0.0307	0.0001	0.0600	0.0321	-0.0005	0.1519	0.0325	0.0005	0.4562

Tables (3) and (4) show the behavior of estimators by the elemental percentile method for the $PHN(\xi, \eta, \alpha)$ model. As can been seen, these also are asymptotically consistent and their biases are less than the biases of the maximum likelihood estimators for a small sample. However, the bias of the α estimator is still too large. For small sample sizes, Jackknife or Bootstrap estimators can be applied to correct the bias of the MLE estimators (see, Efron 1982, Efron & Tibshirani 1993).

TABLE 2: $\sqrt{MSE}$ of the MLE from PHN model parameters.

	$\alpha = 0.75$			$\alpha = 1.5$			$\alpha = 3.0$		
n	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$
50	1.5939	0.5583	3.6623	1.3763	0.4745	4.3718	1.1312	0.3808	4.5231
100	1.2367	0.4147	1.5684	1.0863	0.3440	3.4739	0.9468	0.2917	3.9510
200	0.8756	0.2945	0.7383	0.8353	0.2585	1.6110	0.7430	0.2169	3.4404
500	0.5102	0.1756	0.3607	0.5313	0.1633	0.7819	0.5374	0.1517	1.9056

TABLE 3: Bias of the PHN model percentile estimators.

	$\alpha = 0.75$			$\alpha = 1.5$			$\alpha = 3.0$		
n	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$
50	0.1448	-0.0628	1.3740	0.0875	-0.0392	1.8700	-0.1013	-0.0740	1.6296
200	0.0902	0.0099	0.3897	0.0829	0.0107	0.7995	0.0533	0.0042	1.2500
500	0.0157	-0.0027	0.1018	0.0253	0.0038	0.2850	0.0511	0.0081	0.6931
1,500	0.0134	0.0020	0.0371	0.0118	0.0017	0.0811	0.0210	0.0039	0.2578
5,000	0.0040	0.0009	0.0104	-0.0003	0.0009	0.0178	0.0019	0.0004	0.0549

TABLE 4: $\sqrt{MSE}$ of the PHN model percentile estimators.

	$\alpha = 0.75$			$\alpha = 1.5$			$\alpha = 3.0$		
n	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$	$\hat{\xi}$	$\hat{\eta}$	$\hat{\alpha}$
50	1.6300	0.6170	3.5602	1.4147	0.4979	4.3722	1.1907	0.4184	4.5468
200	0.8751	0.2966	1.4306	0.8616	0.2674	2.4950	0.7802	0.2269	3.5839
500	0.5135	0.1781	0.5298	0.5466	0.1699	1.2777	0.5313	0.1515	2.4152
1,500	0.2810	0.0983	0.2516	0.2896	0.0904	0.5403	0.3256	0.0919	1.3219
5,000	0.1553	0.0539	0.1304	0.1580	0.0497	0.2708	0.1728	0.0494	0.6281

5. Illustration

In this illustration we use a dataset related to 1,150 heights measured at 1 micron intervals along a roller drum (i.e. parallel to the roller's axis). This was part of an extensive study of the roller's surface roughness. It is available for download at `http://lib.stat.cmu.edu/jasadata/laslett`.

The data set to illustrate the PHN model has the following summary statistics: mean $\bar{x} = 3.535$ and variance $s^2 = 0.422$. Teh quantities $\sqrt{b_1} = -0.986$ and $b_2 = 4.855$ correspond to sample asymmetry and kurtosis coefficients. According to the asymmetry ($\sqrt{b_1}$) and kurtosis (b_2) values there is a strong evidence that an asymmetric model may provide a better fit for these data. We see that the skewness and kurtosis values are outside the range allowed by the SN and PN models, and even though the kurtosis value is greater than the one found in this paper for the PHN model, the latter may provide a better fit than the SN and PN models.

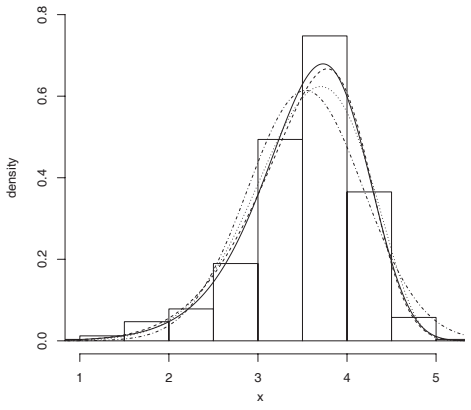
We proceed then to fit the models PN, SN and PHN to the data set. Maximum likelihood estimators for each model are presented in Table (5), with standard errors in parenthesis, obtained by inverting the observed information matrix. The Kolmogorov-Smirnov test rejects the normality assumption (p-value = 0); while the equality hypothesis of the roller variables' mean is not rejected (p-value= 0.1308), which justifies the fitness of the PHN model.

TABLE 5: Parameter estimators (standard error) for N, PN, SN and PHN models.

Estimates	**N**	**PN**	**SN**	**PHN**
$log(lik)$	-1135.866	-1085.241	-1071.362	-1066.994
AIC	2275.488	2176.482	2148.724	2139.988
$\hat{\xi}$	3.5347(0.0191)	4.5495(0.0572)	4.2503(0.0284)	7.0723(0.3194)
$\hat{\eta}$	0.6497(0.0135)	0.1982(0.0279)	0.9694(0.0304)	1.4380(0.0648)
$\hat{\alpha}$	–	0.0479(0.0156	-2.7864(0.2529)	86.8309(28.6166)

To implement model comparison between the models considered above, we use the AIC (Akaike Information Criterion), which penalizes the maximized likelihood function by the excess of model parameters ($AIC = -2\widehat{\ell}(\cdot) + 2k$, where k is the number of parameters in the model), see Akaike (1974).

According to this criterion the model that best fits the data is the one with the lowest AIC. By this criterion the PHN model gives the best fit to the roller data set. Graphs for the fitted models are shown in Figure 6. Figure 7-(a) shows the qqplot calculated with the roller's variable percentiles and the percentile of the PHN variable calculated with the estimates of the parameters, while Figure 7-(b) shows the empirical cumulative distribution functions and the estimated PHN model.



(a)

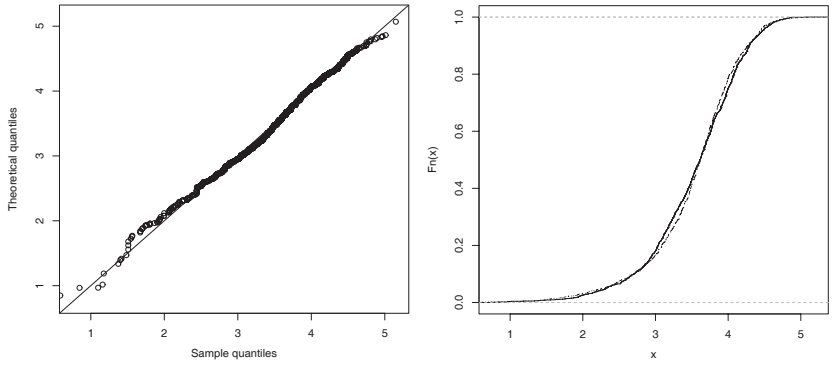
FIGURE 6: Graphs for distributions: $N(3.5347, 0.6497)$ (dashed dotted line), $PN(4.5495, 0.1982, 0.0479)$ (dashed line), $SN(4.2503, 0.9694, -2.7864)$ (dotted line) and $PHN(7.0723, 1.4380, 86.8309)$ (solid line).

We also conducted a hypothesis test to compare the fitness of the normal (N) model versus that of the PHN model. Formally we have the hypotheses

$$H_0 : \alpha = 1 \text{ versus } H_1 : \alpha \neq 1$$

then, using the statistic likelihood of ratio,

$$\Lambda = \frac{\ell_N(\widehat{\xi}, \widehat{\eta})}{\ell_{PHN}(\widehat{\xi}, \widehat{\eta}, \widehat{\alpha})}$$



(a) (b)

FIGURE 7: (a) Q-Qplot roller variable (b) CDF, roller variable (dotted line), PHN variable (solid line).

Substituting the estimated values, we obtain

$$-2\log(\Lambda) = -2(1135.866 - 1066.994) = 137.744$$

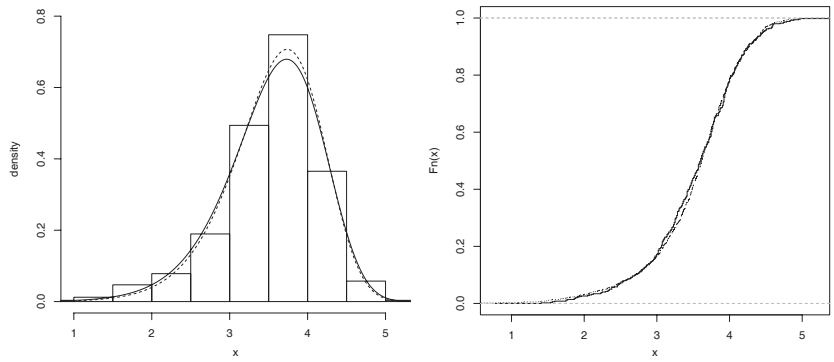
which when compared with the 95% critical value of the $\chi^2_1 = 3.84$ indicate that the null hypotheses is clearly rejected. The PHN model is a good alternative for modelling data.

According to the AIC criterion the PHN model fits the roller data better than the Normal, SN and PN models. So the PHN model captures the asymmetry and kurtosis that the other models fail to capture. A reason for this situation is in the fact that the asymmetry and kurtosis of these particular data are outside the range allowed in the SN and PN models.

We also estimated the model parameters using the two-step percentile method, obtaining: $\widehat{\xi}_p = 6.8219(0.0133)$, $\widehat{\eta}_p = 1.3574(0.0028)$ and $\widehat{\alpha}_p = 75.3801(1.0902)$ (where the estimation errors, in parentheses, were calculated with the Jackknife method). Figure 8-(a) shows the PHN densities from MLE estimation (solid line) and elemental percentile estimation (dash-dot line); Figure 8-(b) shows the corresponding cumulative density functions. Note that this method provides estimates that give a fairly good fit to the PHN model in comparison with the one fitted by maximum likelihood, but the graphs of cumulative distributions give a better fit to the distribution function estimated by maximum likelihood.

6. Concluding Remarks

We have defined a new family of distributions whose hazard function is proportional to hazard function concerning to original distribution function. We discussed several of its properties and provided and estimation of parameters via maximum likelihood, profile likelihood and elemental percentile methods. This is supported



(a) (b)

FIGURE 8: (a) $PHN(\xi, \eta, \alpha)$ density: MLE (solid line), estimation by elemental percentile method (dash-dot line) (b) CDF, roller variable (dotted line), PHN variable from estimation by elemental percentile method (solid line).

with an application to real data in which we show that the PHN model provides consistently better fits than the SN and PN models. The outcome of this practical demonstration shows that the new family is very general, quite flexible and widely applicable.

Acknowledgements

We gratefully acknowledge grants from Universidad de Córdoba (Colombia) and the Mobility Program of the Universidad Industrial de Santander (Colombia). We thank two anonymous referees for helpful comments.

[Recibido: junio de 2012 — Aceptado: abril de 2013]

References

Akaike, H. (1974), 'A new look at statistical model identification', *IEEE Transaction on Automatic Control* **AU–19**, 716–722.

Arellano-Valle, R. B., Gómez, H. W. & Quintana, F. (2004), 'A new class of skew-normal distributions', *Communications in Statistics – Theory and Methods* **33**, 1465–1480.

Arellano-Valle, R. B., Gómez, H. W. & Quintana, F. (2005), 'Statistical inference for a general class of asymmetric distributions', *Journal of Statistical Planning and Inference* **128**, 427–443.

Arnold, B., Gómez, H. & Salinas, H. (2009), 'On multiple constraint skewed models', *Statistics* **43–3**, 279–293.

Azzalini, A. (1985), 'A class of distributions which includes the normal ones', *Scandinavian Journal of Statistics* **12**, 171–178.

Barndorff-Nielsen, O. (1983), 'On a formula for the distribution of the maximum likelihood estimator', *Biometrika* **70**(2), 343–365.

Castillo, E. & Hadi, A. (1995), 'A method for estimating parameters and quantiles of distributions of continuous random variables', *Computational Statistics and Data Analysis* **20**(4), 421–439.

Chiogna, M. (1998), 'Some results on the scalar skew-normal distribution', *Journal of the Italian Statistical Society* **1**, 1–14.

Durrans, S. R. (1992), 'Distributions of fractional order statistics in hydrology', *Water Resources Research* **28–6**, 1649–1655.

Efron, B. (1979), 'Bootstrap methods: another look at the Jackknife', *Annals of Statistics* **7**, 1–26.

Efron, B. (1982), 'The Jackknife, the Bootstrap, and other Resampling Plans', *CBMS 38, SIAM-NSF* .

Efron, B. & Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Eugene, N., Lee, C. & Famoye, F. (2002), 'Beta-normal distribution and its applications', *Communications in Statistics – Theory and Methods* **31**, 497–512.

Farias, R., Moreno, G. & Patriota, A. (2009), 'Reducción de modelos en la presencia de parámetros de perturbación', *Revista Colombiana de Estadística* **32**(1), 99–121.

Fernandez, C. & Steel, M. (1998), 'On Bayesian modeling of fat tails and skewness', *Journal of the American Statistical Association* **93–441**, 359–371.

Gómez, H., Venegas, O. & Bolfarine, H. (2007), 'Skew-symmetric distributions generated by the distribution function of the normal distribution', *Environmetrics* **18**, 395–407.

Gupta, A. K., Chang, F. C. & Huang, W. J. (2002), 'Some skew-symmetric models', *Random Operators Stochastic Equations* **10**, 113–140.

Gupta, D. & Gupta, R. (2008), 'Analyzing skewed data by power normal model', *Test* **17**, 197–210.

Henze, N. (1986), 'A probabilistic representation of the skew-normal distribution', *Scandinavian Journal of Statistics* **13**, 271–275.

Lehmann, E. L. (1953), 'A graphical estimation of mixed Weibull parameter in life testing electron tubes', *Technometrics* **1**, 389–407.

Mudholkar, G. S. & Hutson, A. D. (2000), 'The epsilon-skew-normal distribution for analyzing near-normal data', *Journal of Statistical Planning and Inference* **83**, 291–309.

O'Hagan, A. & Leonard, T. (1976), 'Bayes estimation subject to uncertainty about parameter constraints', *Biometrika* **63**, 201–203.

Pewsey, A. (2000), 'Problems of inference for Azzalini's skewnormal distribution', *Journal of Applied Statistics* **27–7**, 859–870.

Pewsey, A., Gómez, H. & Bolfarine, H. (2012), Likelihood based inference for distributions of fractional order statistics., *in* 'II Jornada Internacional de Probabilidad y Estadística', Pontificia Universidad Católica del Perú.

R Development Core Team, R. (2013), 'R: A language and environment for statistical computing', *R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0* .

Roberts, C. (1966), 'A correlation model useful in the study of twins', *Journal of the American Statistical Association* **61**, 1184–1190.

Sen, P. & Singer, J. (1993), *Large Sample Methods in Statistics: An Introdution with Applications*, Chapman and Hall, New York.

Sen, P., Singer, J. & Pedroso de Lima, A. (2010), *From Finite Sample to Asymptotic Methods in Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

Severini, T. (1998), 'An approximation to the modified profile likelihood function', *Biometrika* **85**, 403–411.

Severini, T. A. (2000), *Likelihood Methods in Statistics*, Oxford University press.

Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 115 a 144

Correspondence Analysis of Contingency Tables with Subpartitions on Rows and Columns

Análisis de correspondencias de tablas de contingencia con subparticiones en filas y columnas

Campo Elías Pardo[1,a], Mónica Bécue-Bertaut[2,b], Jorge Eduardo Ortiz[3,c]

[1]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

[2]Departamento de Estadística e Investigación Operativa, Universidad Politécnica de Cataluña, Barcelona, España

[3]Facultad de Estadística, Universidad Santo Tomás, Bogotá, Colombia

Abstract

We present Intra-Table Correspondence Analysis using two approaches: Correspondence Analysis with respect to a model and Weighted Principal Component Analysis. In addition, we use the relationship between Correspondence Analysis and the Log-Linear Models to provide a deeper insight into the interactions that each Correspondence Analysis describes. We develop in detail the Internal Correspondence Analysis as an Intra-Table Correspondence Analysis in two dimensions and introduce the Intra-blocks Correspondence Analysis. Moreover, we summarize the superimposed representations and give some aids to interpret the graphics associated to the subpartition structures of the table. Finally, the methods presented in this work are illustrated by their application to the standardized public test data collected from Colombian secondary education students in 2008.

***Key words*:** Multidimensional contingency table, Principal component analysis.

Resumen

Para presentar los análisis de correspondencias intra-tablas, se usan los enfoques del análisis de correspondencias con respecto a un modelo y del análisis en componentes principales ponderado. Adicionalmente, se utiliza la relación de los análisis de correspondencias con los modelos log-lineales para entender mejor las interacciones que cada análisis de correspondencias

[a]Associate professor. E-mail: cepardot@unal.edu.co
[b]Professor. E-mail: monica.becue@upc.es
[c]Professor. E-mail: jorgeortiz@usantotomas.edu.co

describe. Se desarrolla de manera detallada el análisis de correspondencias interno como un análisis de correspondencias intra-tablas en dos dimensiones y se introduce el análisis de correspondencias intrabloques. Por otra parte, se resumen las representaciones superpuestas y las ayudas para la interpretación de las gráficas asociadas a la estructura de subparticiones de la tabla. Finalmente, se ilustran los procedimientos con el análisis de una tabla de contingencia construida a partir de los resultados de las pruebas de estado realizadas a los estudiantes de educación media en Colombia en el año 2008.

***Palabras clave*:** análisis en componentes principales, tabla de contingencias multidimensional.

1. Introduction

Contingency tables (CT) with sub-partitions on rows and columns have row and column categories defined from two nested factors. We use $B(A) \times D(C)$ to denote the table structure. The rows are formed by factors A and B, with B categories nested into A categories. In the same way, C and D factors form the columns, with D categories nested into C categories. Each A category defines a *row band* and each C category defines a *column band*. A sub-table crossing a row band with a column band is called a *block*.

The nesting may occur naturally, for example, in a table crossing subregions and economic sub-sectors, where the subregions are aggregated into regions and the economic sub-sectors are aggregated into sectors. In this case, we say that the CT has a "true" sub-partition structure. In other applications, the researcher will choose the variable defining the coarsest partition according to the objectives of the study. For example, the notation *age-group(sex)* indicates that the categories of the variables *sex* and *age-group* are codified interactively. The sex variable defines the partition and the age-group categories are nested into the two categories of sex. A four-way CT with factors A, B, C and D can be flattened into a two-way table in different manners; for example, into the two-way CT denoted by $B(A) \times D(C)$.

We cite hereafter several examples of CT with row and column sub-partitions extracted from the literature:

Hydrobiological studies: $species(taxonomic\ groups) \times places(dates)$, i.e. phaunistic tables, with row-species categorized into taxonomic groups and columns $places \times dates$), being the same places observed at different dates (Cazes, Chessel & Doledec 1988).

Genomics: $sequences(species) \times codons(amino\ acids)$, a CT crossing sequences aggregated into species and codons aggregated into amino acids (Lobry & Chessel 2003, Lobry & Necsulea 2006).

Genetics: $objects(populations) \times aleles(loci)$, a CT with objects split in populations described using alleles clustered into several loci (Laloë, Moazami-Gourdarzi & Chessel 2002).

We aim at presenting different strategies, in the framework of Correspondence Analysis (CA; Lebart, Piron & Morineau 2006, Ramírez & Martínez 2010), to describe contingency tables endowed with sub-partition structures both in rows and columns. Having this objective in mind, we do not discuss inferential methods that might be used to analyze this kind of table.

From a contingency table crossing the row categories $B(A)$ with the column categories $D(C)$, several CA can be performed, depending on the sub-partition structures that are considered. Each CA can be seen as a particular CA with respect to a model, using the generalization of CA proposed by Escofier (1983). This point of view allows us to consider the relationship between Log-Linear Models and Correspondence Analysis applied to the analysis of a two way contingency table. This table is obtained through flattening a four way CT, as described in Van der Heijden (1987).

The structure of the CT, as well as the treatments applied to it, are deduced from the objectives. Dolédec & Chessel (1991) lay out the use of these CA in the environmental sciences.

The first example considers a faunal table in hydrobiology field. The row categories are nested as *species*(*group*). The authors apply Intra-group CA (row bands) and argue both that the specialists have different skills to identify species in each taxonomic group, and that, in such a method, the between-groups variability is eliminated. The Intra-date CA (column bands) shows, more clearly, the associations between species and sites. The Internal Correspondence Analysis (ICA) is both Intra-dates and Intra-groups, as proposed by Cazes et al. (1988) to highlight the species-site associations.

Bécue-Bertaut, Pagès & Pardo (2005) present ICA as a double Intra-Table CA and show that it can be computed either as a CA with respect to a model or as a Weighed Principal Component Analysis. Furthermore, they propose to project on the principal planes issued from this ICA, the "partial" rows ("partial" columns), that is, the rows (columns) as seen from the different points of view corresponding to each group of columns (rows). The superimposed representation of the partial rows (partial columns) is obtained following the same rationale that Multiple Factor Analysis (MFA: Escofier & Pagès (1982); Pagès (2004)). These superimposed representations ease the comparison of the different viewpoints and so enrich the interpretation of the results.

In this paper, the theoretical sections presented by Bécue-Bertaut et al. (2005) are extended and Intra-Block Correspondence Analysis (IBCA) is presented. The resulting methodology is applied to a CT built up from the results of the schools standardized test scores answered by last grade Colombian students in secondary education in 2008. The relationship between CA and Log-Linear Models are used to show the interactions described by the different CA.

§2 defines the notation, taking into account the sub-partition structures of the CT. In §3 we present the different CA as specific cases of both CA with respect to a model and Weighted Principal Component Analysis. The superimposed representations are detailed in §4. The interest of the methodology is shown in §5,

by its application to the schools standardized tests scores in Colombia in 2008. In the Appendix, the demonstrations of some formulae are detailed.

2. Notation

The notation adopted in this work is close to this used by Bécue-Bertaut et al. (2005). Let $B(A) \times D(C)$ be a CT with I rows and K columns. The factors A and C have L and J factors, respectively. The L categories from A are sub-partitioned into $I_1, \ldots, I_l, \ldots, I_L$ categories, respectively; and, similarly, the J categories from C into $K_1, \ldots, K_j, \ldots, K_J$ categories. We use the same symbols to indicate sets and their cardinality. Thus, I is both the set and the number of rows, that is, the categories of $B(A)$; K is both the set and the number of the columns. The categories of $D(C)$; I_l is both the set and number of categories that are nested into the category l from A. From the CT, the relative frequencies table $\mathbf{F}$ is built up. It is structured as shown in Figure 1.

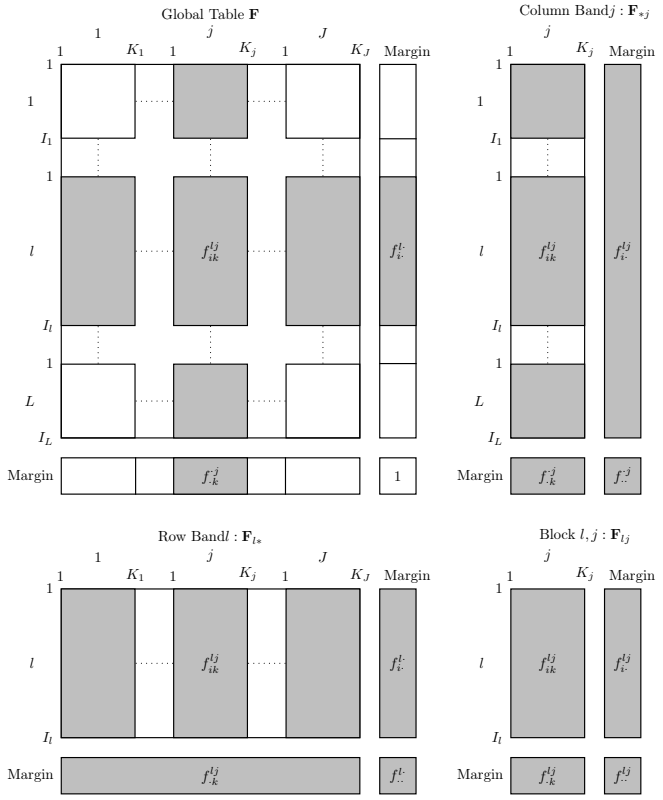


FIGURE 1: Table $\mathbf{F}$ with sub-partition structures in the rows and in the columns.

The general term of $\mathbf{F}$ is noted by f_{ik}^{lj} and its row and column margins by $f_{i\cdot}^{l\cdot}$ and $f_{\cdot k}^{\cdot j}$, respectively. $\mathbf{F}_{l*}$ is the row band l and $\mathbf{F}_{*j}$ the column band j. The

total of the row band $\mathbf{F}_{l*}$ is $f_{..}^{l\cdot} = \sum_{j=1}^{J} f_{..}^{lj}$ and the total of the column band $\mathbf{F}_{*j}$ is $f_{..}^{\cdot j} = \sum_{l=1}^{L} f_{..}^{lj}$.

The block (l, j), noted $\mathbf{F}_{lj}$, has I_l rows and K_j columns. Its row and column margins are $f_{i.}^{lj} = \sum_{k \in K_j} f_{ik}^{lj}$ and $f_{.k}^{lj} = \sum_{i \in I_l} f_{ik}^{lj}$; and its total is $f_{..}^{lj} = \sum_{i \in I_l} \sum_{k \in K_j} f_{ik}^{lj}$.

A cell of $\mathbf{F}$ is identified by the block ^{lj}, as superscript, and the specific cell into the block ${}_{ik}$, as subscript.

$\mathbf{F}$ can be analyzed though the different CA presented in this work: a Simple Correspondence Analysis (SCA); two Intra-Table CA, called here 'analysis in only one dimension': the Intra-Column Bands CA and the Intra-row Bands CA; the Internal Correspondence Analysis (ICA) or double Intra-analysis; the Intra-blocks Correspondence Analysis (IBCA).

To avoid misinterpretations, we use the expression 'Intra-Tables CA', when the structure only concerns one dimension. When the structure concerns the two dimensions, we use the term 'Internal Correspondence Analysis' (ICA) rather than 'Double Intra-Tables CA'. ICA was proposed, with this denomination, by Cazes et al. (1988). Pagès & Bécue-Bertaut (2006) use the term ICA for referring to Intra-Tables CA only in one dimension because, in this case, the two methods are equivalent.

The clouds of points, associated with CA, are noted by using the letter N and a subscript, referring to both the set of points and its cardinality. For example, N_I is the cloud of the I row points and N_{I_l} is the cloud of the I_l points belonging to the row band l.

3. Correspondence Analysis (CA)

We summarize the use of CA to describe a CT endowed with sub-partitions both in rows and columns. Each CA is presented as a Weighted Principal Component Analysis, denoted $PCA(\mathbf{X}, \mathbf{M}, \mathbf{D})$. $\mathbf{X}$ is the data matrix, issued from the original data possibly conveniently transformed; $\mathbf{M}$ is a diagonal matrix corresponding to both the metric in the row space and the column weights. $\mathbf{D}$ is a diagonal matrix corresponding to both the metric in the column space and the row weights. $PCA(\mathbf{X}, \mathbf{M}, \mathbf{D})$ is also called, in French literature, the *general factor analysis* (Lebart, Morineau & Warwick 1984, Escofier & Pagès 1992, Pagès 2004) or *duality diagram* (Cailliez & Pagès 1976, Tenenhaus & Young 1985). This approach emphasizes the geometric point of view of PCA leading to call several statistical measures as in Physics. For example, the barycentre or centroid corresponds to the vector of means, the inertia corresponds to the generalized variance. *Active* and *illustrative* elements are considered; the former are taken into account to compute the principal axes while the latter, if present, are projected on the principal axes previously computed from the active elements.

3.1. Simple Correspondence Analysis (SCA)

SCA describes the residuals of $\mathbf{F}$ with respect to the independence model. The independence model is defined as the product of the marginal terms. SCA applied to $\mathbf{F}$ is also $PCA(\mathbf{X}, \mathbf{M}, \mathbf{D})$ being $\mathbf{X}$ the matrix with the general term:

$$x_{ik}^{lj} = \frac{f_{ik}^{lj} - f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} \tag{1}$$

and $\mathbf{M}$ and $\mathbf{D}$ the matrices:

$$\mathbf{D} = diag(f_{i\cdot}^{l\cdot}) \qquad and \qquad \mathbf{M} = diag(f_{\cdot k}^{\cdot j}) \tag{2}$$

$\mathbf{M}$ (respectively, $\mathbf{D}$) is the metric matrix (matrix of weights) in row (column) space and the matrix of weights (metric matrix) in the column (row) space.

3.1.1. Centroids of the Subclouds as Illustrative Elements

In the row space induced by CA, the cloud N_I can be considered as the union of the L subclouds N_{I_l} formed, each of them by the points belonging to the row band I_l. The weight of the row point (l, i) within the subcloud N_{I_l} is $f_{i\cdot}^{l\cdot}/f_{\cdot\cdot}^{l\cdot}$; thus, the coordinate (j, k) of the centroid of the subcloud N_{I_l} is:

$$\sum_{i \in I_l} \frac{f_{i\cdot}^{l\cdot}}{f_{\cdot\cdot}^{l\cdot}} \left(\frac{f_{ik}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} - 1 \right) = \frac{f_{\cdot k}^{lj}}{f_{\cdot\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} - 1 \tag{3}$$

In the same way, the coordinate (l, i) of the centroid of the subcloud N_{K_j} in the column space is:

$$\sum_{k \in K_j} \frac{f_{\cdot k}^{\cdot j}}{f_{\cdot\cdot}^{\cdot j}} \left(\frac{f_{ik}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} - 1 \right) = \frac{f_{i\cdot}^{lj}}{f_{\cdot\cdot}^{\cdot j} f_{i\cdot}^{l\cdot}} - 1 \tag{4}$$

3.1.2. Inertia Decomposition from the SCA

The partition of the cloud N_K into J subclouds N_{K_j} induces the inertia decomposition into $BetweenInertia + IntraInertia$:

- Between subclouds N_{K_j} Inertia:

$$\sum_{l,i} f_{i\cdot}^{l\cdot} \sum_{j} f_{\cdot\cdot}^{\cdot j} \left(\frac{f_{i\cdot}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}} - 1 \right)^2 = \sum_{l,i} \sum_{j} \frac{\left(f_{i\cdot}^{lj} - f_{i\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j} \right)^2}{f_{i\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}} \tag{5}$$

- Intra subclouds N_{K_j} Inertia:

$$\sum_{l,i} f_{i\cdot}^{l\cdot} \sum_{j} f_{\cdot\cdot}^{\cdot j} \sum_{k \in K_j} \frac{f_{\cdot k}^{\cdot j}}{f_{\cdot\cdot}^{\cdot j}} \left(\frac{f_{ik}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} - \frac{f_{i\cdot}^{lj}}{f_{\cdot\cdot}^{\cdot j} f_{i\cdot}^{l\cdot}} \right)^2 = \sum_{l,i} \sum_{j} \sum_{k \in K_j} \frac{\left(f_{ik}^{lj} - \frac{f_{\cdot k}^{\cdot j} f_{i\cdot}^{lj}}{f_{\cdot\cdot}^{\cdot j}} \right)^2}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} \tag{6}$$

Through exchanging the subscripts i and j, we obtain the decomposition of the inertia of the cloud N_I into between-clouds N_{I_l} inertia and Intra-clouds N_{I_l} inertia.

3.2. Correspondence Analysis with Respect to a Model

Let $\mathbf{A}$ be the model matrix with general term a_{ik}^{lj}, with the same dimensions and margins as $\mathbf{F}$. The CA of $\mathbf{F}$ with respect to the model $\mathbf{A}$, noted $CA(\mathbf{F}, \mathbf{A})$, is equivalent to $PCA(\mathbf{X}, \mathbf{M}, \mathbf{D})$, with $\mathbf{M}$ and $\mathbf{D}$ defined above, in (2), and $\mathbf{X}$ with general term:

$$x_{ik}^{lj} = \frac{f_{ik}^{lj} - a_{ik}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} \tag{7}$$

CA with respect to a model keeps almost all of the properties of the classical CA when the model margins are equal to $\mathbf{F}$ margins (Escofier 1984). This is the case for Intra-Tables CA.

The inertia of both clouds N_I and N_K associated to $CA(\mathbf{F}, \mathbf{A})$ is:

$$Inertia(N_I) = Inertia(N_K) = \sum_{l,j} \sum_{i \in I_l, k \in K_j} \frac{(f_{ik}^{lj} - a_{ik}^{lj})^2}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} \tag{8}$$

The SCA of $\mathbf{F}$ is obtained if the independence model $\mathbf{H} = (f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j})$ is used in the Formula (7).

3.2.1. Decomposition of the Inertia Associated to the SCA when A Model is Considered

Equation (8) is also the *chi-square* distance centered in $\mathbf{H}$ between the conjoint probability distributions $\mathbf{F}$ and $\mathbf{A}$, noted $d^2_{{\chi^2}_H}(\mathbf{F}, \mathbf{A})$ (Cailliez & Pagès 1976, p.449).

It is possible to perform a SCA with respect to model $\mathbf{A}$, denoted $CA(\mathbf{A}, \mathbf{H})$. The associated clouds N_I and N_K have inertia:

$$Inertia(N_I) = Inertia(N_K) = \sum_{l,j} \sum_{i \in I_l, k \in K_j} \frac{(a_{ik}^{lj} - f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j})^2}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} \tag{9}$$

The inertia (9) is also the *chi−square* distance, centered in $\mathbf{H}$, between the conjoint probability distributions $\mathbf{A}$ and $\mathbf{H}$: $d^2_{{\chi^2}_H}(\mathbf{A}, \mathbf{H})$.

If $\mathbf{A}$ and $\mathbf{F}$ have the same margins and

$$\sum_{l,i,j,k} \frac{\left(f_{ik}^{lj} - a_{ik}^{lj}\right) a_{ik}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} = 0 \tag{10}$$

the inertia associated to $CA(\mathbf{F},\mathbf{H})$ is the sum of the inertias associated to $CA(\mathbf{F},\mathbf{A})$ and $CA(\mathbf{A},\mathbf{H})$:

$$d^2_{\chi^2{}_H}(\mathbf{F},\mathbf{H}) = d^2_{\chi^2{}_H}(\mathbf{F},\mathbf{A}) + d^2_{\chi^2{}_H}(\mathbf{A},\mathbf{H}) \tag{11}$$

The demonstration can be found in the Appendix (§Appendix A.1).

In particular, the models associated with CA Intra-bands and ICA, presented hereafter, fulfill the conditions to obtain the inertia decomposition of SCA shown in (11).

3.2.2. Correspondence Analysis and Log-Linear Models

$CA(\mathbf{F},\mathbf{A})$ describes the residuals with respect to model $\mathbf{A}$. Hence, it is possible to perform specific CA to analyze the residuals of a log-linear model or to eliminate some interactions in SCA to better describe the non-eliminated ones (Van der Heijden 1987, Van der Heijden, de Falguerolles & de Leeuw 1989).

The saturated log-linear model associated to a four-way table is:

$$\begin{aligned} ln(\pi^{lj}_{ik}) = u + u_{A(l)} + u_{B(i)} + u_{C(j)} + u_{D(k)} + \\ u_{AB(li)} + u_{CA(lj)} + u_{AD(lk)} + u_{BC(ij)} + u_{BD(ik)} + u_{CD(jk)} + \\ u_{ABC(lij)} + u_{ABD(lik)} + u_{CAD(lkj)} + u_{BCD(ijk)} + u_{ABCD(lijk)} \end{aligned} \tag{12}$$

where π^{lj}_{ik} is the probability of the cell $(.)^{lj}_{ik}$ and the u terms are the model parameters.

If $\mathbf{F}$ (Figure 1) is the "flattened" $B(A) \times D(C)$ of a four-way CT, the independence model $\mathbf{H}$ corresponds to the log-linear model $[AB][CD]$[1] estimation (A and B are jointly independent from C and D). This model is the sum of the four main effects and the first order interactions AB and CD. Then, the CA of $\mathbf{F}$ ($CA(\mathbf{F},\mathbf{H})$) describes the interactions AC, AD, BC, BD and those of superior order.

From a 'true' sub-partition structure, the row factors A and B and the column factors C and D are nested and, therefore, have no interactions between each couple. The saturated model (12) is reduced to:

$$ln(\pi^{lj}_{ik}) = u + u_{A(l)} + u_{B(i)} + u_{C(j)} + u_{D(k)} + u_{CA(lj)} + u_{AD(lk)} + u_{BC(ij)} + u_{BD(ik)} \tag{13}$$

In this case, the $\mathbf{H}$ model represents all the main effects and the SCA is the description of all the interactions in (13).

3.3. Intra-Table Analysis

We denominate *Intra-Row Band/Column Analysis*, the two Intra-Table Analysis that are possible to perform on the $\mathbf{F}$ table. We only summarize the Intra-Column Band Analysis, because the other one can be symmetrically deduced.

[1] With this notation, the model includes the whole interactions between the variables that belong to the same square brackets. For example, the $[AB][C]$ model represents the main effects and the interactions between A and B.

$\mathbf{F}$ is considered as the juxtaposition of the J column bands, as shown by Bécue-Bertaut & Pagès (2004) in the Multiple Factor Analysis of Contingency Tables (MFACT):

$$\mathbf{F} = [\mathbf{F}_{*1} \cdots \mathbf{F}_{*j} \cdots \mathbf{F}_{*J}]$$

The Intra-Column Band CA is the CA of $\mathbf{F}$ with respect to the Intra-Bands Independence Model, denoted $\mathbf{A^J}$, with general term:

$$(a^J)^{lj}_{ik} = \frac{f^{lj}_{i\cdot} f^{\cdot j}_{\cdot k}}{f^{\cdot j}_{\cdot\cdot}} \tag{14}$$

This is the estimation of the log-linear model $[ABC][CD]$ (A and B are jointly independent from D, when C is given). This model includes the interactions AB, AC, BC, CD and ABC; thus, the $CA(\mathbf{F}, \mathbf{A^J})$ describes the interactions, AD, BD, ABD, ACD and $ABCD$. If the subpartition structure is 'true', the $CA(\mathbf{F}, \mathbf{A^J})$ describes the interactions AD and BD (see §3.2.2).

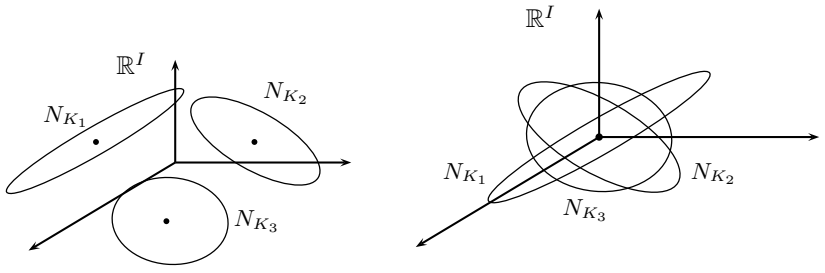
Symmetrically, the Intra-Row Band Independence Model $\mathbf{A^L}$, $[AB][ACD]$ (C and D are jointly independent from B, given A), includes AB, AC, AD, CD and CAD. Thus, the $CA(\mathbf{F}, \mathbf{A^L})$ describes the interactions BC, BD, ABD, ABC and $ABCD$.

The Intra-Column Bands Analysis, $CA(\mathbf{F}, \mathbf{A^J})$, is computed as $PCA(\mathbf{X}, \mathbf{M}, \mathbf{D})$, where $\mathbf{X}$ is the matrix with general term:

$$x^{lj}_{ik} = \frac{f^{lj}_{ik}}{f^{l\cdot}_{i\cdot} f^{\cdot j}_{\cdot k}} - \frac{f^{lj}_{i\cdot}}{f^{l\cdot}_{i\cdot} f^{\cdot j}_{\cdot\cdot}} \tag{15}$$

and $\mathbf{M}$ and $\mathbf{D}$ are metric and weight matrices already defined in (2).

We observe that (15) is equal to (1) - (4): in the Intra-Column Bands CA, the subclouds N_{K_j} in the space $\mathbb{R}^I$ are translated such as their centroids are in the origin. Figure 2a. shows the centroids of the subclouds in SCA and Figure 2b. the same subclouds, but centered in the origin. By centering, the associated inertia to $CA(\mathbf{F}, \mathbf{A^J})$ is the Intra subclouds N_{K_j} inertia from the SCA of $\mathbf{F}$.



a. Subclouds associated to SCA b. Centered subclouds (Intra-Column Bands CA)

FIGURE 2: Subclouds in $\mathbb{R}^I$, associated to the three column bands.

3.3.1. Inertia Decomposition of SCA of F

In the SCA of $\mathbf{F}$, the inertia of the N_K cloud in $\mathbb{R}^I$ can be expressed as the sum of the between and intra-inertias subclouds N_{K_j} obtained replacing $\mathbf{A}$ by $\mathbf{A^J}$ in (11):

$$d^2_{\chi^2}(\mathbf{F},\mathbf{H}) = d^2_{\chi^2}(\mathbf{A^J},\mathbf{H}) + d^2_{\chi^2}(\mathbf{F},\mathbf{A^J}) \tag{16}$$

The two right terms in (16) are associated, respectively, to the following CA (see Appendix Appendix A.2):

- $CA(\mathbf{A^J},\mathbf{H})$, which is also the SCA of the table $\mathbf{T^J}$, with general term $f_{i\cdot}^{lj}$ and dimension $I \times J$.
- $CA(\mathbf{F},\mathbf{A^J})$, which is the Intra-Column Bands CA of $\mathbf{F}$.

3.3.2. Subclouds $N_{I_l} \in \mathbb{R}^K$ from the Intra-Column Bands CA

In the Intra-Column Bands CA it is possible to obtain the centroids of the subclouds $N_{I_l} \in \mathbb{R}^K$ and to project them as illustrative elements. The general term of the coordinate (j,k) of the centroid of the sub cloud N_{I_l} is:

$$\sum_{i\in I_l} \frac{f_{i\cdot}^{l\cdot}}{f_{\cdot\cdot}^{l\cdot}} \left(\frac{f_{ik}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} - \frac{f_{i\cdot}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}} \right) = \frac{f_{\cdot k}^{lj}}{f_{\cdot\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} - \frac{f_{\cdot\cdot}^{lj}}{f_{\cdot\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}} \tag{17}$$

3.4. Internal Correspondence Analysis (ICA)

The Double Intra Bands CA is obtained by centring the subclouds N_{I_l} of the Intra-Column Bands CA. Then, the general term of $\mathbf{X}$ is equal to (15) - (17):

$$x_{ik}^{lj} = \frac{f_{ik}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} - \frac{f_{\cdot k}^{lj}}{f_{\cdot k}^{\cdot j} f_{\cdot\cdot}^{l\cdot}} - \frac{f_{i\cdot}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}} + \frac{f_{\cdot\cdot}^{lj}}{f_{\cdot\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}} \tag{18}$$

The Formula (18) can also been obtained centering the subclouds N_{K_j} in the Intra-Row Bands CA.

The double Intra CA or Internal Correspondence Analysis (ICA) is the $CA(\mathbf{F},\mathbf{C})$, where $\mathbf{C}$ is the model with general term:

$$c_{ik}^{lj} = \frac{f_{\cdot k}^{lj} f_{i\cdot}^{l\cdot}}{f_{\cdot\cdot}^{l\cdot}} + \frac{f_{i\cdot}^{lj} f_{\cdot k}^{\cdot j}}{f_{\cdot\cdot}^{\cdot j}} - \frac{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j} f_{\cdot\cdot}^{lj}}{f_{\cdot\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}} \tag{19}$$

We denote $\mathbf{E}$ the matrix with general term $\dfrac{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j} f_{\cdot\cdot}^{lj}}{f_{\cdot\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}}$, then $\mathbf{C}$ can be written $\mathbf{A^J} + \mathbf{A^L} - \mathbf{E}$ and expressed as:

$$\mathbf{C} = [\mathbf{A^J} - \mathbf{E}] + [\mathbf{A^L} - \mathbf{E}] + \mathbf{E} \tag{20}$$

The inertia of the SCA of $\mathbf{F}$ can be decomposed as follows:

$$d^2_{\chi^2}(\mathbf{F},\mathbf{H}) = d^2_{\chi^2}(\mathbf{E},\mathbf{H}) + d^2_{\chi^2}(\mathbf{A^J},\mathbf{E}) + d^2_{\chi^2}(\mathbf{A^L},\mathbf{E}) + d^2_{\chi^2}(\mathbf{F},\mathbf{C}) \qquad (21)$$

Following Sabatier (1987), the right hand terms in (21) are (see §Appendix A.2 in the Appendix):

- SCA of table $\mathbf{T}$ formed by the sum of the blocks (l, j), with general term $f^{lj}_{..}$ and dimension $L \times J$. This CA describes the interactions AC, i.e. between the factors defining the row and column bands.
- Intra-Tables CA of $\mathbf{T^J}$, with general term $f^{lj}_{i.}$ and dimension $I \times J$. $\mathbf{T^J}$ is a three-way table, since it is the margin of the column bands of $\mathbf{F}$, so factor D disappears. The Intra-Tables CA of $\mathbf{T^J}$ corresponds to the residuals with respect to the model $[AB][AC]$ (B is independed of C, given A). The model contains the interactions AB and AC; thus, the Intra-Tables CA describes the interactions BC and ABC.
- Intra-Tables CA of $\mathbf{T^L}$, with the general term $f^{lj}_{.k}$ and dimension $L \times K$. Table $\mathbf{T^L}$ is the margin of the row bands of $\mathbf{F}$, hence, it is a three way table. This Intra-Table CA describes the interactions AD and ACD, that are the residuals with respect to the model $[AC][CD]$ (A is an independent from D, when C is given).
- ICA of $\mathbf{F}$ ($CA(\mathbf{F},\mathbf{C})$). $\mathbf{C}$ is not the estimation of a log-linear model, its structure is additive instead of multiplicative. Because the four CA contain all of the interactions from the $CA(\mathbf{F})$, the ICA describes the interactions that are not in the three former CA, i.e. BD, ABD, BCD and $ABCD$.

In other words, the SCA of $\mathbf{F}$ is a global analysis that can be decomposed into four CA, where the first order interactions present in the SCA of $\mathbf{F}$, are separated. The inertias associated with the four CA and their relative contributions to the inertia from the SCA are indicators of the importance of these associations.

3.4.1. Intra-Bands CA as Particular Cases of ICA

Intra-Row Bands CA is a particular case of ICA because it can be obtained by considering the L row bands but only one column band with K columns. The Intra-Column Band CA can be obtained considering the J column bands but only one row band with I rows. In the former case, the terms 1 and 3 from (19) cancel one another; in the second case the terms 2 and 3 cancel one another. This justifies the name of "Internal Correspondence Analysis" (ICA) given to one dimension Intra-Tables CA by Pagès & Bécue-Bertaut (2006).

3.4.2. ICA as a Weighted PCA

ICA is the $CA(\mathbf{F},\mathbf{C})$, i.e. the $PCA(\mathbf{X},\mathbf{M},\mathbf{D})$, where $\mathbf{X}$ has the general term given in (18) and $\mathbf{M}$ and $\mathbf{D}$ are defined in (2). In this analysis, the representations in spaces $\mathbb{R}^K$ and $\mathbb{R}^I$ are symmetric: in $\mathbb{R}^K$ the cloud N_I is divided into L

subclouds N_{I_l}; in $\mathbb{R}^I$ the cloud N_K is divided into J subclouds N_{K_j}. Without loss of generality, the properties are presented below in the space $\mathbb{R}^K$.

3.4.3. Row Clouds in $\mathbb{R}^K$

In ICA, the cloud N_I of the I rows is formed by the union of the L subclouds N_{I_l}, each centered in the origin. So, the coordinate of a point (l,i) represents the deviation of the point with respect to the centroid of the subcloud N_{I_l} to which it belongs (Figure 2).

Distances: the square distance between two row points is:

$$d^2[(l,i),(l',i')] = \sum_{j,k} \frac{1}{f_{\cdot k}^{\cdot j}} \left(\frac{f_{ik}^{lj} - c_{ik}^{lj}}{f_{i\cdot}^{l\cdot}} - \frac{f_{i'k}^{l'j} - c_{i'k}^{l'j}}{f_{i'\cdot}^{l'\cdot}} \right)^2 \tag{22}$$

Two points (l,i) and (l',i') are close to one another if their deviations to the respective model, weighted with the inverse of $f_{\cdot k}^{\cdot j}$, are similar for every (j,k). A point (l,i) is located far from the origin if row (l,i) in table **F** differs from the model **C** (Escofier 2003, p. 120).

Transition Formulae: a row coordinate $F_s(l,i)$ on a factorial axis s is a function of the column coordinates $G_s(j,k)$ (see §Appendix A.3):

$$F_s(l,i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \sum_{k \in K_j} \left(\frac{f_{ik}^{lj}}{f_{i\cdot}^{l\cdot}} - \frac{f_{\cdot k}^{lj}}{f_{\cdot\cdot}^{l\cdot}} \right) G_s(j,k) \tag{23}$$

Formula (23) indicates that a row (l,i) lies on the same side that the columns (j,k) whose coordinates are greater than the coordinates of the homologous columns in the l band margin.

Aids to the Interpretation: the contribution to the inertia and the quality of representation on the axes are calculated for each row point. Moreover, aids to the interpretation are defined for each subcloud N_{I_l}:

- Weight subcloud: $f_{\cdot\cdot}^{l\cdot}$.
- Quality of representation on axis s: $Inertia_s(N_{I_l})/Inertia(N_{I_l})$.

  $$Inertia_s(l,i) = f_{i\cdot}^{l\cdot}(F_s(l,i))^2$$

  Therefore:

  $$Inertia_s(N_{I_l}) = \sum_{i \in I_l} f_{i\cdot}^{l\cdot}(F_s(l,i))^2$$

  In $\mathbb{R}^K$ the contribution of a row point to the inertia of cloud N_I is:

  $$Inertia(l,i) = f_{i\cdot}^{l\cdot}\mathbf{x}'_{li}\mathbf{M}\mathbf{x}_{li}$$

  where $\mathbf{x}'_{li}$ is the row (l,i) of $\mathbf{X}$ and $\mathbf{M} = diag(f_{\cdot k}^{\cdot j})$.

- Contribution to axis inertia: the sum of the inertia of the points belonging to the subcloud.

3.5. Intra-blocks Correspondence Analysis (IBCA)

Intra-blocks Correspondence Analysis of $\mathbf{F}$, denoted $IBCA(\mathbf{F})$, is defined as the CA with respect to the Intra-blocks Independence Model $\mathbf{B}$, using the same metrics as the SCA of $\mathbf{F}$. The general term of $\mathbf{B}$ is defined by:

$$b_{ik}^{lj} = \frac{f_{i\cdot}^{lj} f_{\cdot k}^{lj}}{f_{\cdot\cdot}^{lj}} \tag{24}$$

$\mathbf{B}$ is the estimation of the log-linear model $[ABC][ACD]$ (B is independent of D, when AC is given). This model includes the interactions AB, AC, BC, CD, AD, ABC and CAD; thus, the IBCA ($CA(\mathbf{F}, \mathbf{B})$) describes the interactions BD, ABD, BCD and $ABCD$.

If the CT has a 'true' partition structure, the interactions AB, CD and those of superior order including them do not exist. Hence, the model $\mathbf{B}$ contains only the interactions AC, BC and AD and IBCA describes the interactions BD (see §3.2.2).

$IBCA(\mathbf{F})$ is the $PCA(\mathbf{X}, \mathbf{M}, \mathbf{D})$ with:

- $\mathbf{M}$= $diag(f_{\cdot k}^{\cdot j})$
- $\mathbf{D}$= $diag(f_{i\cdot}^{l\cdot})$
- $\mathbf{X}$ with general term given by:

$$x_{ik}^{lj} = \frac{f_{ik}^{lj} - b_{ik}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} = \frac{f_{ik}^{lj} - \dfrac{f_{i\cdot}^{lj} f_{\cdot k}^{lj}}{f_{\cdot\cdot}^{lj}}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} \tag{25}$$

3.5.1. Centered Clouds and Subclouds

The cloud N_I formed by the I points is centered, because the margins of table $\mathbf{F}$ and model $\mathbf{B}$ are equal.

Each subcloud N_{I_l} formed by the I_l points belonging to the row band l are centered, using the weights $\dfrac{f_{i\cdot}^{l\cdot}}{f_{\cdot\cdot}^{l\cdot}}$:

$$\frac{1}{f_{\cdot\cdot}^{l\cdot}} \sum_{i \in \mathcal{I}_l} f_{i\cdot}^{l\cdot} \frac{f_{ik}^{lj} - \dfrac{f_{i\cdot}^{lj} f_{\cdot k}^{lj}}{f_{\cdot\cdot}^{lj}}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} = \frac{1}{f_{\cdot\cdot}^{l\cdot}} \left(\sum_{i \in \mathcal{I}_l} \frac{f_{ik}^{lj}}{f_{\cdot k}^{\cdot j}} - \sum_{i \in \mathcal{I}_l} \frac{f_{i\cdot}^{lj} f_{\cdot k}^{lj}}{f_{\cdot\cdot}^{lj} f_{\cdot k}^{\cdot j}} \right) = \frac{f_{\cdot k}^{lj} - f_{\cdot k}^{lj}}{f_{\cdot\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} = 0$$

3.5.2. Distances

The square distance between two row points is:

$$d^2[(l,i),(l',i')] = \sum_{j,k} \frac{1}{f_{\cdot k}^{\cdot j}} \left(\frac{f_{ik}^{lj} - b_{ik}^{lj}}{f_{i\cdot}^{l\cdot}} - \frac{f_{i'k}^{l'j} - b_{i'k}^{l'j}}{f_{i'\cdot}^{l'\cdot}} \right)^2 \tag{26}$$

Two points (l,i) and (l',i') are close to each other if they similarly differ from the model. Each difference is pondered by $1/f_{\cdot k}^{\cdot j}$. Therefore, a point (l,i) is far from the origin when the row (l,i) of table **F** strongly differs from the model **B** (Escofier 2003, p.120).

3.5.3. Transition Formulae

The formulae allowing the simultaneous representation of row and column points, as well as their interpretation, are:

$$\begin{aligned} F_s(l,i) &= \frac{1}{\sqrt{\lambda_s}} \sum_{j,k} \left(\frac{f_{ik}^{lj} - b_{ik}^{lj}}{f_{i\cdot}^{l\cdot}} \right) G_s(j,k) \ ; \\ G_s(j,k) &= \frac{1}{\sqrt{\lambda_s}} \sum_{l,i} \left(\frac{f_{ik}^{lj} - b_{ik}^{lj}}{f_{\cdot k}^{\cdot j}} \right) F_s(l,i) \end{aligned} \tag{27}$$

Attractions between row and column profiles exist when the observed frequencies are greater than the values in the model.

3.5.4. Aids to the Interpretation

The aids to interpretation used in CA are also available in IBCA, i.e. contribution to the axis inertia and square cosines. Similarly, the aids associated to subclouds N_{K_j} and N_{I_l} are expressed in ICA.

3.5.5. Intra-Blocks CA only in one Dimension

If only one dimension structure is considered, model **B** becomes the intra-row bands independence or the intra-column bands model, depending of the case. Thus, the Intra Bands CA can also be considered as Intra-Blocks Analysis in one single dimension.

IBCA has the advantage of being associated with a log-linear model, while ICA allows us to split the inertia of the clouds associated to SCA in four addends, each corresponding to a CA (see 3.4).

4. Superimposed Representation of the Partial and Global Clouds over a Common Referential

In ICA or IBCA, the global representation of the cloud N_I, in the row space, is obtained considering the whole K coordinates for each row point (l, i). The sub-partition of the columns into J bands also permits to consider each row from the point of view of each J band. Thus, there are J points, denoted $(l, i)^j$ and called *partial points*, considered and projected as illustrative points. The simultaneous projections of global and partial points are denominated superimposed representations.

4.1. Projection of the Partial Clouds

The projections of the partial clouds are defined as done by Pagès (2004) in the frame of multiple factor analysis (MFA).

- Each column j induces the partial cloud $N_I^j \subset \mathbb{R}^{K_j} \subset \mathbb{R}^K = \bigoplus_j \mathbb{R}^{K_j}$, $\mathbf{M}_j$ is the metrics in $\mathbb{R}^{K_j}$ obtained from $\mathbf{M}$, the coordinates of the points N_I^j are the rows of $\mathbf{X}_{*j}$ and the coordinates of these points in $\mathbb{R}^K$ are the rows of the matrix $\widetilde{\mathbf{X}}_{*j}$ defined as:

$$\widetilde{\mathbf{X}}_{*j} = [\mathbf{0} \ \cdots \ \mathbf{0} \ \ \mathbf{X}_{*j} \ \ \mathbf{0} \ \cdots \ \mathbf{0}]$$

- The union of the J partial clouds form the cloud N_I^J with IJ points, that can also be considered as the union of the I clouds $N_{(l,i)}^J$, each with J partial points $(l, i)^j$ belonging to the same row (l, i).

- The inertia of the cloud N_I^J can be expressed as WithinInertia + BetweenInertia subclouds $N_{(l,i)}^J$.

- The cloud of the centroids of the I partial clouds $N_{(l,i)}^J$ is $\frac{1}{J}\sum_j \widetilde{\mathbf{X}}_j$. To force $F_s(i)$ to lie at the centroid of the J partial points $F_s^j(i)$, the rows of $\widetilde{\mathbf{X}}_j$, called partial, are projected as illustrative but dilated by J.

4.2. Restricted Transition Formulae

In (23), each addend j is the restricted formula to the columns K_j belonging to its band. This formula allows us to interpret the position of the partial rows $(l, i)^j$ on the factorial axis s, similarly to the global coordinates:

$$F_s(l,i)^j = \frac{1}{\sqrt{\lambda_s}} \sum_{k \in K_j} \left(\frac{f_{ik}^{lj}}{f_{i\cdot}^{l\cdot}} - \frac{f_{\cdot k}^{lj}}{f_{\cdot\cdot}^{l\cdot}} \right) G_s(j,k) \tag{28}$$

Formula (28) indicates that a row $(l,i)^j$ is placed on the same side that columns $k \in K_j$ whose profile coordinates obtained from the **F** table are greater than the profile coordinates obtained from the margin of its band l. The interpretation of the superimposed representations is mainly supported by these formulae. In the graphic representations, the coordinates are amplified by J.

By exchanging the indices, the restricted transition formulae for the partial columns are deduced.

4.3. Aids to the interpretation of the Partial Clouds

In the superimposed representation, for each factorial axis s there are:

- IJ partial coordinates $F_s^j(l,i)$
- I global coordinates $F_s(l,i)$

These points form different projected clouds:

- I partial clouds $N^J_{(l,i)}$, each with centroid $F_s(l,i)$
- J partial clouds N^j_I
- L clouds N_{I_l}: $\{F_s(l,i);\ i \in I_l\}$.

Since the partial rows $(l,i)^j$ are illustrative they do not contribute to the inertia of the axes. For the partial clouds, the aids to the interpretation are defined as detailed hereafter.

4.3.1. Quality of the Representation of the Partial Clouds

The quality of representation on axis s of each partial cloud N^j_I is computed as the ratio between the projected inertia and the inertia in $\mathbb{R}^K$.

4.3.2. Similarity Measure between Partial Clouds

The total inertia of N^J_I can be decomposed into within and between inertia of clouds $N^J_{(l,i)}$.

The ratio $BetweenInertia/TotalInertia$, computed for each factorial axis s, is a measure of the proximity of the partial points belonging to the same row and therefore of the global similarity between the J partial clouds projected on axis s. If this ratio is close to 1, the homologous points $\{(l,i)^j; j = 1,\ldots,J\}$ are close to each other and the axis s represents a structure common to the different column bands (Pagès 2004, pp.8-9).

4.3.3. Row Contributions to the Within-Inertia

The within-inertia can be decomposed into the contributions of each row, in order to detect differences between the several points of view represented by the column bands. Then, it is possible to identify both the most heterogeneous and homogeneous, in order to interpret the global relations.

It is possible to calculate the contribution to the within-inertia of N_I^J for the cloud associated with a partial row $N_{(l,i)}^J$.

4.4. Zero Partial Points into the Blocks in ICA versus IBCA

Zero Row Inside a Block: in ICA, if the values of the row (l,i) belonging to a column band j of the contingency table are zeros, the partial point does not always lie at the origin. In fact: if $f_{ik}^{lj}=0,\ \ \forall k\in K_j$, then $f_{i\cdot}^{lj}=0$ but the general term of $\mathbf{X}$ (18) for the cells of row (l,i) in column band j is $x_{ik}^{lj}=\dfrac{1}{f_{\cdot\cdot}^{l\cdot}}\left(\dfrac{f_{\cdot\cdot}^{lj}}{f_{\cdot\cdot}^{\cdot j}}-\dfrac{f_{\cdot k}^{lj}}{f_{\cdot k}^{\cdot j}}\right)$, and this term is not necessarily zero.

In this case, the interpretation of the superimposed representations becomes difficult. Some points belonging to null profiles can lie close to points belonging to non-null profiles.

IBCA solves this problem because the partial point associated with a row of zeros lies at the origin: as $f_{ik}^{lj}=0;\ \ \forall k\in K_j$ then $f_{i\cdot}^{lj}=0$, thus the cells of $\mathbf{X}$ (25) belonging to the row (l,i) into the column band j are zeros.

Zero Column Inside a Block: in ICA, if the values of a column (j,k) belonging to a row band l of the contingency table are zeros, the partial point is not at the origin, while in IBCA, it is always at the origin. These results can be obtained by exchanging the indices in the former paragraph.

A Block of Zeros: when all the cells inside block $\mathbf{F}_{lj}$ are zeros, the cells of the model ($\mathbf{C}_{lj}$) inside the block are also zeros; then, the cells of the block $\mathbf{X}_{lj}$ are also zeros. In this block, the cells of model $\mathbf{B}_{lj}$ are not defined, but this problem can be solved defining these cells as zeros.

4.5. Outliers

When few profiles strongly differ from the others, the first axis of the SCA enhance that difference and might hide the differences among the rest of the points. In this case, there are two ways to proceed: 1) to observe the differences on the following axes or 2) to perform the analysis again without the outliers and eventually project them as illustrative elements. These ways to proceed can be used in Intra-Tables CA, ICA and IBCA.

5. Example: Colombia Regional Scores for Secondary Education Standardized Tests

The *Instituto Colombiano para el fomento de la Educación Superior* (ICFES) performs nation-wide secondary education quality assessment based on public standardized tests. Schools are classified into seven levels, according to their scores, ranging from: *very inferior, inferior, low, medium, high, superior* to *very superior*. The first two categories were joined into one named *inferior* and the last three into another named *high*, leaving four levels. Thus, score is a categorical variable with four levels.

To illustrate the application of the methods proposed in the first sections, the schools classification from their scores in the 2008 tests was used, together with the following information:

1. School attendance shifts: full day, morning and afternoon including evening, Saturdays and Sundays;
2. The Colombian administrative system: Colombia is divided into 33 departments, including Bogotá as capital district. The five departments with less than one hundred thousand inhabitants were collapsed to form a "fictitious department" named $P01$, thus leaving 29 departments.
3. Population size: the departments are grouped into 5 categories depending on their population: $P5$ more than two million inhabitants, $P4$ between one and two million, $P3$ between five hundred thousand and one million and $P2$ between one hundred thousand and five hundred thousand. Department $P01$ is included into size-group $P2$.

Our prime objective is the comparison of the departments according to their schools standardized tests scores. The departments are grouped according to their population size, since this variable may hide regional differences. The same rationale leads us to consider the school attendance shifts because, generally, students attending full day tuition present advantages over their peers attending partial shifts.

To achieve the main objective, the contingency table (CT) is structured as $department\ (group) \times score\ (school\ attendance\ shift)$ (Table 1). According to the notations used in the first sections, four factors are considered: A department size-group, B department, C school attendance shift and D score. Since the departments are nested into size-groups, the rows have a "true" sub-partition structure. We have to deal with a CT with $I = 29$ rows and $K = 12$ columns. The 29 rows are the departments divided into $L = 4$ size groups with $I_1 = 7, I_2 = 8, I_3 = 7, I_4 = 7$, according to their population.

The 12 columns correspond to the cross categories of $school\ attendance\ shifts \times scores$. We consider these 12 columns as divided into $J = 3$ groups according to the three school attendance shifts. Each of the 12 blocks corresponds to a subtable with, in rows, the departments of a given size-group and, in columns, the scores of a given school attendance shift.

The profile for the Choco department is an outlier, not considered as active in the analysis. Being located far from the other departments, it is not projected as illustrative. The table shows the high count of Chocó' schools classified into an inferior score. Therefore, the active table has $I = 28$ departments and $I_4 = 6$.

Table 1: Colombian schools classified by departments, school attendance shift and standardized public tests score in 2008.

			SCHOOL ATTENDANCE SHIFT											
			FULL DAY				MORNING				AFTERNOON			
GROUP	COD.	DEPARTMENT	inferior	low	medium	high	inferior	low	medium	high	inferior	low	medium	high
	BOG	Bogotá	5	40	101	309	9	79	219	241	15	171	179	61
	ANT	Antioquia	63	180	116	105	38	105	96	90	125	156	84	29
	VAL	Valle	35	93	72	81	51	140	118	132	62	113	55	19
P5	CUN	Cundi.	19	80	81	103	11	90	114	50	40	84	33	7
	ATL	Atlántico	31	48	22	37	72	62	48	48	106	61	32	15
	SAN	Santander	7	27	51	61	10	40	78	79	30	51	24	20
	BOL	Bolívar	31	31	8	25	90	67	28	30	77	59	14	11
	NAR	Nariño	8	15	21	16	31	50	63	56	21	33	26	10
	COR	Córdoba	18	32	18	11	35	54	16	9	41	38	13	3
	TOL	Tolima	8	19	30	26	28	77	57	28	29	40	21	7
P4	CAU	Cauca	36	56	27	8	24	53	32	24	18	27	13	11
	NSA	NorSantander	6	39	20	23	11	40	37	31	31	20	20	7
	BOY	Boyacá	6	48	74	40	4	31	52	25	14	39	21	8
	MAG	Magdalena	31	19	4	6	58	53	14	12	58	37	6	2
	HUI	Huila	7	37	42	29	1	16	27	12	24	30	14	10
	CAL	Caldas	11	37	26	26	8	38	54	21	10	18	8	2
	CES	César	0	16	8	11	9	50	23	15	36	37	19	3
	RIS	Risaralda	2	14	14	20	0	25	30	24	12	28	14	5
P3	MET	Meta	4	12	15	6	7	45	24	19	22	21	10	7
	SUC	Sucre	9	6	5	3	29	51	19	9	30	28	10	5
	LAG	Guajira	7	6	7	8	12	30	6	7	24	10	3	1
	QUI	Quindio	0	6	7	12	1	17	31	18	10	20	7	1
	CHO	Chocó	23	9	6	1	26	10	3	1	20	5	0	0
	CAQ	Caquetá	8	10	5	1	2	21	11	7	14	9	5	1
	PUT	Putumayo	1	7	10	10	4	6	7	4	4	5	2	0
P2	CAS	Casanare	3	12	10	6	2	16	16	4	8	13	1	1
	ARA	Arauca	3	1	5	2	3	7	12	7	7	5	3	1
	GUV	Guaviare	0	2	2	0	1	4	1	2	0	2	1	0
	PO1	<100 inh.	6	16	7	2	2	7	3	3	3	4	2	0

Department size groups (inhabitants in millions): P5: more than two, P4: between one and two, P3: between 0.5 and one, P2: less than 0.5.

5.1. Simple CA on the Global Table

In the factorial planes, each axis inertia and the corresponding percentage in relation to the global inertia are specified.

The simple CA total inertia is equal to 0.2648. The first three axes retain 84.2%: 0.15 (59.9%), 0.5 (18.7%) and 0.02 (8.6%). The first two axes retain an inertia over than the average. According to the sub-partitions structure associated to the CT, the inertia is decomposed into (see §3.4):

- 0.0062 (2.3%) department group - school attendance shift association;
- 0.0442 (16.7%) department group - score and department group - performance - school attendance shift associations;
- 0.0281 (10.6%) department - school attendance shift association;
- 0.1863 (70.4%) department - score and department - performance - school attendance shift associations.

Figure 3a. shows the representation of the departments, and the size-groups as illustrative, on the first factorial plane issued from simple CA applied to the table crossing departments and *scores* × *school attendance shifts*. Figure 3b. shows the representation of the *scores* × *school attendance shifts* and the *school attendance shifts* as illustrative, on the same plane. A Guttman (parabola) effect is observed, more tidy in the scores trajectories corresponding to full day and morning shifts. The scores are sorted out on the first axis. The second axis opposes full day shift (on the positive part) to morning and afternoon shifts (on the negative part). Since the sub-clouds are not centered on their own centroid, the second axis opposes departments with a high proportion of full day attendance schools to departments with a high proportion of morning or afternoon attendance shifts. The former are mostly concentrated in the less populated size-groups of departments. This opposition is of no interest in the context of our study.

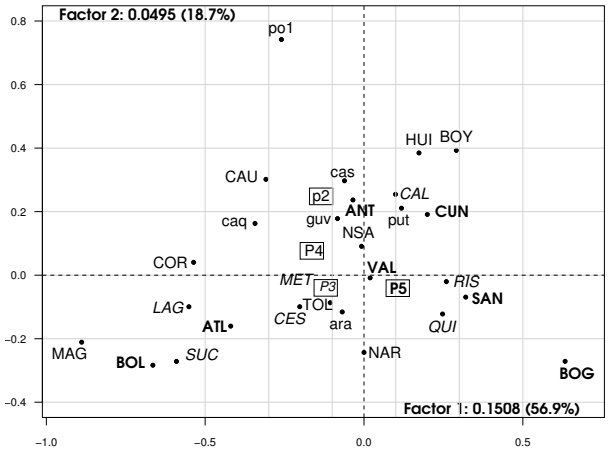
5.2. Intra-School-Attendance-Shift Analysis

The Intra-school-attendance-shift CA allows for removing the variability due to the different profiles of the school-attendance-shifts from one department to another. In this analysis, the inertia is equal to 0.2143 (80.9% of the simple CA's inertia), corresponding basically to the relationship between departments and scores. The first factorial plane retains 83.4% of the inertia and, according to the eigenvalue structure, well synthesizes the results of this analysis.

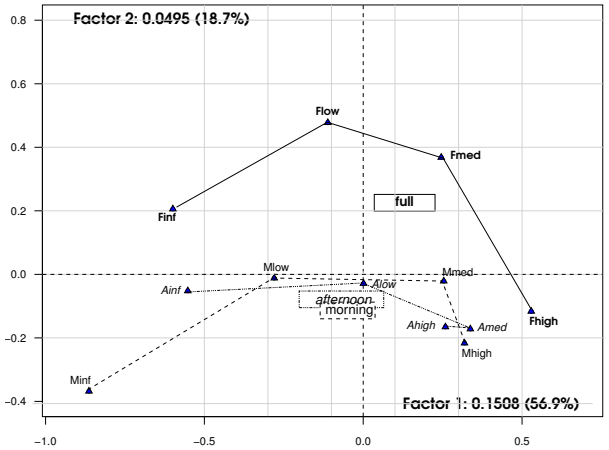
This analysis differs from the simple CA because of the re-centering of the school attendance-shift clouds so that their centroids coincide with the global centroid. In this latter analysis, the departments are more clearly sorted out depending on their schools scores because the interaction between departments and school attendance shifts has been eliminated (Figure 4). The Guttman effect also becomes clearer with this recentering. The departments' scores seem to be linked to their population size. The ICA and the IBCA will allow us compare the scores of the size-groups from an internal point of view.

5.3. ICA and IBCA

The ICA and IBCA results are very similar with total inertias equal to 0.1863 and 0.1856, respectively (70.4 % and 70.1 % of the simple CA's inertia). IBCA main results are described, since they allow for better superimposed representations (§4.4).



a. Departments and centroids of the size-groups: P_5, P_4, P_3 y P_2 as illustrative.



b. *Scores* × *school attendance shifts* and the *school attendance shifts* as illustrative.

FIGURE 3: First factorial plane of the Simple Correspondence Analysis.

Two axes (75.6% of inertia) are retained in the IBCA. Table 2 presents the aids to interpret sub-clouds for both school attendance and department groups. The influence of the sub-clouds in the analysis depends on the weight of each band which is proportional to the percentage of schools that they contain (weight).

Figure 5 shows the simultaneous global representation of rows and columns on the IBCA first factorial plane. This graph synthesizes the CT's analysis. The departments are sorted out by scores issued from the public standardized tests, along a parabola. Bogotá obtained the best scores while Bolivar and Magdalena obtained the worst. Results show that the departments of Bolivar, Magdalena, Atlántico, La Guajira, Sucre and Cordoba, all from the Caribbean Region, obtained inferior results. Cordoba stands out in this region because it has a greater

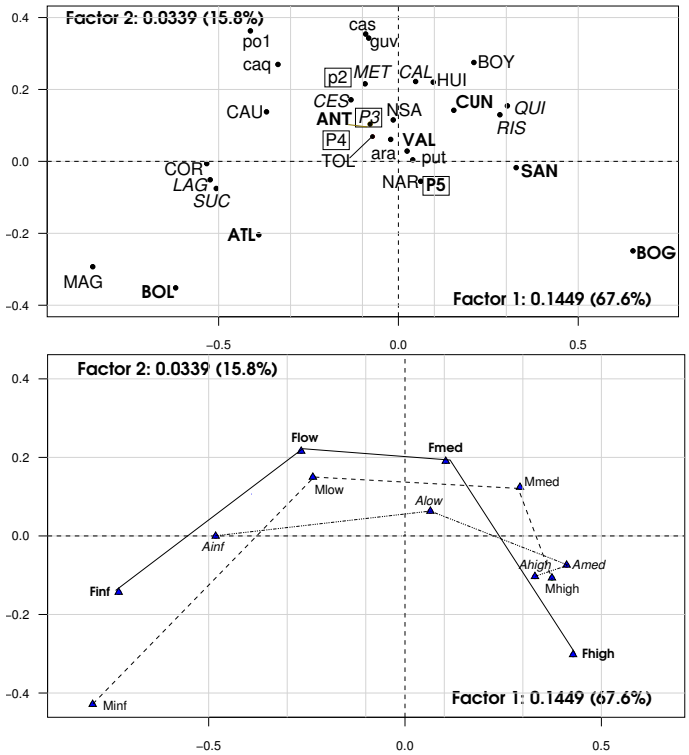


FIGURE 4: First factorial plane of the Intra-school-attendance-shift CA.

TABLE 2: Aids to the interpretation of row and column bands in the IBCA of schools

A. School attendance (column bands)

	Total		Comp1			Comp2			Plane	
	Cont.Inertia		Cont.Inertia		Cali	Cont.Inertia		Cali	Cali	Weight
School day	x10000	%	x10000	%	dad %	x10000	%	dad %	dad %	%
Full	636	34.3	376	29.0	59.2	153	60.3	24.0	83.2	30.5
Morning	806	43.4	611	47.0	75.9	90	35.5	11.2	87.0	40.5
Afternoon	414	22.3	312	24.0	75.3	11	4.2	2.6	77.9	29.0
Total	1856	100.0	1299	100.0		254	100.0			100.0

B. Department groups (row bands)

	Total		Comp1			Comp2			Plane	
	Cont.Inertia		Cont.Inertia		Cali	Cont.Inertia		Cali	Cali	Weight
Group	x10000	%	x10000	%	dad %	x10000	%	dad %	dad %	%
P5	1162	62.6	872	67.1	75.0	197	77.7	17.0	92.0	58.0
P4	459	24.7	310	23.9	67.5	33	12.9	7.1	74.7	25.2
P3	179	9.6	108	8.3	60.6	14	5.6	7.9	68.4	13.1
P2	56	3.0	9	0.7	16.1	10	3.8	17.4	33.5	3.7
Total	1856	100.0	1299	100.0		254	100.0			100.0

percentage of schools in medium levels. As a rule, the most standing out departments belong to the Andean region. Among the less populated size-group of departments, Arauca (Llanos Orientales) and Putumayo (Amazonia) stand out.

One hypothesis is that these last two departments have succeeded in transferring part of the oil production royalties to the educational system.

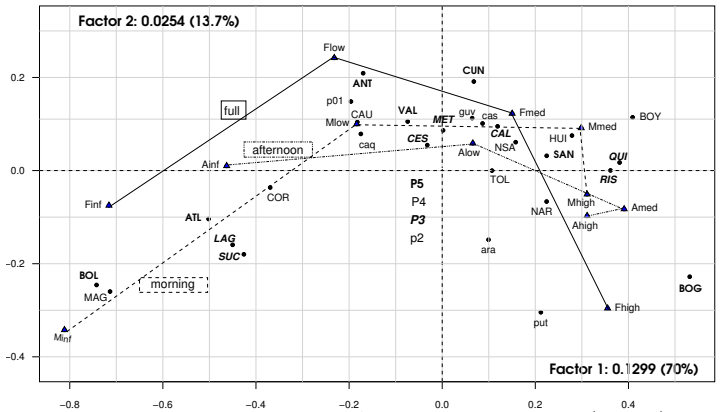


FIGURE 5: First factorial plane of the IBCA for CT $department(group) \times score(shift)$.

5.4. Superimposed Representations in the IBCA

The first axis is a dispersion axis common to the three partial clouds of departments (one for each school attendance shift; ratio BetweenInertia/WithinInertia = 81.1%). Concerning the four score clouds (one for each group of departments, ratio BetweenInertia/WithinInertia = 46.9%), this axis corresponds both to common and specific effects of the clouds. On the second axis, only one third of the inertia corresponds to a common effect to both shift (ratio = 29.9%) and groups of departemts (ratio = 28.8%) (see §4.3.2).

5.4.1. Departments

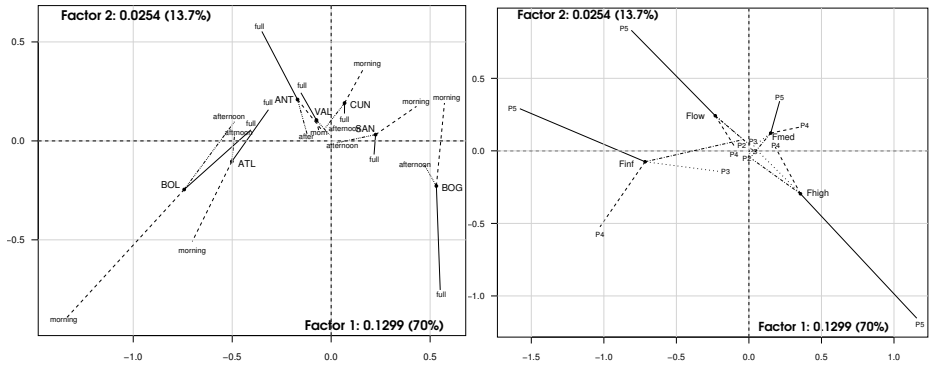
Three partial points correspond to each department in the superimposed representations, one for each shift. The global point is the average of those three points. Figure 6a. shows this representation for the most populated department group.

Bolivar is the partial cloud with the greatest dispersion (0.0028 of within inertia), morning shift has the lowest average score among the departments of the most populated size-group. Table 3 shows that 41.9% of the schools with morning shifts are classified into the inferior category, more than three times the average of the most populated departments (12.6%). In Bolivar, the percentages of these schools classified in medium and high categories are less than half the average. Bogotá has obtained the best results in all three shifts, being the full day shift the most outstanding with 67.9% of its schools in high level, over the average equal to 38.7%. Bogotá and Santander are more alike from the morning shift point of view than for the full day shift. Atlántico and Bolivar's scores are similar in full day

and afternoon shifts but they differ from the morning shift, with better scores for Atlántico.

5.4.2. Columns: $score(shift)$

Each global point represents one shift and one score category. The global point is the centroid of four partial points (one for each group of departments). Figure 6b. shows the superimposed representation of the score categories corresponding to full day shift (see §4). In this shift, the scores differ mostly from the most populated departments ($P5$ and $P4$).



a. The most populated department group b. Full day shift

FIGURE 6: Superimposed representation on the IBCA first factorial plane.

5.5. Separate CA for the 12 Blocks

ICBA can be considered as a comparison method for the profiles corresponding to each block. For a given block, the Intra-blocks independence model is the independence model of the table considered separately. The changes in the IBCA, in relation to the separate CA of each block, concerning metrics and weights, are the price that must be paid to have a common reference framework. Figure 7 shows some of the 12 separate correspondence analyses. The axes have been rotated for an easier comparison to one another and to IBCA.

For instance, the planes corresponding to the three shifts in the most populated departments ($P5$) show similar trends that are kept in IBCA.

6. Conclusions

Various correspondence analyses, useful for the description of contingency tables with sub-partition structures both in rows and columns were demonstrated. An extension concerning the theoretical sections presented in Bécue-Bertaut et al. (2005), was made, putting emphasis on the Double Intra Correspondence Analysis

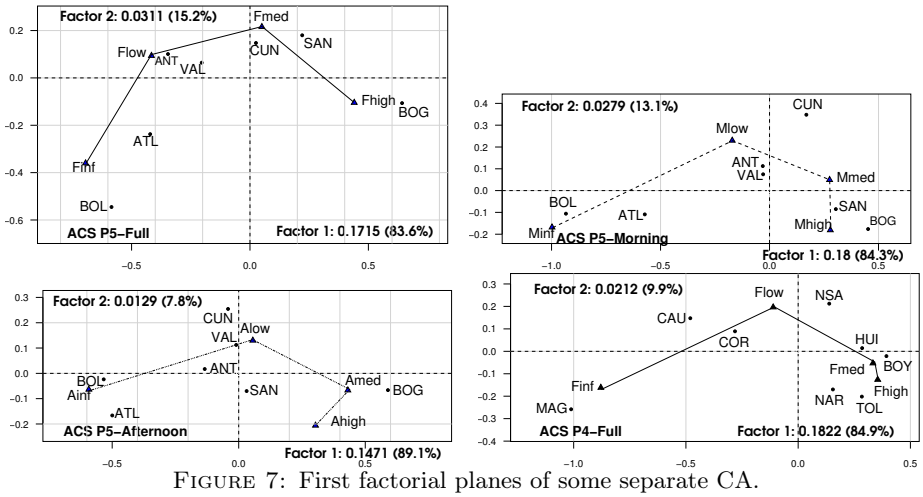


FIGURE 7: First factorial planes of some separate CA.

or Internal Correspondence Analysis (ICA) (§3.4). The Intra-Block Correspondence Analysis (IBCA) was proposed (§3.5). The Intra-Row Bands and Intra-Column-Bands CA are particular cases of ICA and IBCA, whenever the sub-partition structure is considered in one single dimension (§3.4.1 and §3.5.5).

The decomposition of a simple CA for a CT with sub-partition structure in rows and columns, presented by Sabatier (1987), was demonstrated, based on the inertia decomposition of the simple CA into four addends. Thus, the four CA to each addend were derived (§3.1.2 and Appendix A.2).

The relation between correspondence analysis and log-linear models applied to multiple ways contingency tables were used to show the interactions described by the ICA and IBCA (§3.2.2).

In the superimposed representation of ICA, some points belonging to null profiles, inside a block, which can lie close to points belonging to non-null profiles. IBCA solves this problem because the partial points belonging to null profiles which always lay at the origin (§4.4).

The methods presented in this work were illustrated by their application to the standardized public test data collected from the Colombian secondary education students in 2008. IBCA provides, in a synthesized way, the regional differences between departments regarding the schools scores. The superimposed representations allow us to compare, on the one hand, the departments scores through the different shifts and, on the other hand, the score-shifts through the groups of departments (§5).

Acknowledgements

We thank both anonymous judges and the Editor in Chief for carefully reading the article and for their valuable feedback which contributed to its improvement;

Professor Jerôme Pagès for his comments about the superimposed representations; D. Castin for her tidy revision of the English version. This document is derived from the doctoral thesis in Statistics presented at the Faculty of Science of the Universidad Nacional de Colombia Sede Bogotá (Pardo 2011), directed and co-directed by the second and third co-authors respectively. The research was funded by the Universidad Nacional de Colombia through a study grant. Computation was done using pamctdp developed in R (R Development Core Team 2010) using functions of the ade4 package (Thioulouse, Chessel, Dolédec & Olivier 1997). FactoMineR (Husson, Josse, Le & Mazet 2009) and FactoClass (Pardo & DelCampo 2007) packages were used for parts of the graphic functions.

[Recibido: junio de 2011 — Aceptado: mayo de 2013]

References

Bécue-Bertaut, M. & Pagès, J. (2004), 'A principal axes method for comparing multiple contingency tables: MFACT.', *Computational Statistics & Data Analysis* **45**(3), 481–503.

Bécue-Bertaut, M., Pagès, J. & Pardo, C. (2005), Contingency table with a double partition on rows and columns. Visualization and comparison of the partial and global structures, *in* J. Janssen & P. Lenca, eds, 'Proceedings ASMDA 2005', Applied Stochastic Models and Data Analysis. Brest, France. May,17-20, 2005, ENST Bretagne, pp. 355–364.
*http://conferences.telecom-bretagne.eu/asmda2005/IMG/pdf/proceedings/355.pdf

Cailliez, F. & Pagès, J. (1976), *Introduction à l'Analyse des Données*, Smash, Paris.

Cazes, P., Chessel, D. & Doledec, S. (1988), 'L'analyse des correspondances internes d'un tableau partitionné. Son usage en hydrobiologie', *Revue de Statistique Appliquée* **36**(1), 39–54.

Dolédec, S. & Chessel, D. (1991), 'Recent developments in linear ordination methods for environmental sciences', *Advances in Ecology* **1**, 133–155.

Escofier, B. (1983), Generalisation de l'analyse des correspondances a la comparaison de tableaux de frequence, Rapports de Recherche 207, Institut National de Recherche en Informatique et en Automatique, Centre de Rennes. IRISA. *http://hal.inria.fr/inria-00076351

Escofier, B. (1984), 'Analyse factorielle en référence à un modèle. Application à l'analyse de tableaux d'échanges', *Revue de Statistique Appliquée* **32**(4), 25–36.

Escofier, B. (2003), *Analyse des Correspondances. Recherches au Coeur de l'Analyse des Données*, Presses Universitaires de Rennes - Société Francaise de Statistique, Rennes, France.

Escofier, B. & Pagès, J. (1982), Comparaison de groupes de variables définies sur le même ensemble d'individus, Rapports de Recherche 149, INRIA-IRISA, Rennes, France.
*http://hal.inria.fr/inria-00076411

Escofier, B. & Pagès, J. (1992), *Análisis Factoriales Simples y Múltiples. Objetivos, Métodos e Interpretación*, Universidad del País Vasco, Bilbao.

Husson, F., Josse, J., Le, S. & Mazet, J. (2009), *FactoMineR: Factor Analysis and Data Mining with R*. R package version 1.12.
*http://CRAN.R-project.org/package=FactoMineR

Laloë, D., Moazami-Gourdarzi, K. & Chessel, D. (2002), Contribution of individual markers to the analysis of the relationships among breeds by correspondence analysis, *in* '7th World Congress on Genetics Applied to Livestock Production', Montpellier, France.
*http://pbil.univ-lyon1.fr/R/articles/arti110.pdf

Lebart, L., Morineau, A. & Warwick (1984), *Multivariate Descriptive Statistical Analysis*, Wiley, New York.

Lebart, L., Piron, M. & Morineau, A. (2006), *Statistique exploratoire multidimensionnelle. Visualisation et inférence en fouilles de données*, 4 edn, Dunod, Paris.

Lobry, J. & Necsulea, A. (2006), 'Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes', *Gene* **385**, 128 – 136.

Lobry, J. R. & Chessel, D. (2003), 'Internal Correspondence Analysis of Codon and Amino-Acid Usage in Thermophilic Bacteria', *Journal of Applied Genetics* **44**(2), 235–261.

Pagès, J. (2004), 'Multiple Factor Analysis: Main Features and Application to Sensory Data', *Revista Colombiana de Estadística* **27**(1), 1–26.

Pagès, J. & Bécue-Bertaut, M. (2006), Multiple Factor Analysis for Contingency Tables, *in* M. Greenacre & J. Blasius, eds, 'Multiple Correspondence Analysis and Related Methods', Chapman and Hall/CRC, chapter 13, pp. 299–326.

Pardo, C. & DelCampo, P. (2007), 'Combinacion de metodos factoriales y de analisis de conglomerados en r: el paquete factoclass', *Revista Colombiana de Estadistica* **30**(2), 231–245.
*www.matematicas.unal.edu.co/revcoles

Pardo, C. E. (2011), Métodos en ejes principales para tablas de contingencia con estructuras de partición en filas y columnas, Tesis para optar al título de Doctor en Ciencias-Estadística, Universidad Nacional de Colombia. Facultad de Ciencias. Departamento de Estadística, Bogotá.

R Development Core Team (2010), *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/

Ramírez, J. R. & Martínez, G. (2010), 'Analisis de correspondencia a partir de una muestra peobabilistica', *Revista Colombiana de Estadística* **33**, 273–293.

Sabatier, R. (1987), Methodes factorielles en analyse des données: aproximations et prise en compte de variables concomitantes, Doctorat d'Etat, Universite des Sciences et Techniques du Languedoc, Montpeiller.

Sabatier, R., Lebreton, J. & Chessel, D. (1989), Principal Component Analysis with Instrumental Variables as a Tool for Modelling Composition Data, *in* R. Coppi & S. Bolasco, eds, 'Multiway Data Analysis', Elsevier, Amsterdam, pp. 341–350.

Tenenhaus, M. & Young, F. (1985), 'An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Homogeneity Analysis and other Methods for Quantifying Categorical Multivariate Data', *Psychometrika* **50**(1), 91–119.

Thioulouse, J., Chessel, D., Dolédec, S. & Olivier, J. (1997), 'ADE-4: a multivariate analysis and graphical display software', *Statistical and Computing* **7**, 75–83.
*http://pbil.univ-lyon1.fr/ADE-4/ADE-4F.html

Van der Heijden, P. (1987), *Correspondence Analysis of Longitudinal Categorical Data*, DSWO Press, Leiden.

Van der Heijden, P., de Falguerolles, A. & de Leeuw, J. (1989), 'A combined approach to contingency table analysis using correspondence analysis and log-linear analysis', *Applied Statistics* pp. 249–292.

Appendix A. Some Proofs

Appendix A.1. Proofs of Formulae (11) and (10), p.122

The CA inertia can be decomposed as:

$$\sum_{l,i,j,k} \frac{\left((f_{ik}^{lj} - a_{ik}^{lj}) + (a_{ik}^{lj} - f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j})\right)^2}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} = \sum_{l,i,j,k} \frac{(f_{ik}^{lj} - a_{ik}^{lj})^2}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} + \sum_{l,i,j,k} \frac{(a_{ik}^{lj} - f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j})^2}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} +$$

$$+ \sum_{l,i,j,k} \frac{(f_{ik}^{lj} - a_{ik}^{lj})(a_{ik}^{lj} - f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j})}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}}$$

Then (11) is true if the last addend is equal to zero. We have:

$$\sum_{l,i,j,k} \frac{(f_{ik}^{lj} - a_{ik}^{lj})a_{ik}^{lj} - (f_{ik}^{lj} - a_{ik}^{lj})f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} = \sum_{l,i,j,k} \frac{(f_{ik}^{lj} - a_{ik}^{lj})a_{ik}^{lj}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} - \sum_{l,i,j,k} (f_{ik}^{lj} - a_{ik}^{lj})$$

The last term is zero because the totals of $\mathbf{F}$ and $\mathbf{A}$ are equal to one another.

Appendix A.2. Decomposition of the Inertia Associated to the SCA

Sabatier, Lebreton & Chessel (1989) demonstrated the decomposition of the inertia using the correspondence analysis with respect to instrumental variables. We show the inertia decomposition by expressing $\mathbf{F} - \mathbf{H}$ as the sum of differences and using these differences expressed as general terms to calculate the inertia. Then, each inertia term is associated to a CA.

$$\mathbf{F} - \mathbf{H} = (\mathbf{F} - \mathbf{C}) + (\mathbf{A^J} - \mathbf{E}) + (\mathbf{A^L} - \mathbf{E}) + (\mathbf{E} - \mathbf{H})$$

The inertia associated to the SCA of $\mathbf{F}$ is:

$$\sum_{l,i,j,k} \frac{(f_{ik}^{lj} - h_{ik}^{lj})^2}{h_{ik}^{lj}} = \frac{\left((f_{ik}^{lj} - c_{ik}^{lj}) + ((a^J)_{ik}^{lj} - e_{ik}^{lj}) + ((a^L)_{ik}^{lj} - e_{ik}^{lj}) + (e_{ik}^{lj} - h_{ik}^{lj})\right)^2}{h_{ik}^{lj}}$$

$$= \sum_{l,i,j,k} \frac{(f_{ik}^{lj} - c_{ik}^{lj})^2}{h_{ik}^{lj}} + \sum_{l,i,j,k} \frac{((a^J)_{ik}^{lj} - e_{ik}^{lj})^2}{h_{ik}^{lj}} + \sum_{l,i,j,k} \frac{((a^L)_{ik}^{lj} - e_{ik}^{lj})^2}{h_{ik}^{lj}} + \sum_{l,i,j,k} \frac{(e_{ik}^{lj} - h_{ik}^{lj})^2}{h_{ik}^{lj}} \tag{29}$$

since all the crossed products are equal to zero. In what follows, the equality to zero for the last crossed product is proved:

$$\sum_{l,i,j,k} \frac{((a^L)_{ik}^{lj} - e_{ik}^{lj})(e_{ik}^{lj} - h_{ik}^{lj})}{h_{ik}^{lj}} = \sum_{l,i,j,k} \frac{((a^L)_{ik}^{lj} - e_{ik}^{lj})e_{ik}^{lj}}{h_{ik}^{lj}} - \sum_{l,i,j,k} \frac{((a^L)_{ik}^{lj} - e_{ik}^{lj})h_{ik}^{lj}}{h_{ik}^{lj}}$$

The last term is zero because the totals of $\mathbf{A^L}$ and $\mathbf{E}$ are both zero. The other term is also zero, since:

$$\sum_{l,i,j,k} \frac{((a^L)_{ik}^{lj} - e_{ik}^{lj})\frac{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j} f_{\cdot\cdot}^{lj}}{f_{\cdot\cdot}^{\cdot j} f_{\cdot\cdot}^{l\cdot}}}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} = \sum_{l,j} \frac{f_{\cdot\cdot}^{l\cdot}}{f_{\cdot\cdot}^{\cdot j} f_{\cdot\cdot}^{l\cdot}} \sum_{i,k} ((a^L)_{ik}^{lj} - e_{ik}^{lj}) = 0$$

Now, the CA associated to the inertias contained in the formula can be seen (29):

1. $CA(\mathbf{F}, \mathbf{C})$, i.e, $ICA(\mathbf{F})$.

2. $CA(\mathbf{A^J}, \mathbf{E})$, but the expression can be reduced adding over k:

$$\sum_{l,i,j,k} \frac{\left(\frac{f_{i\cdot}^{lj} f_{\cdot k}^{\cdot j}}{f_{\cdot\cdot}^{\cdot j}} - \frac{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j} f_{\cdot\cdot}^{lj}}{f_{\cdot\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}}\right)^2}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} = \sum_{l,i,j} \frac{\left(f_{i\cdot}^{lj} - \frac{f_{i\cdot}^{l\cdot} f_{\cdot\cdot}^{lj}}{f_{i\cdot}^{l\cdot}}\right)^2}{f_{i\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}}$$

   We note $\mathbf{T^J}$ the table of dimension $I \times J$ and with general term $f_{i\cdot}^{lj}$, this inertia is associated to the Intra-Tables CA of $\mathbf{T^J}$.

3. We can obtain an analogous result if we add on the subscript i, i.e. the $CA(\mathbf{A^L}, \mathbf{E})$ is the Intra-Tables CA of $\mathbf{T^L}$, with dimension $L \times K$ and general term $f_{\cdot k}^{lj}$.

4. The last addend is associated to the $CA(\mathbf{E}, \mathbf{H})$. In this case, it is possible to add to both subscripts i and k:

$$\sum_{l,i,j,k} \frac{\left(\frac{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j} f_{\cdot\cdot}^{lj}}{f_{\cdot\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}} - f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}\right)^2}{f_{i\cdot}^{l\cdot} f_{\cdot k}^{\cdot j}} = \sum_{l,j} \frac{(f_{\cdot\cdot}^{lj} - f_{\cdot\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j})^2}{f_{\cdot\cdot}^{l\cdot} f_{\cdot\cdot}^{\cdot j}}$$

   This inertia is associated to the SCA of $\mathbf{T}$, with dimension $L \times J$ and general term $f_{\cdot\cdot}^{lj}$, i.e. the table formed by the totals of the blocks (l, j).

Appendix A.3. Proof of Formula (23), p.126

The coordinate of the row point over the s-axis, as a function of the coordinates of the column points is (Escofier 1984):

$$F_s(l,i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j,k} \left(\frac{f_{ik}^{lj}}{f_{i\cdot}^{l\cdot}} - \frac{c_{ik}^{lj}}{f_{i\cdot}^{l\cdot}}\right) G_s(j,k)$$

Replacing c_{ik}^{lj} (Formula (19), p.124), three sums appear but the two last are zero, because the coordinates $G_s(j,k)$ from each subcloud N_{K_j} are centered with the weights $\frac{f_{\cdot k}^{\cdot j}}{f_{\cdot\cdot}^{\cdot j}}$:

$$\sum_j f_{i\cdot}^{lj} \sum_{k \in K_j} \frac{f_{\cdot k}^{\cdot j}}{f_{\cdot\cdot}^{\cdot j}} G_s(j,k) = 0 \quad \textit{and} \quad \sum_j \frac{f_{\cdot\cdot}^{lj}}{f_{\cdot\cdot}^{l\cdot}} \sum_{k \in K_j} \frac{f_{\cdot k}^{\cdot j}}{f_{\cdot\cdot}^{\cdot j}} G_s(j,k) = 0$$

then,

$$F_s(l,i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j,k} \left(\frac{f_{ik}^{lj}}{f_{i\cdot}^{l\cdot}} - \frac{f_{\cdot k}^{lj}}{f_{\cdot\cdot}^{l\cdot}}\right) G_s(j,k)$$

Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 145 a 152

Improved Exponential Type Ratio Estimator of Population Variance

Estimador tipo razón exponencial mejorado para la varianza poblacional

Subhash Kumar Yadav[1,a], Cem Kadilar[2,b]

[1]Department of Mathematics & Statistics (A Centre of Excellence), Dr. Rml Avadh University, Faizabad, India

[2]Department of Statistics, Hacettepe University, Ankara, Turkey

Abstract

This article considers the problem of estimating the population variance using auxiliary information. An improved version of Singh's exponential type ratio estimator has been proposed and its properties have been studied under large sample approximation. It is shown that the proposed exponential type ratio estimator is more efficient than that considered by the Singh estimator, conventional ratio estimator and the usual unbiased estimator under some realistic conditions. An empirical study has been carried out to judge the merits of the suggested estimator over others.

***Key words*:** Auxiliary variable, Bias, Efficiency, Mean squared error.

Resumen

Este artículo considera el problema de estimar la varianza poblacional usando información auxiliar. Una versión mejorada de un estimador exponencial tipo razón de Singh ha sido propuesta y sus propiedades han sido estudiadas bajo aproximaciones de grandes muestras. Se muestra que el estimador exponencial tipo razón propuesto es más eficiente que el estimador de Singh, el estimador de razón convencional y el estimador insesgado usual bajo algunas condiciones realísticas. Un estudio empírico se ha llevado a cabo con el fin de juzgar los méritos del estimador sugerido sobre otros disponibles.

***Palabras clave*:** eficiencia, error cuadrático medio, sesgo, variable auxiliar.

[a]Assistant professor. E-mail: drskystats@gmail.com
[b]Professor. E-mail: kadilar@hacettepe.edu.tr

1. Introduction

The auxiliary information in sampling theory is used for improved estimation of parameters thereby enhancing the efficiencies of the estimators.

Let $(x_i,\, y_i),\ i = 1,\, 2,\, \ldots,\, n$ be the n pair of sample observations for the auxiliary and study variables, respectively, drawn from the population of size N using Simple Random Sampling without Replacement. Let $\overline{X}$ and $\overline{Y}$ be the population means of auxiliary and study variables, respectively, and let $\overline{x}$ and $\overline{y}$ be the respective sample means. Ratio estimators are used when the line of regression of y on x passes through the origin and the variables x and y are positively correlated to each other, while product estimators are used when x and y are negatively correlated to each other; otherwise, regression estimators are used.

The sample variance estimator of the population variance is defined as

$$t_0 = s_y^2 \tag{1}$$

which is an unbiased estimator of population variance $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$ and its variance is

$$V\left(t_0\right) = \gamma\, S_y^4\left(\lambda_{40} - 1\right) \tag{2}$$

where $\lambda_{rs} = \frac{\mu_{rs}}{\mu_{20}^{r/2}\,\mu_{02}^{s/2}}$, $\mu_{rs} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^r (X_i - \overline{X})^s$, and $\gamma = \frac{1}{n}$.

Isaki (1983) proposed the ratio type estimator for estimating the population variance of the study variable as

$$t_R = s_y^2\left(\frac{S_x^2}{s_x^2}\right) \tag{3}$$

where

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2, s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i, \overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i, \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The Bias (B) and Mean Squared Error (MSE) of the estimator in (3), up to the first order of approximation, are given, respectively, as

$$B(t_R) = \gamma S_y^2\left[\left(\lambda_{40} - 1\right) - \left(\lambda_{22} - 1\right)\right] \tag{4}$$

$$MSE(t_R) = \gamma S_y^4\left[\left(\lambda_{40} - 1\right) + \left(\lambda_{04} - 1\right) - 2\left(\lambda_{22} - 1\right)\right] \tag{5}$$

Singh, Chauhan, Sawan & Smarandache (2011) proposed the exponential ratio estimator for the population variance as

$$t_{Re} = s_y^2 \exp\left[\frac{S_x^2 - s_x^2}{S_x^2 + s_x^2}\right] \tag{6}$$

The bias and MSE, up to the first order of approximation, respectively, are

$$B(t_{Re}) = \gamma\, S_y^2 \left[\frac{3}{8}(\lambda_{04}-1) - \frac{1}{2}(\lambda_{22}-1)\right] \tag{7}$$

$$MSE(t_{Re}) = \gamma\, S_y^4 \left[(\lambda_{40}-1) + \frac{(\lambda_{04}-1)}{4} - (\lambda_{22}-1)\right] \tag{8}$$

The usual linear regression estimator for population variance is

$$\widehat{S}_{lr}^2 = s_y^2 + b(S_x^2 - s_x^2) \tag{9}$$

where $b = \frac{s_y^2(\widehat{\lambda}_{22}-1)}{s_x^2(\widehat{\lambda}_{04}-1)}$ is the sample regression coefficient.

The MSE of $\widehat{S}_{lr}^2$, to the first order of approximation, is

$$MSE(\widehat{S}_{lr}^2) = \gamma\, S_y^4 \left[(\lambda_{40}-1) - \frac{(\lambda_{22}-1)^2}{\lambda_{04}-1}\right] \tag{10}$$

Many more authors, including Singh & Singh (2001, 2003), Nayak & Sahoo (2012), among others, have contributed to variance estimation.

2. Improved Exponential Type Ratio Estimator

Motivated by Upadhyaya, Singh, Chatterjee & Yadav (2011) and following them, we propose the improved ratio exponential type estimator of the population variance as follows:

The ratio exponential type estimator due to Singh et al. (2011) is given by

$$t_{Re} = s_y^2 \exp\left[\frac{S_x^2 - s_x^2}{S_x^2 + s_x^2}\right] = s_y^2 \exp\left[1 - \frac{2s_x^2}{S_x^2 + s_x^2}\right]$$

which can be generalized by introducing a positive real constant 'α' (i.e. $\alpha \geq 0$) as

$$t_{Re}^{(\alpha)} = s_y^2 \exp\left[1 - \frac{\alpha s_x^2}{S_x^2 + (\alpha-1)s_x^2}\right] = s_y^2 \exp\left[\frac{S_x^2 - s_x^2}{S_x^2 + (\alpha-1)s_x^2}\right] \tag{11}$$

Here, we note that: (i) For $\alpha = 0$, $t_{Re}^{(\alpha)}$ in (11) reduces to

$$t_{Re}^{(0)} = s_y^2 \exp[1] \tag{12}$$

which is a biased estimator with the MSE larger than s_y^2 utilizing no auxiliary information as the value of 'e' is always greater than unity.

(ii) For $\alpha = 1$, $t_{Re}^{(\alpha)}$ in (11) reduces to

$$t_{Re}^{(1)} = s_y^2 \exp\left[\frac{S_x^2 - s_x^2}{S_x^2}\right] \tag{13}$$

(iii) For $\alpha = 2$, $t_{Re}^{(\alpha)}$ in (11) reduces to Singh et al. (2011) ratio exponential type estimator

$$t_{Re} = s_y^2 \exp\left[\frac{S_x^2 - s_x^2}{S_x^2 + s_x^2}\right] \tag{14}$$

Then, we have investigated for which value of α, the proposed estimator, $t_{Re}^{(\alpha)}$, is more efficient than the estimators, t_0, t_R, and t_{Re}.

3. The First Degree Approximation to the Bias and Mean Squared Error of the Suggested Estimator

In order to study the large sample properties of the proposed class of estimator, $t_{Re}^{(\alpha)}$, we define $s_y^2 = S_y^2(1+\varepsilon_0)$ and $s_x^2 = S_x^2(1+\varepsilon_1)$ such that $E(\varepsilon_i) = 0$ for $(i = 0, 1)$ and $E\left(\varepsilon_0^2\right) = \gamma(\lambda_{40} - 1)$, $E\left(\varepsilon_1^2\right) = \gamma(\lambda_{04} - 1)$, $E(\varepsilon_0\varepsilon_1) = \gamma(\lambda_{22} - 1)$.

To the first degree of approximation, the bias and the MSE of the estimator, $t_{Re}^{(\alpha)}$, are respectively given by

$$B(t_{Re}^{(\alpha)}) = \gamma S_y^2 \frac{(\lambda_{04} - 1)}{2\alpha^2}\left[2\alpha(1-\lambda) - 1\right] \tag{15}$$

$$MSE(t_{Re}^{(\alpha)}) = \gamma\, S_y^4 \left[(\lambda_{40} - 1) + \frac{(\lambda_{04} - 1)}{\alpha^2}(1 - 2\alpha\lambda)\right] \tag{16}$$

where $\lambda = \frac{\lambda_{22}-1}{\lambda_{04}-1}$.

The $MSE(t_{Re}^{(\alpha)})$ is minimum for

$$\alpha = \frac{1}{\lambda} = \alpha_0 \text{ (say)} \tag{17}$$

Substituting $\alpha = \frac{1}{\lambda}$ into (11), we get the asymptotically optimum estimator (AOE) in the class of estimators $(t_{Re}^{(\alpha)})$ as

$$(t_{Re}^{(\alpha_0)}) = s_y^2 \exp\left[\frac{\lambda(S_x^2 - s_x^2)}{\lambda S_x^2 + (1-\lambda)s_x^2}\right] \tag{18}$$

The value of λ can be obtained from the previous surveys or the experience gathered in due course of time, for instance, see Murthy (1967), Reddy (1973, 1974) and Srivenkataramana & Tracy (1980), Singh & Vishwakarma (2008), Singh & Kumar (2008) and Singh & Karpe (2010).

The mean square error of AOE $(t_{Re}^{(\alpha_0)})$, to the first degree of approximation, is given by

$$MSE(t_{Re}^{(\alpha_0)}) = \gamma\, S_y^4 \left[(\lambda_{40} - 1) - \frac{(\lambda_{22} - 1)^2}{\lambda_{04} - 1}\right] \tag{19}$$

which equals to the approximate MSE of the usual linear regression estimator of population variance.

In case the practitioner fails to guess the value of 'λ' by utilizing all his resources, it is worth advisable to replace λin (18) by its consistent estimate

$$\widehat{\lambda} = \frac{\widehat{\lambda}_{22} - 1}{\widehat{\lambda}_{04} - 1} \tag{20}$$

Thus, the substitution of $\widehat{\lambda}$ in (18) yields an estimator based on estimated 'λ' as

$$(t_{Re}^{(\widehat{\alpha}_0)}) = s_y^2 \exp\left[\frac{\widehat{\lambda}(S_x^2 - s_x^2)}{\widehat{\lambda}S_x^2 + (1-\widehat{\lambda})s_x^2}\right] \tag{21}$$

It can be shown to the first degree of approximation that

$$MSE(t_{Re}^{(\alpha_0)}) = MSE(t_{Re}^{(\widehat{\alpha}_0)}) = \gamma\, S_y^4 \left[(\lambda_{40} - 1) - \frac{(\lambda_{22} - 1)^2}{\lambda_{04} - 1}\right] \tag{22}$$

Thus, the estimator $t_{Re}^{(\widehat{\alpha}_0)}$, given in (21), is to be used in practice as an alternative to the usual linear regression estimator.

4. Efficiency Comparisons of the Proposed Estimator with the Mentioned Existing Estimators

From (16) and (2), we have $MSE(t_0) - MSE(t_{Re}^{(\alpha)}) = \gamma\, S_y^4 \frac{(\lambda_{04}-1)}{\alpha^2}\,(1 - 2\alpha\lambda) > 0$, if

$$\alpha > \frac{1}{2\lambda} \tag{23}$$

From (16) and (5), we have $MSE(t_R) - MSE(t_{Re}^{(\alpha)}) = \gamma\, S_y^4(\lambda_{04}-1)(1-\frac{1}{\alpha})\,(1 + \frac{1}{\alpha} - 2\lambda) > 0$, if either

$$\min\left\{1,\ \frac{1}{2\lambda - 1}\right\} < \alpha < \max\left\{1,\ \frac{1}{2\lambda - 1}\right\}, \quad \lambda > \frac{1}{2} \tag{24}$$

or

$$\alpha > 1, \quad 0 \leq \lambda \leq \frac{1}{2}. \tag{25}$$

From (16) and (8), we have $MSE(t_{Re}) - MSE(t_{Re}^{(\alpha)}) = \gamma\, S_y^4(\lambda_{04} - 1)(\frac{1}{2} - \frac{1}{\alpha})\,(\frac{1}{2} + \frac{1}{\alpha} - 2\lambda) > 0$, if either

$$\min\left\{2,\ \frac{2}{4\lambda - 1}\right\} < \alpha < \max\left\{2,\ \frac{2}{4\lambda - 1}\right\}, \quad \lambda > \frac{1}{4} \tag{26}$$

or

$$\alpha > 2, \quad 0 \leq \lambda \leq \frac{1}{4} \tag{27}$$

From (16) and (10), we have

$$MSE(t_{Re}^{(\alpha)}) - MSE(\widehat{S}_{lr}^2) = \gamma\, S_y^4(\lambda_{04}-1)\left[\frac{(\lambda_{04}-1)}{\alpha} - (\lambda_{22}-1)\right]^2 < 0$$

if

$$\lambda_{04} < 1 \tag{28}$$

5. Numerical Illustrations

The appropriateness of the proposed estimator has been verified with the help of the four data sets, given in Table 1 (Subramani & Kumarapandiyan 2012). In Table 2, which gives the range of α and also the optimal value, α_0, for the efficiency condition of the proposed estimator, we see that $(t_{Re}^{(\alpha)})$, is quite wide as t_{Re}; whereas, from Table 3, which provides the Percent Relative Efficiencies (PREs) of different estimators of the population variance with respect to the sample variance, we observe that the proposed estimator is more efficient than t_{Re}.

Table 1: Parameters of populations.

Parameters	**Population 1**	**Population 2**	**Population 3**	**Population 4**
N	103	103	80	49
n	40	40	20	20
$\overline{Y}$	626.2123	62.6212	51.8264	116.1633
$\overline{X}$	557.1909	556.5541	11.2646	98.6765
ρ	0.9936	0.7298	0.9413	0.6904
S_y	913.5498	91.3549	18.3569	98.8286
C_y	1.4588	1.4588	0.3542	0.8508
S_x	818.1117	610.1643	8.4563	102.9709
C_x	1.4683	1.0963	0.7507	1.0435
λ_{04}	37.3216	17.8738	2.8664	5.9878
λ_{40}	37.1279	37.1279	2.2667	4.9245
λ_{22}	37.2055	17.2220	2.2209	4.6977
λ	0.9969	0.9635	0.7748	0.7846

Table 2: Range of 'α' for $(t_{Re}^{(\alpha)})$ to be more efficient than different estimators of the population variance.

	Populations			
Estimators	**1**	**2**	**3**	**4**
t_0	$\alpha > 0.50$	$\alpha > 0.52$	$\alpha > 0.65$	$\alpha > 0.64$
t_R	$\alpha \in (1.00, 1.01)$	$\alpha \in (1.00, 1.08)$	$\alpha \in (1.00, 1.82)$	$\alpha \in (1.00, 1.76)$
t_{Re}	$\alpha \in (0.67, 2.00)$	$\alpha \in (0.70, 2.00)$	$\alpha \in (0.95, 2.00)$	$\alpha \in (0.94, 2.00)$
Common Range of α for $\left(t_{Re}^{(\alpha)}\right)$ to be more efficient than t_0, t_R, t_{Re}	$\alpha \in (0.67, 2.00)$	$\alpha \in (0.70, 2.00)$	$\alpha \in (0.95, 2.00)$	$\alpha \in (0.94, 2.00)$
Optimum value of α	$\alpha_0 = 1.003$	$\alpha_0 = 1.038$	$\alpha_0 = 1.291$	$\alpha_0 = 1.275$

TABLE 3: Percent relative efficiencies (PREs) of different estimators of population variance with respect to sample variance $t_0 = s_y^2$.

	Populations			
Estimators	**1**	**2**	**3**	**4**
$t_0 = s_y^2$	100.00	100.00	100.00	100.00
t_R	93,838.70	175.74	183.23	258.72
t_{Re}	401.30	149.76	247.21	266.29
$t_{Re}^{(\widehat{\alpha}_0)}$	**94,749.28**	**175.96**	**270.63**	**331.68**

6. Conclusion

We have suggested an improved exponential ratio estimator for estimating the population variance. From theoretical discussions, given in Section 4 and results in Table 3, we infer that the proposed estimator is better than the mentioned existing estimators in literature, the usual sample variance , traditional ratio estimator due to Isaki (1983) and Singh et al. (2011) exponential ratio estimator in the sense of having lesser mean squared error. We have also given the range of α along with its optimum value for to be more efficient than different estimators. Hence, the proposed estimator is recommended for its practical use for estimating the population variance when the auxiliary information is available. In future articles, we hope to adapt the proposed estimator here to the combined and separate methods in the stratifed random sampling.

Acknowledgements

The authors are grateful to the referees and the Editor-in-Chief for providing some useful comments on an earlier draft of the paper.

[Recibido: agosto de 2012 — Aceptado: mayo de 2013]

References

Isaki, C. (1983), 'Variance estimation using auxiliary information', *Journal of the American Statistical Association* **78**, 117–123.

Murthy, M. (1967), *Sampling Theory and Methods*, Calcutta Statistical Publishing Society, Kolkatta, India.

Nayak, R. & Sahoo, L. (2012), 'Some alternative predictive estimators of population variance', *Revista Colombiana de Estadistica* **35**(3), 507–519.

Reddy, V. (1973), 'On ratio and product methods of estimation', *Sankhya Serie B* **35**(3), 307–316.

Reddy, V. (1974), 'On a transformed ratio method of estimation', *Sankhya Serie C* **36**, 59–70.

Singh, H. & Karpe, N. (2010), 'Estimation of mean, ratio and product using auxiliary information in the presence of measurement errors in sample surveys', *Journal of Statistical Theory and Practice* **4**(1), 111–136.

Singh, H. & Kumar, S. (2008), 'A general family of estimators of finite population ratio, product and mean using two phase sampling scheme in the presence of non-response', *Journal of Statistical Theory and Practice* **2**(4), 677–692.

Singh, H. & Singh, R. (2001), 'Improved ratio-type estimator for variance using auxiliary information', *Journal of Indian Society of Agricultural Statistics* **54**(3), 276–287.

Singh, H. & Singh, R. (2003), 'Estimation of variance through regression approach in two phase sampling', *Aligarh Journal of Statistics* **23**, 13–30.

Singh, H. & Vishwakarma, G. (2008), 'Some families of estimators of variance of stratified random sample mean using auxiliary information', *Journal of Statistical Theory and Practice* **2**(1), 21–43.

Singh, R., Chauhan, P., Sawan, N. & Smarandache, F. (2011), 'Improved exponential estimator for population variance using two auxiliary variables', *Italian Journal of Pure and Applied Mathematics* **28**, 101–108.

Srivenkataramana, T. & Tracy, D. (1980), 'An alternative to ratio method in sample surveys', *Annals of the Institute of Statistical Mathematics* **32**, 111–120.

Subramani, J. & Kumarapandiyan, G. (2012), 'Variance estimation using median of the auxiliary variable', *International Journal of Probability and Statistics* **1**(3), 36–40.

Upadhyaya, L., Singh, H., Chatterjee, S. & Yadav, R. (2011), 'Improved ratio and product exponential type estimators', *Journal of Statistical Theory and Practice* **5**(2), 285–302.

Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 153 a 176

Response Surface Optimization in Growth Curves Through Multivariate Analysis

Optimización de superficies de respuesta en curvas de crecimiento a través de análisis multivariado

Felipe Ortiz[1,a], Juan C. Rivera[2,b], Oscar O. Melo[2,c]

[1]Facultad de Estadística, Universidad Santo Tomás, Bogotá, Colombia

[2]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

A methodology is proposed to jointly model treatments with quantitative levels measured throughout time by combining the response surface and growth curve techniques. The model parameters, which measure the effect throughout time of the factors related to the second-order response surface model, are estimated. These estimates are made through a suitable transformation that allows to express the model as a classic MANOVA model, so the traditional hypotheses are formulated and tested. In addition, the optimality conditions throughout time are established as a set of specific combination factors by the fitted model. As a final step, two applications are analyzed using our proposed model: the first was previously analyzed with growth curves in another paper, and the second involves two factors that are optimized over time.

***Key words*:** Growth curves, Multiple optimization, Response surfaces, Second order models.

Resumen

En este artículo se propone una metodología para modelar conjuntamente tratamientos con niveles cuantitativos medidos en el tiempo, mediante la combinación de técnicas de superficies de respuesta con curvas de crecimiento. Se estiman los parámetros del modelo, los cuales miden el efecto en el tiempo de los factores relacionados con el modelo de superficie de respuesta de segundo orden. Estas estimaciones se realizan a través de una transformación que permite expresar el modelo como un modelo clásico de MANOVA; de esta manera, se expresan y juzgan las hipótesis tradicionales.

[a]Lecturer. E-mail: andresortiz@usantotomas.edu.co
[b]MsC in Statistics. E-mail: jcriverar@unal.edu.co
[c]Associate professor. E-mail: oomelom@unal.edu.co

Además, las condiciones de optimización a través del tiempo son establecidas para un conjunto de factores específicos por medio del modelo ajustado. Como paso final, se analizan dos aplicaciones utilizando el modelo propuesto: la primera fue analizada mediante curvas de crecimiento en otro artículo, y la segunda consiste en dos factores que son optimizados a lo largo del tiempo.

***Palabras clave*:** curvas de crecimiento, optimización múltiple, superficies de respuesta, modelos de segundo orden.

1. Introduction

Sometimes in experimentation, researchers interest focuses on analyzing data over time to know the tendencies of an individual or groups of individuals. In other cases, the goal is not only the trend but also to know what combination of factors can optimize the process over time. This latter context is the starting point for analysis of growth curves and response surface methodology (RSM). Response surface and growth curves are statistical methods frequently used in the analysis of experiments. The purpose of the first is to determine the optimum operating conditions of a process, whereas the latter method is used to model the effect of treatments throughout time.

Two applications of the above hybrid model approach are analyzed in this paper. The first is an experiment to analyze the effect of dietary ingestion of sodium Zeolite A (SZA) on the growth and physiology of sixty horses reported by Frey, Potter, Odom, Senor, Reagan, Weir, Elsslander, Webb, Morris, Smith & Weigand (1992). The horses were randomly assigned to four treatments: control and three levels of dietary SZA (0.66%, 1.32% and 2%). In addition, plasma silicon concentration was measured at the times: $t = 0, 1, 3, 6, 9$ hours after ingestion on eighty four days into the diet. The second study is an experiment about the waste-water treatment, in which is common adding inhibitory agents to reduce the negative environmental impact generated by these substances discharged into the receiving water bodies. In such cases, we study the biological oxygen demand (BOD) as a water pollution measure. Montoya & Gallego (2012) performed a central composite rotatable design adding combinations of detergent (D) and animal fat (AF) to the residual water. The BOD, biomass growth and substrate consumption at $t = 24, 48, 72, 96, 120$ hours after of mixture were observed.

In both experiments, we are interested in studying the optimum combination of factors throughout time that optimizes our response variable. Therefore, in these kinds studies, we want to observe if the growth curves can be represented by a cubic, quadratic or linear polynomial in time, and if the response surface can be expressed by a quadratic or linear polynomial in the treatments. Furthermore, we want to obtain the confidence band(s) for the expected combination of factors over time (response surface throughout growth curves).

A growth curve is a model of the evolution of a quantity over time. Growth curves are widely used in biology for quantities such as population size, body height or biomass. Growth curve experiments have been considered from various angles by Rao (1959), Potthoff & Roy (1964), Khatri (1966), Khatri (1973), Verbyla

& Venables (1988), Kshirsagar & Boyce (1995), Srivastava (2002), Pan & Fang (2002), Chiou, Müller, Wang & Carey (2003) and Kahm, Hasenbrink, Lichtenberg-Fraté, Ludwig & Kschischo (2010). All of these growth curve studies involve successive and correlated measurements on the same individuals which are divided into two or more groups of treatments. We use in this paper treatments that are combinations of quantitative factors which are based on polynomial models in the response surface.

RSM uses statistical models and therefore practitioners need to be aware that even the best statistical model is an approximation to reality. In this way, if researches are interested in modeling and analyzing situations to determine optimum operating conditions for a process; this particular analysis is performed through the RSM. It is widely applicable in the biological sciences, chemistry, social experimentation agriculture, engineering, food sciences, quality control and economics, among others. The RSM has been developed in experimental and industrial production by Box & Wilson (1951), Hill & Hunter (1966), Mead & Pike (1975), Lucas (1976), Box & Draper (1982), Draper & Ying (1994), Chiou, Müller & Wang (2004) and Box & Draper (2007). These authors discussed some first-order and second-order response surface designs from the point of view of their ability to detect certain likely kinds of lack of fit for a higher's degree polynomial than has been fitted.

The two previous approaches to growth curve and RSM problems are now mixed to give a solution to our two applications because we need to know what is the combination of factors over time that best works in the optimization process. Our methodology is derived from the theory of multivariate normal analysis of variance, and it is based on polynomial models for both growth curve and response surface. Moreover, we provide both confidence bands and the over-all tests of significance for various kinds of compound hypotheses that involve the parameters of the proposed model. Furthermore, we find the optimal operating conditions over time.

This kind of problem was previously studied by Guerrero & Melo (2008) providing a solution where they combined the response surface and the growth curve techniques using an univariate analysis. In this paper, the same is done to obtain the functional relationship that exists between the treatment and time in order to predict its effect in any future time. Although, there are several phenomena of this kind where these two techniques may be used simultaneously, a procedure that combines them at the same time is not known using multivariate analysis. This analysis works better than the univariate approximation presented by Guerrero & Melo (2008) because the different statistics for hypothesis testing are exact, which does not always happen in the univariate approach.

The experiments to be considered are characterized by the presence of k fixed quantitative factors, $\zeta_1, \zeta_2, \ldots, \zeta_k$, associated with a continuous variable of interest $\boldsymbol{Y}$, where the observed levels of each factor are equally spaced and the response variable is measured on the same experimental units in several moments.

The plan of the paper is the following: Section 2 presents the response surfaces model in growth curves. Then in Section 3, parameter estimation, hypotheses testing and test statistics are presented. Section 4 is dedicated to locating the optimum; at first the model is reparametrized (Section 4.1), and then the optimal point is found (Section 4.2). Finally, two applications of our procedure are showed in Section 5, and the conclusions are exposed in Section 6.

2. Response Surfaces Model in Growth Curves

The growth curve model implies that there are g different groups or treatments and a single growth variable y, which is measured at q time points $t_1, t_2, \ldots, t_q$ on n_j individuals chosen at random from the j-th group ($j = 1, 2, \ldots, g$). A polynomial regression of degree $(p-1)$ for y on the time variable t is defined. Thus,

$$E(y_t) = \phi_{j0}t^0 + \phi_{j1}t^1 + \cdots + \phi_{j(p-1)}t^{p-1} \tag{1}$$

$t = t_1, t_2, \ldots, t_q, q > p-1, j = 1, 2, \ldots, g$. The observations $y_{t_1}, \ldots, y_{t_q}$ on the same individual are correlated, and come from a multivariate normal distribution with unknown variance-covariance matrix $\boldsymbol{\Sigma}$, equal for all the individuals. Let $\boldsymbol{Y}_j$ denote the $n_j \times q$ matrix of the observations for the j-th group, and let

$$\mathbf{Y}' = [\boldsymbol{Y}'_1, \boldsymbol{Y}'_2, \ldots, \boldsymbol{Y}'_g]$$

denote the $q \times n$ matrix for all the $n = n_1 + n_2 + \cdots + n_g$ individuals. Then from (1)

$$E(\boldsymbol{Y}_j) = \begin{pmatrix} \boldsymbol{\phi}_j\mathbf{G} \\ \boldsymbol{\phi}_j\mathbf{G} \\ \vdots \\ \boldsymbol{\phi}_j\mathbf{G} \end{pmatrix} = \mathbf{J}_{n_j 1}\boldsymbol{\phi}_j\mathbf{G}, \quad j = 1, 2, \ldots, g \tag{2}$$

where $\boldsymbol{\phi}_j = [\phi_{j0}, \phi_{j1}, \ldots, \phi_{j(p-1)}]'$ denotes the vector of the regression or growth curve coefficients for the j-th group, and

$$\mathbf{G} = \begin{pmatrix} t_1^0 & t_2^0 & \ldots & t_q^0 \\ t_1^1 & t_2^1 & \ldots & t_q^1 \\ \vdots & \vdots & \ddots & \vdots \\ t_1^{p-1} & t_2^{p-1} & \ldots & t_q^{p-1} \end{pmatrix}$$

and $\mathbf{J}_{a\times b}$ denotes, in general, an $a \times b$ matrix with all unit elements. Furthermore, the matrix $\mathbf{G}_{p\times q}$ relates the parameters of the curve with the corresponding polynomial degree.

Combining (2) for all g groups, we now have

$$E(\mathbf{Y}) = \begin{pmatrix} \mathbf{J}_{n_1 1}\boldsymbol{\phi}_1\mathbf{G} \\ \mathbf{J}_{n_2 1}\boldsymbol{\phi}_2\mathbf{G} \\ \vdots \\ \mathbf{J}_{n_g 1}\boldsymbol{\phi}_g\mathbf{G} \end{pmatrix} = \mathbf{A}\boldsymbol{\Phi}\mathbf{G} \tag{3}$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \vdots \\ \boldsymbol{\phi}_g \end{pmatrix}$$

is the $g \times p$ matrix of the growth curve coefficients, and $\boldsymbol{A} = diag[\mathbf{J}_{n_1 1}, \mathbf{J}_{n_2 1}, \ldots \mathbf{J}_{n_g 1}]$ is a block diagonal matrix of order $n \times g$ 'group indicator'. Therefore, assuming independence between individuals, we have that

$$Var(Vec(\mathbf{Y})) = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_q \tag{4}$$

where $\otimes$ denotes the Kronecker product of two matrices (see Magnus (1988)).

The equations (3) and (4) conform to the growth curve model introduced by Potthoff & Roy (1964), and later analyzed by Khatri (1966), Grizzle & Allen (1969), Kabe (1974), and Khatri (1988), among many others.

2.1. Construction of Proposed Model

With the idea of making a joint modeling of growth curves and response surfaces, a couple of aspects were considered:

1. The matrix $\mathbf{A}_{n\times g}$, whose columns contain information about treatments, was changed by a new matrix $\mathbf{X}_{n\times s}$, whose columns register the levels of a factor and their interactions for each of the n individuals, just like in second order response surfaces designs with k quantitative fixed factors and $s = 1 + k + (k + \binom{k}{2})$ parameters in the surface.

2. For relating the parameters from the response surface with each of the groups, a new matrix of parameters $\boldsymbol{\theta}_{s\times g}$ was included in the model where θ_{lj} measures the effect of the l-th parameter in the surface for the $j-th$ group. Let $\boldsymbol{\Phi}_{g\times p}$ be the matrix that relates the groups with the growth curve coefficients, i.e., ϕ_{jm} is the parameter associated to the degree coefficient m in the growth curve for the j-th group ($l = 1, 2, \ldots, s; j = 1, 2, \ldots, g; m = 0, 1, \ldots, p-1$).

Under the usual assumptions described above and maintaining the same structure and interpretation for the matrices $\mathbf{Y}_{n\times q}$ and $\mathbf{G}_{p\times q}$, the proposed model is given by

$$E(\mathbf{Y}_{n\times q}) = \mathbf{X}_{n\times s}\,\boldsymbol{\theta}_{s\times g}\,\boldsymbol{\Phi}_{g\times p}\,\mathbf{G}_{p\times q} \tag{5}$$

Notice that the model (5) is a classic model Potthoff & Roy (1964) adaptation in which the matrix $\boldsymbol{\xi}_{s\times p} = \boldsymbol{\theta}_{s\times g}\boldsymbol{\Phi}_{g\times p}$, whose components are given by

$$\xi_l^m = \sum_{j=1}^{g} \theta_{lj}\phi_{jm}$$

which is the parameter associated with the l-th component of the surface in the m-th growth curve degree ($l = 1, 2, \ldots, s$ and $m = 0, 1, \ldots, p-1$). This allows to write (5) as

$$E(\mathbf{Y}_{n\times q}) = \mathbf{X}_{n\times s}\,\boldsymbol{\xi}_{s\times p}\,\mathbf{G}_{p\times q} \tag{6}$$

Another form for writing this model is

$$E(y_{ia}) = \sum_{m=0}^{p-1}\left(\xi_{\,0}^{\,m} + \sum_{r=1}^{k}\xi_{\,r}^{\,m}x_{ir} + \sum_{r=1}^{k}\sum_{r'=1}^{k}\xi_{\,rr'}^{\,m}\,x_{ir}x_{ir'}\right)t_a^m \tag{7}$$

or equivalently the model (6) can be written as

$$E(y_{ia}) = \left(\sum_{m=0}^{p-1}\xi_0^m t_a^m\right) + \sum_{r=1}^{k}x_{ir}\left(\sum_{m=0}^{p-1}\xi_r^m t_a^m\right) + \sum_{r=1}^{k}\sum_{r'=1}^{k}x_{ir}x_{ir'}\left(\sum_{m=0}^{p-1}\xi_{rr'}^m t_a^m\right) \tag{8}$$

with $a = 1, 2, \ldots, q$ and $i = 1, \ldots, n$, and where $\xi_{\,rr'}^{\,m}$ is the parameter that denotes the effect of the interaction between the factors r and r' in the m-th growth curve degree ($r, r' = 1, 2, \ldots, k$ and $m = 0, 1, \ldots, p-1$), x_{ir} and $x_{ir'}$ are encoded explanatory variables associated to the factors r-th and r'-th, respectively, and y_{ia} is the response variable associated to the i-th individual in the a-th time.

Note that the model (7) is in fact a growth curve whose coefficients are themselves a response surface of order two, and the model (8) is a response surface whose parameters are growth curves. Moreover in (7), it is necessary to point out that for a fixed m, all the parameters of the form ξ_0^m, ξ_r^m, $\xi_{rr'}^m$ ($r, r' = 1, 2, \ldots, k$) belong to the m-th column of $\boldsymbol{\xi}$. Similarly, in (8), each set of parameters of the form ξ_0^m, ξ_r^m, $\xi_{rr'}^m$ with $m = 0, 1, \ldots, p-1$ and fixed r, r', conforms the rows of $\boldsymbol{\xi}$. The remarks above are of great utility in section 3.2 for building the hypotheses of interest on the model parameters.

3. Inference on the Model

3.1. Parameter Estimation

Parameter estimation is achieved by expressing the model (6) as a MANOVA classic model, using the following transformation

$$\mathbf{Y}^{\triangle} = \mathbf{Y}\mathbf{P}^{-1}\mathbf{G}'(\mathbf{G}\mathbf{P}^{-1}\mathbf{G}')^{-1} \tag{9}$$

with $\mathbf{P}$ any symmetric positive definite matrix, such that $(\mathbf{G}\mathbf{P}^{-1}\mathbf{G}')^{-1}$ exists.

By applying the transformation (9) in (6), the next expression is obtained

$$\begin{aligned} E(\mathbf{Y}^{\triangle}_{n\times p}) &= \mathbf{X}_{n\times s}\,\boldsymbol{\xi}_{s\times p} \\ Var(\mathbf{Y}^{\triangle}_{i}) &= (\mathbf{G}\mathbf{P}^{-1}\mathbf{G}')^{-1}\mathbf{G}\mathbf{P}^{-1}\boldsymbol{\Sigma}\mathbf{P}^{-1}\mathbf{G}'(\mathbf{G}\mathbf{P}^{-1}\mathbf{G}')^{-1} \\ &= \boldsymbol{\Sigma}^{\triangle}_{p}, \qquad\qquad i = 1, 2, \ldots, n \end{aligned} \tag{10}$$

Potthoff & Roy (1964) found that taking $\mathbf{P} = \boldsymbol{\Sigma}$ produces the minimum variance estimator for $\boldsymbol{\xi}$; however, since $\boldsymbol{\Sigma}$ is unknown, in practice $\mathbf{P}$ is given by

$$\mathbf{P} = \mathbf{S} = \mathbf{Y}'\{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\}\mathbf{Y} \tag{11}$$

Note that $\mathbf{P}$ can take different forms which depend of the data correlation structure; a complete discussion about $\mathbf{P}$ can be found in Davis (2002), Molenberghs & Verbeke (2005), and Davidian (2005).

Then, for model (10), the parameter estimators obtained with the maximum likelihood method are given by

$$\widehat{\boldsymbol{\xi}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}^{\triangle} \tag{12}$$

From a slight extension of the result given by Rao (1967) in equation 50, we can find that the unconditional covariance matrix of the elements of $\widehat{\boldsymbol{\xi}}$ can be expressed as

$$Var\left(\widehat{\boldsymbol{\xi}}\,'\right) = \tfrac{n-s-1}{n-s-q+p-1}(\mathbf{X}'\mathbf{X})^{-1} \otimes \boldsymbol{\Sigma}^{\triangle} \tag{13}$$

where $\otimes$ is the Kronocker product, and $Var(\widehat{\boldsymbol{\xi}}\,')$ denotes the covariance matrix of the elements of $\widehat{\boldsymbol{\xi}}$ taken in a columnwise manner.

It is easily shown that $E(\widehat{\boldsymbol{\xi}}) = \boldsymbol{\xi}$, and using a result given by Grizzle & Allen (1969), we find that $E((\mathbf{G}\mathbf{S}^{-\mathbf{1}}\mathbf{G}')^{-1}) = (n-s-q+p)\boldsymbol{\Sigma}^{\triangle}$. From this last equation and equation (13), it follows that an unbiased estimator of the variance of $\widehat{\boldsymbol{\xi}}$ is

$$\widehat{Var}\left(\widehat{\boldsymbol{\xi}}\,'\right) = \tfrac{n-s-1}{n-s-q+p-1}(\mathbf{X}'\mathbf{X})^{-1} \otimes \widehat{\boldsymbol{\Sigma}}^{\triangle} \tag{14}$$

where

$$\widehat{\boldsymbol{\Sigma}}^{\triangle} = \frac{1}{n-s-q+p}(\mathbf{G}\mathbf{S}^{-\mathbf{1}}\mathbf{G}')^{-\mathbf{1}}$$

In next Subsection, we will present a classic technique for testing a hypothesis of the form $\mathbf{C}\,\boldsymbol{\xi}\,\mathbf{U} = \mathbf{0}$ under the generalized expectation model (6), and also we will obtain related confidence bounds.

3.2. Hypothesis of Interest and Test Statistics

As shown in the section 2.1, the model (6) can be written by expressions (7) and (8), where it can be observed that the hypotheses of interest lie mainly on the

rows or the columns of the matrix $\boldsymbol{\xi}$. These and many other can be written in the conventional general linear hypothesis form

$$H_0: \quad \mathbf{C}\,\boldsymbol{\xi}\,\mathbf{U} = \mathbf{0} \qquad vs \qquad H_1: \quad \mathbf{C}\,\boldsymbol{\xi}\,\mathbf{U} \neq \mathbf{0} \tag{15}$$

where $\mathbf{C}_{c\times s}$ and $\mathbf{U}_{p\times u}$ are known matrices of ranges c ($\leq s$) and u ($\leq p$), respectively. The matrices that define the main hypotheses, together with their corresponding interpretation, are shown in Table 1.

Table 1: Hypotheses more common over treatments and times.

H_0	Interpretation	$\mathbf{C}$	$\mathbf{U}$
$\boldsymbol{\xi} = \mathbf{0}$	The time-parameter interaction adjusted by the intercepts is not significant.	$\begin{pmatrix} 0 & \mathbf{0}_{1\times s-1} \\ \mathbf{0}_{s-1\times 1} & \mathbf{I}_{s-1} \end{pmatrix}$	$\begin{pmatrix} 0 & \mathbf{0}_{1\times p-1} \\ \mathbf{0}_{p-1\times 1} & \mathbf{I}_{p-1} \end{pmatrix}$
$\boldsymbol{\xi}^{(m)} = \mathbf{0}$	The m-th column of ξ is zero, indicating that the degree m coefficient is not important in the growth curve.	$\mathbf{I}_s$	$(0,\ldots,\underset{\substack{\downarrow \\ m-th}}{1},\ldots,0)'_{p\times 1}$
$\boldsymbol{\xi}_{(l)} = \mathbf{0}$	The l-th row of ξ is zero, indicating that the parameter of the surface is not significant.	$(0,\ldots,\underset{\substack{\downarrow \\ l-th}}{1},\ldots,0)_{1\times s}$	$\mathbf{I}_p$
$\xi_l^m = 0$	The l-th component of the surface does not exercise influence in the m-th degree of the curve.	$(0,\ldots,\underset{\substack{\downarrow \\ l-th}}{1},\ldots,0)_{1\times s}$	$(0,\ldots,\underset{\substack{\downarrow \\ m-th}}{1},\ldots,0)'_{p\times 1}$

For the construction of the test statistics, the following two matrices should be kept in mind

$$\mathbf{H} = \mathbf{U}'\widehat{\boldsymbol{\xi}}'\mathbf{C}'[\mathbf{C}\mathbf{R}_\mathbf{1}\mathbf{C}']^{-\mathbf{1}}\mathbf{C}\widehat{\boldsymbol{\xi}}\mathbf{U}$$
$$\mathbf{E} = \mathbf{U}'(\mathbf{G}\mathbf{S}^{-\mathbf{1}}\mathbf{G}')^{-\mathbf{1}}\mathbf{U}$$

where

$$\mathbf{R_1} = \left\{\mathbf{I} + (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}\mathbf{S}^{-\mathbf{1}}\left[\mathbf{I} - \mathbf{G}'(\mathbf{G}\mathbf{S}^{-\mathbf{1}}\mathbf{G}')^{-\mathbf{1}}\mathbf{G}\mathbf{S}^{-\mathbf{1}}\right]\mathbf{Y}'\mathbf{X}\right\}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$$

$\mathbf{H}$ and $\mathbf{E}$ play a decisive role in building the four classic multivariate test statistics used in testing hypothesis (15) under the model (10): the Roy's test uses the largest characteristic root of $(\mathbf{H}\mathbf{E}^{-\mathbf{1}})$, the Lawley-Hotelling $T^2 = tr(\mathbf{H}\mathbf{E}^{-\mathbf{1}})$, the trace of Bartlett-Nanda-Pillai $V = tr(\mathbf{H}(\mathbf{H}+\mathbf{E})^{-\mathbf{1}})$, and the statistic proposed

by Wilks (1932), $|\mathbf{E}|/|\mathbf{H}+\mathbf{E}| \sim \Lambda_{(u,c,m)}$ with $m = n-[s+(q-p)]$, which is known as the λ-criterion.

The hypotheses in Table 1 involves one row vector or one column vector. Therefore, we now state the general rule of rejecting a null hypothesis based on Wilks's Λ, using a level of significance α. To test the null hypothesis $\boldsymbol{\xi}^{(m)} = \mathbf{0}$, we use the following test presented in Kshirsagar & Boyce (1995)

$$u_1 = \frac{1-\Lambda}{\Lambda}\frac{ddf}{ndf}$$

where *ddf* is the denominator degree of freedom and *ndf* is the numerator degrees of freedom. Then, the null hypothesis $\boldsymbol{\xi}^{(m)} = \mathbf{0}$ is rejected if $u_1 > F_{(\alpha,ndf,ddf)}$.

To test the null hypothesis $\boldsymbol{\xi}_{(l)} = \mathbf{0}$, we use the following test presented in Kshirsagar & Boyce (1995)

$$c_1 = \frac{1-\Lambda}{\Lambda}\left(\frac{c+ddf-u}{u}\right)$$

Then, the null hypothesis $\boldsymbol{\xi}_{(l)} = \mathbf{0}$ is rejected if $c_1 > F_{(\alpha,u,c+ddf-u)}$.

On the other hand, simultaneous 100(1-α)% confidence bounds for the function $\mathbf{b}'\mathbf{C}\boldsymbol{\xi}\mathbf{U}\mathbf{f}$, $\forall\ \mathbf{b}_{(c\times 1)}$ and $\mathbf{f}_{(u\times 1)}$, are given by

$$\mathbf{b}'\mathbf{C}\widehat{\boldsymbol{\xi}}\mathbf{U}\mathbf{f} \pm \left\{\left(\tfrac{h_\alpha}{1-h_\alpha}\right)(\mathbf{b}'\mathbf{C}\mathbf{R}_1\mathbf{C}'\mathbf{b})(\mathbf{f}'\mathbf{E}\mathbf{f})\right\}^{1/2} \tag{16}$$

where the prediction is, of course, the first term of the equation (16) and h_α stands for the α fractile of the distribution for the Roy's largest characteristic root statistic tabulated by Heck (1960) with its three parameters (denoted by s, m and n in Heck's notation, but here denoted, respectively, by s^*, m^* and n^*) equal to $s^* = \min(c,u)$, $m^* = (|c-u|-1)/2$ and $n^* = (n-s-(q-p)-u-1)/2$.

Other test statistics are presented in some works; for instance, Grizzle and Allen's statistic (1969) which considers a variant for the matrix associated the hypothesis (relating to the herein presented). Singer & Andrade (1994) remarked on the appropriate selection of error terms and presented a test statistic that follows an exact F distribution (under H_0). This was also used in the application of Section 5, since it yielded the same decisions as the test statistics exposed there.

4. Location of the Optimum

The crucial goal of the response surface methodology is to find the optimal operating conditions for the variable of interest, and in this scenario, their behavior throughout time is added.

4.1. Reparameterization of the Model

In order to find the optimal operating conditions in presence of multiple responses, it is convenient to find an expression that us allows to distinguish the

terms of order zero, one and two of the model (10). This model can be reparametrized as

$$\begin{aligned}\widehat{\mathbf{Y}}^{\triangle}_{i_{1\times p}} &= \mathbf{b}_{0_{1\times p}} + (\mathbf{x}'_{1\times k}\mathbf{b}_{k\times p}) + (\mathbf{x}'\mathbf{B}^{(0)}\mathbf{x}, \mathbf{x}'\mathbf{B}^{(1)}\mathbf{x}, \ldots, \mathbf{x}'\mathbf{B}^{(p-1)}\mathbf{x}) \\ &= \mathbf{b}_{0_{1\times p}} + (\mathbf{x}'_{1\times k}\mathbf{b}_{k\times p}) + (\mathbf{x}'_{1\times k}\mathbf{B}_{k\times kp})(\mathbf{I}_p \otimes \mathbf{x}_{k\times 1})\end{aligned} \tag{17}$$

where $\mathbf{x}_{k\times 1}$ is the vector associated to the k factors of the response surfaces, $\mathbf{b}_{0_{1\times p}}$ is the vector whose components are the intercepts of each curve degree, $\mathbf{b}_{k\times p}$ is the matrix that contains the coefficients associated to the k linear terms of the response surface for each of the curve degrees, and $\mathbf{B}_{k\times kp}$ is the matrix $(\mathbf{B}^{(0)}, \mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(p-1)})$ with $\mathbf{B}^{(m)}$ $(m = 0, 1, \ldots, p-1)$ being the $k \times k$ matrix associated to the quadratic form of the response surface for the m-th growth curve degree.

4.2. Optimization

The location of the optimal point is obtained by solving the equation system resulting from the expression

$$\frac{\partial \widehat{\mathbf{Y}}^{\triangle}_{i}}{\partial \mathbf{x}} = \mathbf{b}_{k\times p} + 2[\mathbf{B}^{(0)}\mathbf{x} \vdots \mathbf{B}^{(1)}\mathbf{x} \vdots \cdots \vdots \mathbf{B}^{(p-1)}\mathbf{x}] = \mathbf{0} \tag{18}$$

which is demonstrated using properties of differential matrix calculus.

By applying the *vec* operator in system (18), the following system of k variables and kp equations is obtained

$$vec(\mathbf{b}_{k\times p}) + 2\begin{pmatrix} \mathbf{B}^{(0)}\mathbf{x} \\ \mathbf{B}^{(1)}\mathbf{x} \\ \vdots \\ \mathbf{B}^{(p-1)}\mathbf{x} \end{pmatrix} = \mathbf{0}$$

$$\mathbf{B}'\mathbf{x} = -\frac{1}{2}vec(\mathbf{b}_{k\times p}),$$

which is solved by appending to it a pre-matrix $\mathbf{B}$; hence, the stationary point is

$$\mathbf{x}_0 = -\frac{1}{2}(\mathbf{B}\mathbf{B}')^{-1}\mathbf{B}vec(\mathbf{b}_{k\times p}) \tag{19}$$

and the non-singularity of $\mathbf{B}\mathbf{B}'$ is guaranteed by the linear independence of the columns of $\mathbf{X}'\mathbf{X}$.

Let $\gamma_1, \gamma_2, \ldots, \gamma_k$ be the characteristic roots of the matrix $\mathbf{B}\mathbf{B}'$, then the nature of the stationary point is determined by

- If $\gamma_v > 0$ $\forall v = 1, 2, \ldots, k$, then $\mathbf{x}_0$ is minimum.
- If $\gamma_v < 0$ $\forall v = 1, 2, \ldots, k$, then $\mathbf{x}_0$ is maximum.

- In any other case, $\mathbf{x}_0$ is a saddle point.

Using (16) with $\mathbf{C} = \mathbf{I}$, $\mathbf{U} = \mathbf{I}$, $\mathbf{b}' = \mathbf{x}_0$ and $\mathbf{f} = \mathbf{G}^{(m)}$; the confidence bounds for the predicted values of the optimal point in each moment are given by

$$\mathbf{x}_0'\widehat{\boldsymbol{\xi}}\mathbf{G}^{(m)} \pm \left\{ \left(\tfrac{1}{2n^*+2} F_{(\alpha,1,2n^*+2)} \right) (\mathbf{x}_0'\mathbf{R}_1\mathbf{x}_0) \left[\mathbf{G}^{(m)\prime}\mathbf{E}\mathbf{G}^{(m)} \right] \right\}^{1/2} \tag{20}$$

where $\mathbf{G}^{(m)}$ is a column of the matrix $\mathbf{G}$ and $n^* = (n - s - 2)/2$.

5. Applications

Two applications are analyzed in this Section: the first is an experiment to analyze the plasma silicon concentration and its effect over the dietary ingestion of SZA on the growth of sixty horses (Frey et al. 1992), and the second is an experiment about the waste-water treatment, where the biological oxygen demand (BOD) as a water pollution is studied (Montoya & Gallego 2012).

5.1. Plasma Silicon Concentration

An experiment to analyze the effect of dietary ingestion of SZA on the growth and physiology of sixty horses was reported by Frey et al. (1992). The horses were randomly assigned to four treatments: control (0%) and three levels of dietary SZA (0.66%, 1.32% and 2%). In addition, the plasma silicon concentration was measured in the times: $t = 0, 1, 3, 6$ and 9 hours after ingestion at eighty four days into the diet. This data was previously analyzed by Kshirsagar & Boyce (1995) employing growth curves, but they did not consider the surface responses part. However, Guerrero & Melo (2008) presented an optimization process that combines response surface and growth curves from a univariate approach. The last analysis differs from the work in this paper because we make a parameter estimation which does not depend on the transformation of equation (9). Additionally, the test statistics used in Guerrero & Melo (2008) follow a F distribution approximately, while under the multivariate perspective employed throughout this paper, these tests follow an exact distribution of Wilks's Λ.

Figure 1 shows profiles plot for these data. In this Figure, we see that the silicon concentration in the plasma can be modeled as a cubic polynomial over time. Also, the control group (0%) seems to have a different behavior than other concentrations which suggest a difference among the four treatments.

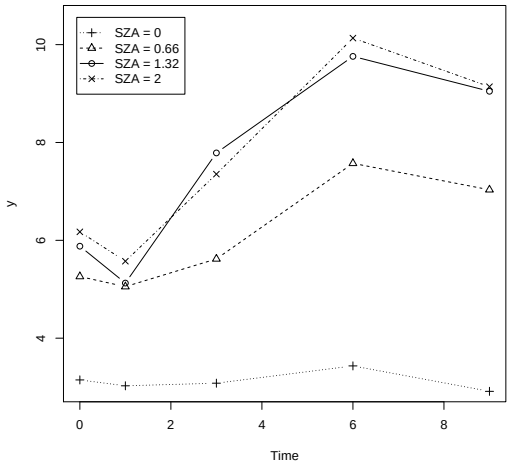


FIGURE 1: Profiles by time for plasma silicon concentration growth.

Fitting the model (6) to this data set, the parameter estimates given by (12) and $\mathbf{P}$ as (11) are

$$\widehat{\boldsymbol{\xi}} = \begin{bmatrix} 3.267 & -0.324 & 0.118 & -0.010 \\ 3.169 & 0.151 & 0.192 & -0.017 \\ -0.921 & -0.127 & -0.024 & 0.002 \end{bmatrix}$$

where the rows are growth curves for the different parameters of the response surface, and the columns correspond to response surfaces for the different growth curve degrees. So, the first row contains the intercepts of the surfaces for the polynomial degrees, the second row contains the linear component of the factor (SZA), and the third row contains the quadratic component of the factor.

Now, the results of the hypotheses testing on the rows (surface parameters) and the columns (curve coefficients) of $\boldsymbol{\xi}$ are shown in Table 2. The hypothesis $H_0 : \boldsymbol{\xi}^{(4)} = \mathbf{0}$ yields a $p - value < 0.001$; therefore, the hypothesis is rejected. This means that the third-order coefficient of the fitted growth curve is significant in the model (see right panel of Figure 2). The hypothesis $H_0 : \boldsymbol{\xi}_{(3)} = \mathbf{0}$ also yields a $p - value < 0.001$, denoting that the quadratic component of the factor is important, too (see left panel of Figure 2). The hypothesis $H_0 : \boldsymbol{\xi}^{(2)} = \mathbf{0}$ is the only one that is not rejected, it corresponds to the linear component of the curve. However, since the degree of the cubic growth curve is significant, the linear component is also included due to the hierarchy of the fitted growth curve.

TABLE 2: Hypotheses testing on the rows and columns for the effect of dietary ingestion of SZA.

Hypothesis	**C**	**U**	Λ	F_c	ndf	ddf	$p-value$
$\boldsymbol{\xi}^{(1)} = \mathbf{0}$	$\mathbf{I}_3$	$(1,0,0,0)'$	0.099	169.45	3	56	< 0.001
$\boldsymbol{\xi}^{(2)} = \mathbf{0}$	$\mathbf{I}_3$	$(0,1,0,0)'$	0.928	1.44	3	56	0.2407
$\boldsymbol{\xi}^{(3)} = \mathbf{0}$	$\mathbf{I}_3$	$(0,0,1,0)'$	0.584	13.29	3	56	< 0.001
$\boldsymbol{\xi}^{(4)} = \mathbf{0}$	$\mathbf{I}_3$	$(0,0,0,1)'$	0.471	20.98	3	56	< 0.001
$\boldsymbol{\xi}_{(1)} = \mathbf{0}$	$(1,0,0)$	$\mathbf{I}_4$	0.350	24.57	4	53	< 0.001
$\boldsymbol{\xi}_{(2)} = \mathbf{0}$	$(0,1,0)$	$\mathbf{I}_4$	0.290	32.38	4	53	< 0.001
$\boldsymbol{\xi}_{(3)} = \mathbf{0}$	$(0,0,1)$	$\mathbf{I}_4$	0.510	12.74	4	53	< 0.001

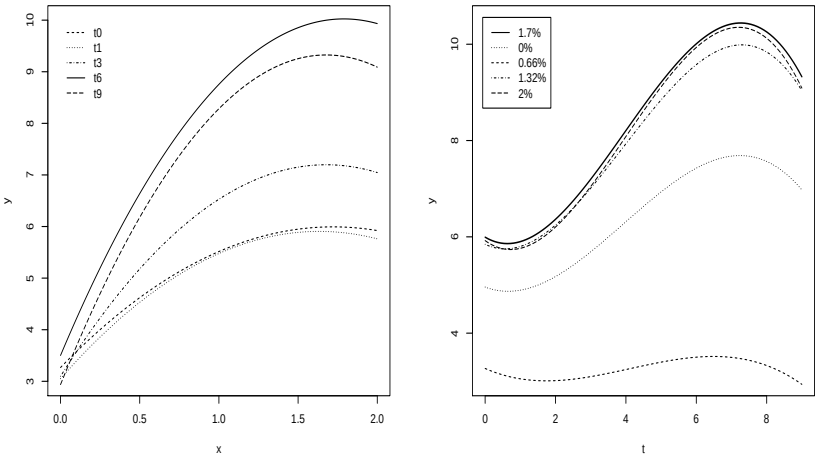


FIGURE 2: Fitted response surfaces (left panel) and growth curves (right panel).

From the matrix of estimated parameters, it is possible to construct the estimated growth curves for the four treatments of the experimental design. For example, for treatment 0.66%, the growth curve is given by the equation

$$\begin{pmatrix} 1 & 0.66 & 0.66^2 \end{pmatrix} \begin{pmatrix} 3.267 & -0.324 & 0.118 & -0.010 \\ 3.169 & 0.151 & 0.192 & -0.017 \\ -0.921 & -0.127 & -0.024 & 0.002 \end{pmatrix} \begin{pmatrix} 1 \\ t \\ t^2 \\ t^3 \end{pmatrix}$$
$$= 4.957 - 0.279t + 0.233t^2 - 0.019t^3 \qquad (21)$$

For the four treatments, the fitted growth curves are summarized in Table 3 and on Figure 2 (right panel). In the same way, we can find the estimation of the

response surface parameters at each point in time. From product matrix, $\widehat{\boldsymbol{\xi}}\mathbf{G}$, the equations are derived and summarized in Table 4, and they are plotted on Figure 2 (left panel).

Table 3: Fitted growth curves for the effect of dietary ingestion of SZA.

Treatment (SZA)	Growth curve
0.00	$3.266 - 0.323t + 0.117t^2 - 0.009t^3$
0.66	$4.957 - 0.279t + 0.233t^2 - 0.019t^3$
1.32	$5.845 - 0.345t + 0.328t^2 - 0.027t^3$
2.00	$5.921 - 0.529t + 0.403t^2 - 0.034t^3$

Table 4: Fitted response surfaces for the effect of dietary ingestion of SZA.

Time	Response surfaces
t_0	$3.266 + 3.169(\text{SZA}) - 0.92(\text{SZA})^2$
t_1	$3.051 + 3.495(\text{SZA}) - 1.07(\text{SZA})^2$
t_3	$3.097 + 4.88(\text{SZA}) - 1.456(\text{SZA})^2$
t_6	$3.051 + 7.292(\text{SZA}) - 2.038(\text{SZA})^2$
t_9	$2.936 + 7.616(\text{SZA}) - 2.27(\text{SZA})^2$

On the other hand, the level of SZA that maximizes the plasma silicon concentration regularly well throughout time obtained with (19) is 1.70%, where $\mathbf{b} = (3.169, 0.151, 0.192, -0.017)$ and $\mathbf{B} = (-0.921, -0.127, -0.024, 0.002)$. The confidence bounds in the optimal point constructed using (20) and $\mathbf{x}_0 = (1, 1.7, 1.7^2)$ are shown in Table 5.

Table 5: Parameter estimation and confidence bounds in the optimal point for the effect of dietary ingestion of SZA.

	t_0	t_1	t_3	t_6	t_9
Estimated value	5.99	5.9	7.19	10.01	9.32
Lower limit	5.72	5.66	6.93	9.74	9.06
Upper limit	6.25	6.13	7.45	10.27	9.59

Under the same reasoning used in equation (21), it is possible to construct the growth curve for the optimum point (1.7%), which is given by $5.99 - 0.43t + 0.37t^2 - 0.03t^3$. Figure 2 (right panel) shows the optimum supremacy over all treatments throughout time.

According to the results obtained in this application, we can stand out three facts:

1. in the solution via univariate developed by Guerrero & Melo (2008), in which one time ($t = 1$) was removed to get that the remaining times ($t = 0, 3, 6, 9$)

were equally spaced; the quadratic component SZA factor in the response surface was not significant, and also a linear polynomial for the growth curve was fitted.

2. the hypothesis $H_0 : \boldsymbol{\xi}_{(3)} = \mathbf{0}$ is rejected, justifying the inclusion of the quadratic component in the response surface to fit the plasma silicon concentration. Note that in our proposal the test statistic follows an exact F distribution.

3. Figure 1 clearly suggests that we should fit a cubic model in the growth curve, which is corroborated by the results of the hypothesis $H_0 : \boldsymbol{\xi}^{(4)} = \mathbf{0}$.

5.2. Environmental Pollution

During waste-water treatment it is common inhibitory agents to reduce the negative environmental impact generated by to add substances discharged into the receiving water bodies. Montoya & Gallego (2012) performed a central composite rotatable design adding combinations of detergent (D in ppm) and animal fat (AF in ppm). They studied the residual water BOD and biomass growth and substrate consumption at $t = 12, 24, 36, 48, 60$ hours after the mixture. These components interfere with the biological degradation of organic material during the process of waste-water treatment. In this case, we study the biomass (in mg/l) growth as a water pollution measure. According to Montoya & Gallego (2012), the presence of detergents and animal fat in the affluent waste-water affect the size and shape of the resulting floccules, which produces as a result a decrease in biomass concentration demanding more time for the system retention that translates into a low BOD elimination.

A description of the behavior of the four factorial points (treatments) of the experimental design throughout time is shown in Figure 3 (left panel). This Figure shows a slight increase of biomass between 12 and 24 hours after that the treatments were applied. Furthermore, we see an accelerated growth between 24 and 48 hours and a slight decrease from 48 until 60 hours. This behavior can be approximated by a cubic polynomial throughout time. Moreover, it is noted that the profiles for the four treatments have a very similar behavior, which suggests that there is not a differential effect for factors D and AF.

In order to observe the behavior of biomass growth at each time point, we fitted the univariate response surfaces for each time (see Figure 4). We can see that the fitted surfaces for the first two times ($t = 12, 24$) have a convex shape unlike the three last times ($t = 36, 48, 60$), which have concave shape. The points that optimize each response surface are shown in Table 6; there is a change in the optimal location point between two convex curves and three concave curves.

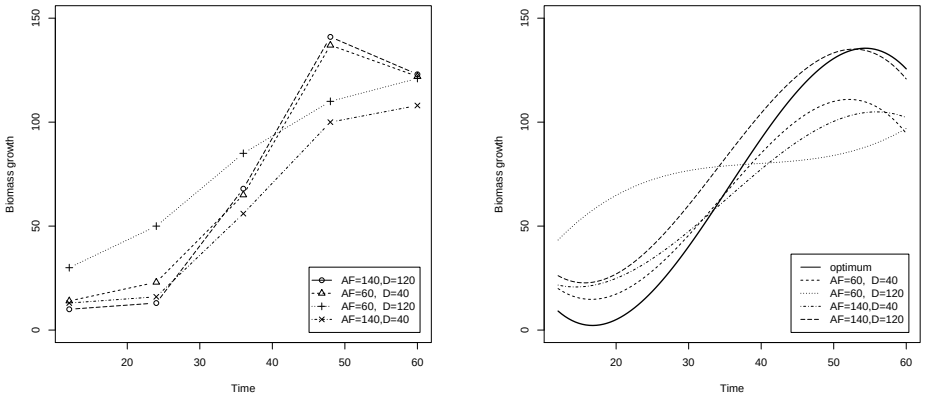


FIGURE 3: Profiles throughout time for the biomass growth (left panel), and fitted growth curves for the four treatments and the optimal point throughout time (right panel).

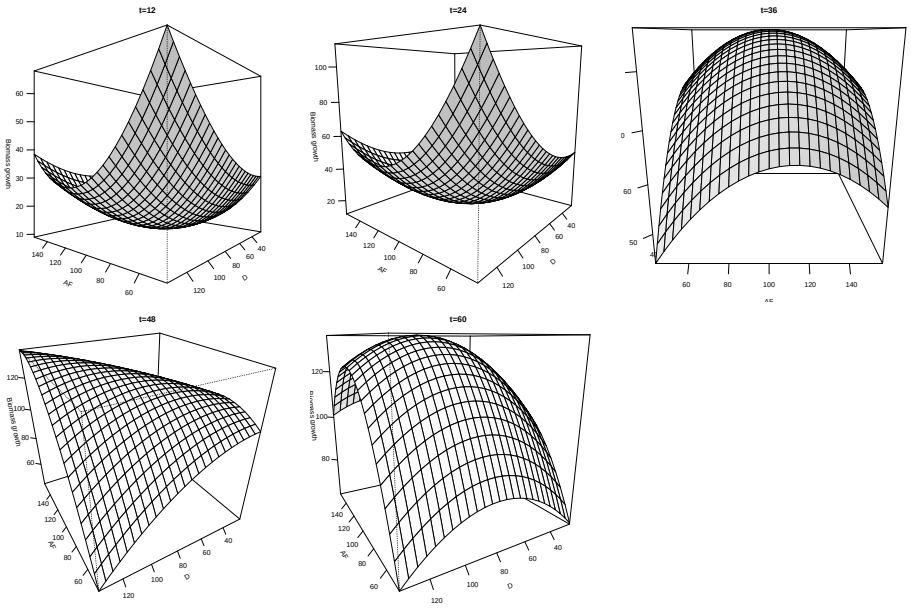


FIGURE 4: Fitted univariate response surfaces.

Table 6: Univariate optimal surfaces.

Time	AF	D	Characterization
t_{12}	100.7	50.8	Minimum
t_{24}	102.7	54.7	Minimum
t_{36}	100.3	104.6	Maximum
t_{48}	183.2	188.7	Maximum
t_{60}	109.2	92.7	Maximum

To fit the model (6), we first analyzed the structure of the matrix $\mathbf{P}$ given in equation (9). Then, we evaluated several possible covariance structures considering the fit to the data and comparing them in terms of Akaike information criterion (AIC). So, it was found that the best covariance structure was an AR(1) with parameter estimates: $\widehat{\phi} = 0.676$ and $\widehat{\sigma} = 23.51$, and the smallest AIC was 492.1. Thus, the estimated parameters matrix using the equation (12) is

$$\widehat{\boldsymbol{\xi}} = \begin{pmatrix} 66.3998 & 0.5581 & 0.0241 & -0.0010 \\ 1.4689 & -0.3724 & 0.0131 & -0.0001 \\ -1.1473 & 0.1479 & -0.0050 & 0.0000 \\ -0.0146 & 0.0029 & -0.0001 & 0.0000 \\ -0.0111 & 0.0018 & 0.0000 & 0.0000 \\ 0.0245 & -0.0036 & 0.0001 & 0.0000 \end{pmatrix}$$

whose first row represents the estimates of the response surface intercepts (ξ_0^m) in the four degrees of growth curve following the expression (7). The second and third rows are associated with the linear effects of factors AF and D, while the third and the fourth rows are associated with the quadratic effects of the factors, and finally; the sixth row estimates the interaction of two factors in all degrees of the growth curve.

Once the above is done, we show in Table 7 the results of the hypotheses testing on the rows (surface parameters) and the columns (curve parameters) of $\boldsymbol{\xi}$. According to the hypothesis $\boldsymbol{\xi}^{(4)} = \mathbf{0}$, a cubic polynomial fit to the growth curve is suitable ($p - value = 0.0039$), while for the hypothesis $\boldsymbol{\xi}_{(2)} = \mathbf{0}, \ldots, \boldsymbol{\xi}_{(6)} = \mathbf{0}$, we do not find evidence of a significant difference between the effects generated by AF and D factors in the experimental design. This is consistent with the behavior seen in Figure 3 (left panel) for the four treatments of central composite rotatable design; however, these factors could be interacting with the time (see Figure 3, left panel) so these components will be kept in the model.

Table 7: Hypotheses testing on the rows and columns for the biomass.

Hypothesis	**C**	**U**	Wilks	F_c	ndf	ddf	$p-value$
$\boldsymbol{\xi}^{(1)} = \mathbf{0}$	$\mathbf{I}_6$	$(1,0,0,0)'$	0.057	11.074	6	4	0.018
$\boldsymbol{\xi}^{(2)} = \mathbf{0}$	$\mathbf{I}_6$	$(0,1,0,0)'$	0.046	13.849	6	4	0.012
$\boldsymbol{\xi}^{(3)} = \mathbf{0}$	$\mathbf{I}_6$	$(0,0,1,0)'$	0.028	22.996	6	4	0.005
$\boldsymbol{\xi}^{(4)} = \mathbf{0}$	$\mathbf{I}_6$	$(0,0,0,1)'$	0.026	24.865	6	4	0.004
$\boldsymbol{\xi}_{(1)} = \mathbf{0}$	$(1,0,0,0,0,0)$	$\mathbf{I}_4$	0.768	0.075	4	1	0.978
$\boldsymbol{\xi}_{(2)} = \mathbf{0}$	$(0,1,0,0,0,0)$	$\mathbf{I}_4$	0.438	0.320	4	1	0.848
$\boldsymbol{\xi}_{(3)} = \mathbf{0}$	$(0,0,1,0,0,0)$	$\mathbf{I}_4$	0.831	0.051	4	1	0.988
$\boldsymbol{\xi}_{(4)} = \mathbf{0}$	$(0,0,0,1,0,0)$	$\mathbf{I}_4$	0.271	0.671	4	1	0.710
$\boldsymbol{\xi}_{(5)} = \mathbf{0}$	$(0,0,0,0,1,0)$	$\mathbf{I}_4$	0.492	0.258	4	1	0.879
$\boldsymbol{\xi}_{(6)} = \mathbf{0}$	$(0,0,0,0,0,1)$	$\mathbf{I}_4$	0.327	0.514	4	1	0.764

From the matrix for the estimated parameters the estimated growth curves for the four treatments are constructed. For example, for the treatment AF=140 and D=120, the growth curve is given by the equation

$$\begin{aligned}&\begin{pmatrix} 1 & 140 & 120 & 140^2 & 120^2 & 140(120) \end{pmatrix} \widehat{\boldsymbol{\xi}} \begin{pmatrix} 1 \\ t \\ t^2 \\ t^3 \end{pmatrix} \\ &= 99.30 - 10.86t + 0.45t^2 - 0.0043t^3 \end{aligned} \tag{22}$$

For the four treatments, the estimated growth curves are summarized in Table 8 and Figures 5 and 3 (right panel). Figure 5 compares the estimated curve fitting with the observed profiles where we see that the fitted growth curves provide a good fitting for the data.

Table 8: Fitted growth curves for the biomass.

AF	D	Growth curve
60	40	$96.99 - 11.10t + 0.44t^2 - 0.004t^3$
60	120	$-19.37 + 7.10t - 0.17t^2 + 0.001t^3$
140	40	$58.61 - 5.74t + 0.25t^2 - 0.002t^3$
140	120	$99.30 - 10.86t + 0.45t^2 - 0.004t^3$

In the same way, we can find estimates for the response surfaces at each point in time; these equations are derived from product matrix $\widehat{\boldsymbol{\xi}}\mathbf{G}$ and are summarized in Table 9 and in Figure 6. This Figure shows contour plots constructed for the biomass growth at each point in time.

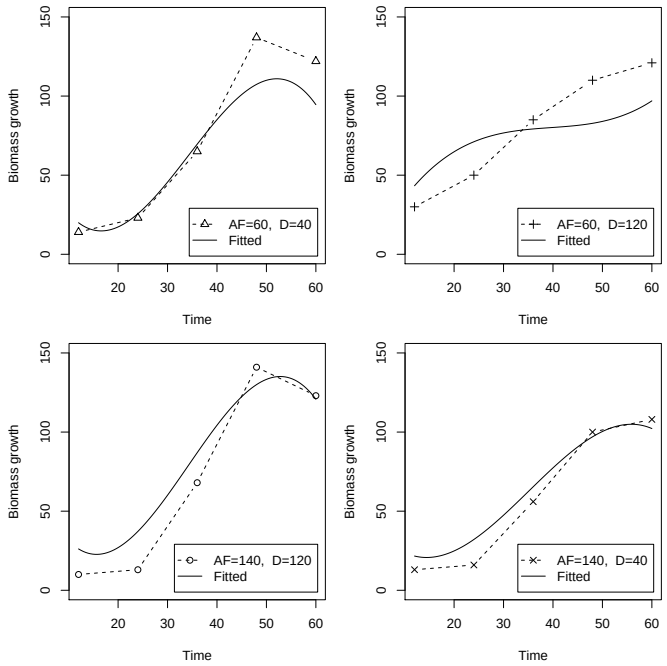


FIGURE 5: Observed and fitted profiles growth curves for the biomass.

It is stressed that the fitted surfaces in the times $t = 12, 24, 48, 60$ capture the behavior concave or convex observed in univariate surface plots (see Figure 4). Moreover, we can conclude from the contour plots that a point located approximately at the coordinate (100, 60) optimizes the process regularly well throughout time, minimizing the fitted surfaces at $t = 12, 24$ and maximizing them at $t = 48, 60$.

TABLE 9: Fitted response surfaces for the biomass

Time	Response surface
t_{12}	$74.75 - 1.30AF - 0.01D + 0.0072AF^2 + 0.0029D^2 - 0.0029AF * D$
t_{24}	$79.20 - 1.44AF + 0.22D + 0.0080AF^2 + 0.0045D^2 - 0.0063AF * D$
t_{36}	$68.88 - 0.10AF + 0.08D - 0.0001AF^2 - 0.0004D^2 + 0.0018AF * D$
t_{48}	$32.99 + 1.59AF + 0.11D - 0.0104AF^2 - 0.0059D^2 + 0.0089AF * D$
t_{60}	$-39.43 + 2.48AF + 0.84D - 0.0124AF^2 - 0.0060D^2 + 0.0025AF * D$

When the model is reparameterized using the expression (17), we obtain the following matrices

$$\mathbf{b} = \begin{pmatrix} 1.4689 & -0.3724 & 0.0131 & -1.104e^{-4} \\ -1.1473 & 0.1479 & -0.0050 & 5.190e^{-5} \end{pmatrix}$$

$$\mathbf{B}^{(0)} = \begin{pmatrix} -0.0146 & 0.0123 \\ 0.0123 & -0.0111 \end{pmatrix} \quad \mathbf{B}^{(1)} = \begin{pmatrix} 0.0029 & -0.0018 \\ -0.0018 & 0.0018 \end{pmatrix}$$

$$\mathbf{B}^{(2)} = \begin{pmatrix} -1.030e^{-4} & 6.371e^{-5} \\ 6.371e^{-5} & -6.445e^{-5} \end{pmatrix} \quad \mathbf{B}^{(3)} = \begin{pmatrix} 9.144e^{-7} & -6.066e^{-7} \\ -6.066^{-7} & 5.798e^{-7} \end{pmatrix}$$

where $\mathbf{b}$ is constructed using the linear effect estimations for the two factors (second and third rows of the estimated parameters matrix, $\widehat{\boldsymbol{\xi}}$). $\mathbf{B}^{(0)}$, $\mathbf{B}^{(1)}$, $\mathbf{B}^{(2)}$ and $\mathbf{B}^{(3)}$ are conformed by the elements of the estimated parameters matrix and kept the reparameterization structure used in the univariate response surface model i.e. the diagonal terms are equivalent to the quadratic effects for each factor, and the off-diagonal elements are equivalent to half of the estimated interaction effects.

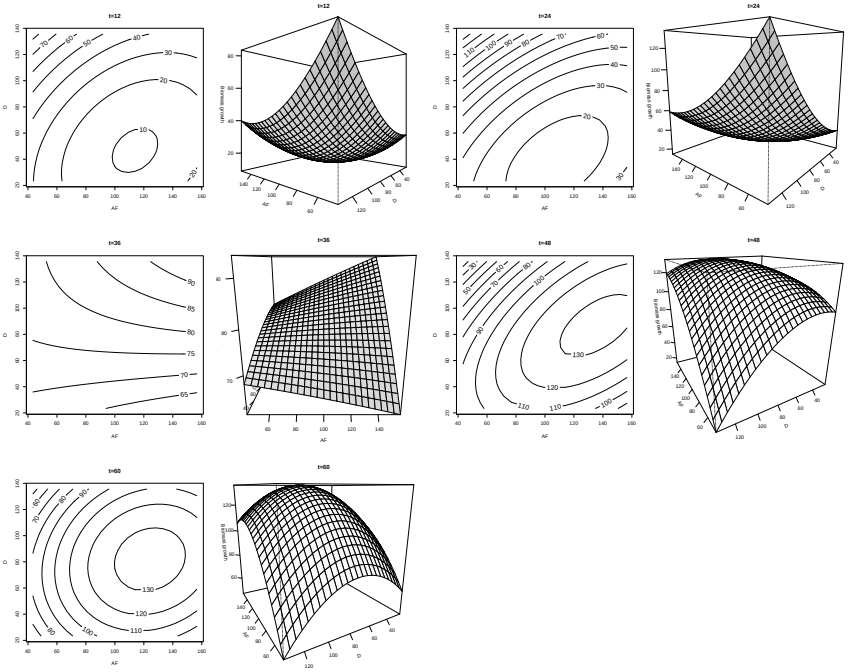


FIGURE 6: Fitted surface and contour plots for each time in the biomass study.

Thus, following the expression (19), the coordinates for the optimal point that optimizes the process throughout time are found. These are AF= 96.6 and D= 55.3 which are within the observation region of the central composite rotatable design and are in accordance with the behavior seen in the previous contour plots. The confidence bounds for optimum constructed using (20) and $\mathbf{x}_0 = (1, 96.6, 55.3, 96.6^2, 55.3^2, 96.6(55.3))$ are shown in Table 10.

Under the same reasoning used in equation (22), it is possible to construct the growth curve for the optimum found (AF= 96.6 and D= 55.3), which is given by

equation $105.2 - 13.7t + 0.53t^2 - 0.005t^3$. This growth curve allows evaluation of the optimization achieved throughout time, minimizing the growth of biomass for times $t = 12, 24$ and maximizing relatively well for time $t = 36, 48, 60$ (see Figure 3, right panel).

Table 10: Parameter estimation and confidence bounds in the optimal point for the biomass.

	t_{12}	t_{24}	t_{36}	t_{48}	t_{60}
Estimated value	9.14	15.16	71.26	125.42	125.62
Lower limit	0.00	0.00	48.84	101.63	101.34
Upper limit	33.42	38.95	93.68	149.21	149.90

6. Conclusions

A joint modelling procedure that gives additional information regarding the interaction of the studied methodologies, as opposed to analyzing them independently, was proposed. In this way, the functional relationship of the response surface parameters with time was modeled by condensing the information of the groups of the usual growth curves analysis. Also, parameter estimation, hypothesis testing, test statistics and confidence bounds were obtained. Finally, under the proposed model, the optimal point that optimizes the response variable regularly well throughout time was found.

In both applications, we studied the optimum combination of factors that optimized our response variable throughout time. Therefore, we fitted a cubic growth curve and a quadratic response surface for the treatments in both situations. In plasma silicon concentration study, it was optimized at a level of dietary ingestion of SZA 1.7% throughout time, so we can say that the plasma silicon concentration has a good growth in horses using this level. In biomass growth, we found that the optimum condition was in the combination of animal fat at a level 96.6 ppm and detergent at a level 55.3 ppm; consequently, using this combination between animal fat and detergent, we optimize this inhibitory behavior during aerobic treatment of waste-water.

Acknowledgments

The authors gratefully acknowledge the comments and suggestions of the anonymous referees that helped to improve the paper immensely. This work was partially supported by Applied Statistics in Experimental Research, Industry and Biotechnology (Universidad Nacional de Colombia).

[Recibido: septiembre de 2011 — Aceptado: mayo de 2013]

References

Box, G. E. P. & Draper, N. R. (1982), 'Measures of lack of fit for response surface designs and predictor variable transformations', *Technometrics* **24**, 1–8.

Box, G. E. P. & Draper, N. R. (2007), *Response Surfaces, Mixtures, and Ridge Analyses*, Wiley Series in Probability and Statistics, New York.

Box, G. E. P. & Wilson, K. B. (1951), 'On the experimental attainment of the optimum conditions', *Journal of the Royal Statistical Society* **13**, 1–45.

Chiou, J., Müller, H. & Wang, J. (2004), 'Functional response models', *Statistica Sinica* **14**, 675–693.

Chiou, J., Müller, H., Wang, J. & Carey, J. (2003), 'A functional multiplicative effects model for longitudinal data, with application to reproductive histories of female medflies', *Statistica Sinica* **13**, 1119–1133.

Davidian, M. (2005), *Applied Longitudinal Data Analysis*, Chapman and hall, North Carolina state university.

Davis, C. S. (2002), *Statistical Methods for the Analysis of Repeated Measurements*, Springer-Verlag, New York.

Draper, N. & Ying, L. H. (1994), 'A note on slope rotatability over all directions', *Journal of Statistical Planning and Inference* **41**, 113–119.

Frey, K. S., Potter, G. D., Odom, T. W., Senor, M. A., Reagan, V. D., Weir, V. H., Elsslander, R. V. T., Webb, M. S., Morris, E. L., Smith, W. B. & Weigand, K. E. (1992), 'Plasma silicon and radiographic bone density on weanling quarter horses fed sodium zelolite A', *Journal of Equine Veterinary Science* **12**, 292–296.

Grizzle, J. E. & Allen, D. M. (1969), 'Analysis of growth and dose response curves', *Biometrics* **25**, 357–381.

Guerrero, S. C. & Melo, O. O. (2008), 'Optimization process of growth curves through univariate analysis', *Revista Colombiana de Estadística* **31**(2), 193–209.

Heck, D. L. (1960), 'Charts of some upper percentage points of the distribution of the largest characteristic root', *The Annals of Mathematical Statistics* **31**(3), 625–642.

Hill, W. J. & Hunter, W. G. (1966), 'A review of response surface methodology: A literature review', *Technometrics* **8**, 571–590.

Kabe, D. G. (1974), 'Generalized Sverdrup's lemma and the treatment of less than full rank regression model', *Canadian Mathematical Bulletin* **17**, 417–419.

Kahm, M., Hasenbrink, G., Lichtenberg-Fraté, H., Ludwig, J. & Kschischo, M. (2010), 'grofit: Fitting biological growth curves with R', *Journal of Statistical Software* **7**, 1–21.

Khatri, C. A. (1966), 'A note on a MANOVA model applied to problems in growth curves', *Annals of the Institute of Statistical Mathematics* **18**, 75–86.

Khatri, C. A. (1973), 'Testing some covariance structures under growth curve model', *Journal Multivariate Analysis* **3**, 102–116.

Khatri, C. A. (1988), 'Robustness study for a linear growth model', *Journal Multivariate Analysis* **24**, 66–87.

Kshirsagar, A. M. & Boyce, S. (1995), *Growth Curves*, Marcel Dekker, New York.

Lucas, J. M. (1976), 'Which response surfaces is best?', *Technometrics* **18**, 411–417.

Magnus, J. R. (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, New York.

Mead, R. & Pike, D. J. (1975), 'A review of responses surface methodology from a biometric viewpoint', *Biometrics* **31**, 830–851.

Molenberghs, G. & Verbeke, G. (2005), *Models for Discrete Longitudinal Data*, Springer, New York.

Montoya, C. & Gallego, D. (2012), Modelo matemático que permita evaluar el cambio de la DBO_5 soluble debido a agentes inhibitorios en un proceso de lodos activados, Master's thesis, Facultad de minas. Universidad Nacional de Colombia.

Pan, J. & Fang, K. (2002), *Growth Curve Models and Statistical Diagnostics*, Springer Series in Statistics, New York.

Potthoff, R. & Roy, S. (1964), 'A generalized multivariate analysis of variance model useful especially for growth curve problems', *Biometrika* **51**, 313–326.

Rao, C. R. (1959), 'Some problems involving linear hypothesis in multivariate analysis', *Biometrika* **46**, 49–58.

Rao, C. R. (1967), Least squares theory using an estimated dispersion matrix and its applications to mesurement of signals, *in* 'Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 355–372.

Singer, J. M. & Andrade, D. F. (1994), 'On the choice of appropiate error terms in profile analysis', *Royal of Statistical Society* **43**(2), 259–266.

Srivastava, M. S. (2002), 'Nested growth curves models', *Sankhyã: The Indian Journal of Statistics, Series A, Selected Articles from San Antonio Conference in Honour of C. R. Rao* **64**(2), 379–408.

Verbyla, A. P. & Venables, W. N. (1988), 'An extension of the growth curve models', *Biometrika* **75**, 129–138.

Wilks, S. S. (1932), 'Certain generalizations in the analysis of variance', *Biometrika* **24**, 471–494.

Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 177 a 192

Partial Least Squares Regression on Symmetric Positive-Definite Matrices

Regresión de mínimos cuadrados parciales sobre matrices simétricas definidas positiva

Raúl Alberto Pérez[1,a], Graciela González-Farias[2,b]

[1]Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia

[2]Departamento de Probabilidad y Estadística, CIMAT-México Unidad Monterrey, Monterrey Nuevo León, México

Resumen

Recently there has been an increased interest in the analysis of different types of manifold-valued data, which include data from symmetric positive-definite matrices. In many studies of medical cerebral image analysis, a major concern is establishing the association among a set of covariates and the manifold-valued data, which are considered as responses for characterizing the shapes of certain subcortical structures and the differences between them.

The manifold-valued data do not form a vector space, and thus, it is not adequate to apply classical statistical techniques directly, as certain operations on vector spaces are not defined in a general Riemannian manifold. In this article, an application of the partial least squares regression methodology is performed for a setting with a large number of covariates in a euclidean space and one or more responses in a curved manifold, called a Riemannian symmetric space. To apply such a technique, the Riemannian exponential map and the Riemannian logarithmic map are used on a set of symmetric positive-definite matrices, by which the data are transformed into a vector space, where classic statistical techniques can be applied. The methodology is evaluated using a set of simulated data, and the behavior of the technique is analyzed with respect to the principal component regression.

***Palabras clave*:** Matrix theory, Multicollinearity, Regression, Riemann manifold.

[a]Assistant Professor. E-mail: raperez1@unal.edu.co
[b]Associate Professor. E-mail: gmfarias@cimat.mx

Abstract

Recientemente ha habido un aumento en el interés de analizar diferentes tipos de datos variedad-valuados, dentro de los cuáles aparecen los datos de matrices simétricas definidas positivas. En muchos estudios de análisis de imágenes médicas cerebrales, es de interés principal establecer la asociación entre un conjunto de covariables y los datos variedad-valuados que son considerados como respuesta, con el fin de caracterizar las diferencias y formas en ciertas estructuras sub-corticales.

Debido a que los datos variedad-valuados no forman un espacio vectorial, no es adecuado aplicar directamente las técnicas estadísticas clásicas, ya que ciertas operaciones sobre espacio vectoriales no están definidas en una variedad riemanniana general. En este artículo se realiza una aplicación de la metodología de regresión de mínimos cuadrados parciales, para el entorno de un número grande de covariables en un espacio euclídeo y una o varias respuestas que viven una variedad curvada llamada espacio simétrico Riemanniano. Para poder llevar a cabo la aplicación de dicha técnica se utilizan el mapa exponencial Riemanniano y el mapa log Riemanniano sobre el conjunto de matrices simétricas positivas definida, mediante los cuales se transforman los datos a un espacio vectorial en donde se pueden aplicar técnicas estadísticas clásicas. La metodología es evaluada por medio de un conjunto de datos simulados en donde se analiza el comportamiento de la técnica con respecto a la regresión por componentes principales.

***Key words*:** multicolinealidad, regresión, teoría de matrices, variedad Riemanniana.

1. Introduction

In studies of diffusion tensor magnetic resonance imaging (TD-MRI), a diffusion tensor (DT) is calculated in each voxel of an imaging space, which describes the local diffusion of water molecules in various directions over this region of the brain. A sequence of images is used to measure this diffusion. The sequence includes a noise that produces uncertainty in the tensor estimation and in the estimation of certain quantities inherent to water molecules, such as eigenvalues, eigenvectors, the anisotropic fraction rate (FA) and the fiber trajectories, which are constructed based on these last parameters. The diffusion-tensor imaging (DTI) is a powerful tool to quantitatively evaluate the integrity of the anatomic connectivity in the white matter of clinical populations. The methods used for the analysis of DTI at a group level include the statistical analysis of certain invariant measures, such as eigenvalues, eigenvectors or principal directions, the anisotropic fraction, and the average diffusivity, among others. However, these invariant measures do not capture all of the information about the complete DTs, which leads to a decrease in the statistical power of the DTI to detect subtle changes in white matter. Hence, new statistical methods are being developed to fully analyze the DTs as responses and to establish their association with a set of covariates (Li, Zhu, Chen, Ibrahim, An, Lin, Hall & Shen 2009, Zhu, Chen, Ibrahim, Li & Lin 2009, Yuan, Zhu, Lin & Marron 2012). In some of these development the log-euclidean metric has been used with the transformation of the DTs from a non-linear space into their

logarithmic matrices on a Euclidean space. Semi-parametrical models have been proposed to study the relationship between the set of covariates and the DTs as responses. Estimation processes and hypotheses test based on test statistics and resampling methods have been developed to simultaneously evaluate the statistical significance of linear hypotheses throughout large regions of interest (ROI).

An appropriate statistical analysis of DTs is important to understand the normal development of the brain, the neurological bases of neuropsychiatric disorders and the joined effects of environmental and genetic factors on the brain structure and function. In addition, any statistical method for complete diffusion tensors can be directly applied to positive-definitive tension matrices in computational anatomy to understand the variations in shapes of brain-structure imaging (Grenander & Miller 1998, Lepore, Brun, Chou, Chiang, Dutton, Hayashi, Luders, Lopez, Aizenstein, Toga, Becker & Thompson 2008).

Symmetric positive-definite (SPD) matrix-valued data occur in a wide variety of applications, such as DTI for example, where a SPD 3x3 DT, which tracks the effective diffusion of the water molecules in certain brain regions, is estimated at each voxel of an imaging space. Another application of SPD matrix-valued data can be seen in studies on functional magnetic resonance imaging (fMRI), where an SPD covariance matrix is calculated to delineate the functional connectivity between different neuronal assembles involved in the execution of certain complex cognitive tasks or in perception processes (Fingelkurts & Kahkonen 2005). Despite the popularity of SPD matrix-valued data, there are few statistical methods to analyze SPD matrices as response variables in a Riemannian manifold. Data considered as responses with a small number of covariates of interest in a Euclidian space can be found from the following studies in the literature for statistical analysis using regression models of SPD matrices: Batchelor, Moakher, Atkinson, Calamante & Connelly (2005), Pennec, Fillard & Ayache (2006), Schwartzman (2006), Fletcher & Joshi (2007), Barmpoutis, Vemuri, Shepherd & Forder (2007), Zhu et al. (2009) and Kim & Richards (2010). However, because the SPD matrix data do not form a vector space, classical multivariate regression techniques cannot be applied directly to establish the relationship between these types of data and a set of covariates of interest.

In a setting with a large number of covariates with a high multicollinearity presence and few available observations, no regression methods have been previously proposed to study the relationship between such covariates and the response variables of SPD matrices living in non-Euclidian spaces. In this article, partial least squares (PLS) regression is proposed using a strategy of exponential and Riemannian logarithmic maps to transform data into Euclidian spaces. The development of the technique is similar to the scheme for the analysis of SPD matrices data as responses in a classical regression model and in a local polynomial regression model, as proposed in Zhu et al. (2009) and Yuan et al. (2012). The PLS regression model is initially evaluated using a set of simulated data and statistical validation techniques which currently exist, such as cross validation techniques. The behavior of the PLS regression technique is analyzed by comparing it to that of the classic dimension-reduction technique, called principal component (PC) regression.

The article is structured as follows: In Section 2, a brief revision of the existing theory for PC and the PLS regression classical model is outlined. In Section 3, some properties of the Riemannian geometric structure of SPD matrices are reviewed. An outline of the regression models, as well as the estimation methods of their regression coefficients are also presented. In Section 4, our PLS regression model is presented, along with the estimation process used and their application and evaluation on a simulated-data set. In Section 5, conclusions and recommendations for future work are given.

2. Regression Methods

2.1. Classical Regression

We will examine the following data set $\{(\mathbf{x}_i, y_i) \ : \ i = 1, 2, \ldots, n\}$, composed of a response y_i and a $k \times 1$ covariate vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T$, where the response can be continuous, discrete, or qualitative observations, and the covariates can be qualitative or quantitative. A regression model often includes two key elements: A link function $\mu_i(\boldsymbol{\beta}) = E[y|\mathbf{x}_i] = g(\mathbf{x}_i, \boldsymbol{\beta})$ and a residual $\epsilon_i = y_i - \mu_i(\boldsymbol{\beta})$, where $\boldsymbol{\beta}_{q\times 1}$ is a regression-coefficients vector and $g(.\ ,\ .)$: from $\mathbb{R}^k \times \mathbb{R}^q \to \mathbb{R}$, $(\mathbf{x}_i, \boldsymbol{\beta}) \to g(\mathbf{x}_i, \boldsymbol{\beta})$ with $q = k + 1$, can be known or unknown according to the type of model: Parametric, not-parametric or semi-parametric. The parametric regression model can be defined as: $y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + \epsilon_i$, with $g(\mathbf{x}_i, \boldsymbol{\beta})$: Known and $E[\epsilon_i|\mathbf{x}_i] = 0, \forall i = 1, 2, \ldots, n$, where the expectation is taken with respect to the conditional distribution of ϵ given $\mathbf{x}$. The non-parametric model can be defined as $y_i = g(\mathbf{x}_i) + \epsilon_i$, with $g(\mathbf{x}_i)$: Unknown and $E[\epsilon_i|\mathbf{x}_i] = 0$.

For inference on $\boldsymbol{\beta}$ in the parametric case (or on $g(.)$, in the non-parametric case), at least three statistical procedurals are needed. First, an estimation method needs to be developed to calculate the estimate of the coefficients of vector $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$. Second, it needs to be proven that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ and that it has certain asymptotic properties. Third, test statistics need to be developed for testing hypotheses with the form:

$$H_0 \ : \mathbf{H}\boldsymbol{\beta} = \mathbf{b}_0 \quad \text{v.s} \quad H_a \ : \mathbf{H}\boldsymbol{\beta} \neq \mathbf{b}_0$$

where normally $\mathbf{H}_{r\times s}$, $\boldsymbol{\beta}_{s\times 1}$ and $\mathbf{b_0}_{r\times 1}$ are a constant matrix, a regression-coefficients vector and a constant vector respectively.

2.2. Regression in Sub-Spaces of Variables

In many practical situations, the number of variates is much greater than the quantity of available observations in the data set for a regression model, causing the problem of multicollinearity between the predictors. Among the available options for handling this problem are techniques based in explicit or implicit subspaces and the Bayesian approach, which includes additional information about the parameters of the model. In the case of the sub-spaces, the regression is

realized within a feasible space of a lesser dimension. The sub-space may be constructed explicitly with a geometric-type motivation derived from the use of latent variables, or implicitly using regularization techniques to avoid the problem of multicollinearity. A latent variable is a non-observable variable that is inferred from other variables by being directly observed and measured. The introduction of latent variables allows to capture more relevant information about the covariates matrix, denoted by $\mathbf{X}$, or information about the structure of the interaction between $\mathbf{X}$ and the response variables matrix, denoted by $\mathbf{Y}$.

In this approach, latent, non-correlated variables are introduced, denoted by $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_a$ and $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_a$, where a is the number of componets retained. The use of latent variables allows for the factorization of low ranges of the predictor and/or the response matrix, which allows for the adjustment of a linear regression model by least squares upon this set of latent variables.

The vectors loadings $\mathbf{p}_k$ and $\mathbf{q}_k$, with $k = 1, 2, \ldots, a$, generate a-dimensional spaces, where the coefficients $\mathbf{t}_{k\ n\times 1}$ and $\mathbf{u}_{k\ n\times 1}$ are considered as latent variables. Among the approaches based on latent variables are PCR and PLS regression, which are briefly described below.

In PC regression, which was introduced in Massy (1965), latent variates called principal components are obtained out of the correlation matrix $\mathbf{X}$, denoted by $\mathbf{R}$. PC regression avoids the problem of multicollinearity by reducing the dimension of the predictors. The loadings $\{\mathbf{p}_k\}_{k=1}^{a}$ are taken as a-first eigenvectors of the spectral decomposition of $\mathbf{R}$ matrix, and these vectors are the directions that maximize the variance of the principal components. The principal components are defined using the projections of the $\mathbf{X}$'s upon these directions. That is, the ith principal component of $\mathbf{X}$ is defined as $\mathbf{t}_k = \mathbf{X}\mathbf{p}_k$ so that $\mathbf{p}_k$ maximizes the variance of $\mathbf{t}_k$,

$$\max_{\mathbf{p}_k} \langle \mathbf{X}\mathbf{p}_k, \mathbf{X}\mathbf{p}_k \rangle = \max_{\mathbf{p}_k} \mathbf{p}_k^T \mathbf{X}^T \mathbf{X}\mathbf{p}_k$$

with $\mathbf{p}_k^T\mathbf{p}_k = 1$ y $\mathbf{p}_k^T\mathbf{p}_l = 0$, $l < k$. The principal components represent the selection of a new coordinate system obtained when rotating the original system of axes, $X_1, X_2, \ldots, X_p$. All of the loadings or principal directions are then obtained, $\mathbf{P} = [\mathbf{p}_1|\mathbf{p}_2|\cdots|\mathbf{p}_a]_{p\times a}$, as are the projections of the $X_i's$ on $\mathbf{p}_k's$, that is, all of the principal components, $\mathbf{T} = [\mathbf{t}_1|\mathbf{t}_2|\cdots|\mathbf{t}_a]_{n\times a}$, with the restrictions $\langle \mathbf{t}_k, \mathbf{t}_l \rangle = 0$ and $\langle \mathbf{t}_k, \mathbf{t}_k \rangle = Var(\mathbf{t}_k) = \lambda_k$, with λ_k: the eigenvalues associated with the eigenvectors P_k with $\lambda_1 \geq \lambda_2 \geq \ldots, \lambda_a$. A regression model of $\mathbf{Y}$ is then adjusted against the latent variates $\mathbf{T}$. Then, the response for $\mathbf{Y}$-new ones is predicted associated with new observations of the predictors vector. In PC regression, the principal components in the predictor space $\mathbf{X}$'s are used without taking into account the information of the responses $\mathbf{Y}$'s.

PLS regression was introduced in Wold (1975) and applied in the economic and social sciences fields. However, due to the contributions made by his son in Wold, Albano, Dunn, Edlund, Esbensen, Geladi, Hellberg, Johansson, Lindberg & Sjöström (1984), it gained great popularity in the area of chemometrics, where data characterized by many predictor variables with multicollinearity problems and few available observations are analyzed. This happens in many studies of imaging analysis. The PLS regression methodology generalizes and combines

characteristics of Principal Component Analysis (PCA) and Multiple Regression Analysis (MLR). Its demand and evidence has increased and it is being applied in many scientific areas. PLS regression is similar to the canonic correlation analysis (CCA), but instead of maximizing the correlation, it maximizes the covariance between the components. That is, $\mathbf{p}$ and $\mathbf{q}$ directions are found so that

$$\max_{\mathbf{p},\mathbf{q}} \langle \mathbf{Xp}, \mathbf{Yq} \rangle = \max_{\mathbf{p},\mathbf{q}} \mathbf{p}^T \mathbf{X}^T \mathbf{Yq}$$

subject to $\|\mathbf{p}\| = \|\mathbf{q}\| = 1$

In general, the PLS regression is a two-phase process. First, the predictor matrix $\mathbf{X}$ is transformed with the help of the vector of response variables, $\mathbf{Y}$, in a matrix of latent, non-correlated variables $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p)$, called PLS components. This distinguishes it from the PLS regression, in which the components are obtained using only the predictor matrix, $\mathbf{X}$. Second, the estimated regression model is adjusted using the original response vector and the PLS components as predictors, and then, response for $\mathbf{Y}$'new ones associated with future observations of the repetition vector are of predict. A reduction of dimensionality is obtained directly on the PLS components because they are orthogonal, and the number of components necessary for the regression analysis is much lower than the number of original predictors. The process of maximizing the covariance instead of the correlation prevents the possible problem of numeric instability that can appear when using correlation, which is due to the division of covariances by variances that may be too small. The directions of the maximum covariance p and q among the PLS components can be found by the following eigen-decomposition problem:

$$\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{Xp} = \lambda \mathbf{p} \quad \text{and} \quad \mathbf{Y}^T \mathbf{X} \mathbf{X}^T \mathbf{Yq} = \lambda \mathbf{q}$$

with $\|\mathbf{p}\| = \|\mathbf{q}\| = 1$. The latent variates (or PLS components) are calculated by projecting the $\mathbf{X}$ and $\mathbf{Y}$ data in the $\mathbf{p}$ and $\mathbf{q}$ directions, that is, $\mathbf{t} = \mathbf{Xp}$ and $\mathbf{u} = \mathbf{Yq}$ results in all latent components being obtained such that $\mathbf{T} = \mathbf{XP}$ and $\mathbf{U} = \mathbf{YQ}$.

3. Geometrical Structure of $\mathrm{Sym}^+(m)$

A summary will now be given of some of the basic results of (Schwartzman 2006) on the geometric structure of the $\mathrm{Sym}^+(m)$ set as a Riemannian manifold. The $\mathrm{Sym}^+(m)$ space is a sub-manifold of the Euclidian space $\mathrm{Sym}(m)$. Geometrically, the $\mathrm{Sym}^+(m)$ and $\mathrm{Sym}(m)$ spaces are differential manifolds of $m(m+1)/2$ dimensions, and they are homeomorphically related by an exponential and logarithmic transformation matrix. For any matrix $\mathbf{A} \in \mathrm{Sym}(m)$, its exponential matrix is given by $\exp(\mathbf{A}) = \sum_{k=1}^{\infty} \frac{\mathbf{A}^k}{k!} \in \mathrm{Sym}^+(m)$. Reciprocally, for any matrix $\mathbf{S} \in \mathrm{Sym}^+(m)$, there is a $\log(\mathbf{S}) = \mathbf{A} \in \mathrm{Sym}(m)$, such that $\exp(\mathbf{A}) = \mathbf{S}$.

For responses in Euclidian spaces in non-parametric standard regression models, $E[\mathbf{S}|X = x]$ is estimated. However, for responses on a curved space, the conditional expectancy of $\mathbf{S}$, given $x = x$, cannot be defined. For $\mu(x) = E[\mathbf{S}|X = x]$,

a tangent vector is introduced in $\mu(x)$ on $\text{Sym}^+(m)$. For a small scalar $\delta > 0$, the differentiable map $C : (-\delta, \delta) \longrightarrow \text{Sym}^+(m), \quad t \to C(t)$, is considered such that $C(0) = \mu(x)$. A tangent vector in $\mu(x)$ is defined as the derivative of the soft curve $C(t)$, with respect to t, valued at $t = 0$. The set of all tangent vectors in $\mu(x)$ is called the tangent space of $\text{Sym}^+(m)$ in $\mu(x)$, and it is denoted by $T_{\mu(x)}\text{Sym}^+(m)$. This space can be identified by a copy of $\text{Sym}(m)$. The $T_{\mu(x)}\text{Sym}^+(m)$ space is equipped with an internal product $\langle \ . \ , \ . \ \rangle$, called a Riemannian metric, which varies softly from point to point. For example, the Frobenius metric can be used as a Riemannian metric. For a given Riemannian metric, $\langle \mathbf{u} \ , \ \mathbf{v} \rangle$ is calculated for any $\mathbf{u}$ and $\mathbf{v}$ in $T_{\mu(x)}\text{Sym}^+(m)$, and then, the length of the soft curve $C(t) : [t_0, t_1] \longrightarrow \text{Sym}^+(m)$ is calculated, which is equal to: $\|C(t)\| = \int_{t_0}^{t_1} \sqrt{\langle \dot{C}(t), \dot{C}(t) \rangle} dt$, where $\dot{C}(t)$ is the derivative of $C(t)$, with respect to t. A geodesic is a soft curve in $\text{Sym}^+(m)$ with tangent vectors that do not change in length or direction along the curve. For any $\mathbf{u} \in T_{\mu(x)}\text{Sym}^+(m)$, there is a single geodesic, denoted by $\gamma_{\mu(x)}(t; \mathbf{u})$, with a dominion that contains the range $[0, 1]$, such that $\gamma_{\mu(x)}(0; \mathbf{u}) = \mu(x)$ and $\dot{\gamma}_{\mu(x)}(0; \mathbf{u}) = \mathbf{u}$.

The exponential Riemannian map is defined as

$$\text{Exp}_{\mu(x)} : T_{\mu(x)}\text{Sym}^+(m) \longrightarrow \text{Sym}^+(m) \ \ ; \ \mathbf{u} \longrightarrow \text{Exp}_{\mu(x)}(\mathbf{u}) = \gamma_{\mu(x)}(1; \mathbf{u}) \quad (1)$$

The inverse of the exponential Riemannian map, called a Riemannian logarithmic map, is defined as

$$\text{Log}_{\mu(x)} : \text{Sym}^+(m) \longrightarrow \ T_{\mu(x)}\text{Sym}^+(m) \ \ ; \ \mathbf{S} \longrightarrow \text{Log}_{\mu(x)}(\mathbf{S}) = \mathbf{u} \quad (2)$$

such that $\text{Exp}_{\mu(x)}(\mathbf{u}) = \mathbf{S}$. Finally, the shortest distance between 2 points $\mu_1(x)$ and $\mu_2(x)$ in $\text{Sym}^+(m)$, is called the geodesic distance and is denoted by $g(\mu_1(x), \mu_2(x))$, which satisfies

$$d_g^2(\mu_1(x), \mu_2(x)) = \langle \text{Log}_{\mu_1(x)}\mu_2(x), \text{Log}_{\mu_1(x)}\mu_2(x) \rangle = \|\text{Log}_{\mu_1(x)}\mu_2(x)\|_g^2 \quad (3)$$

where $d_g^2(. \ , \ .)$, denoted the geodesic distance.

The residual from $\mathbf{S}$ with respect to $\mu(x)$, denoted by $\varepsilon_\mu(x)$, is defined as $\varepsilon_\mu(x) = \text{Log}_{\mu(x)}\mathbf{S} \in T_{\mu(x)}\text{Sym}^+(m)$. The vectorization of $C = [c_{ij}] \in \text{Sym}(m)$ is defined as $\text{Vecs}(C) = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1m} & c_{22} & \dots & c_{2m} & \dots & c_{mm} \end{bmatrix}^T \in \mathbb{R}^{\frac{m(m+1)}{2}}$. The conditional expectancy of $\mathbf{S}$, given $\mathbf{x} = x$, is defined as the matrix $\mu(x) \in \text{Sym}^+(x)$, such that

$$E[\text{Log}_{\mu(x)}\mathbf{S}|X = x] = E[\varepsilon_\mu(x)|X = x] = \mathbf{0}_{m \times m} \quad (4)$$

where the expectancy is taken component by component with respect to the $m(m+1)$-vector aleatory multivaried $\text{Vecs}[\text{Log}_{\mu(x)}S] \in \mathbb{R}^{\frac{m(m+1)}{2}}$.

3.1. Regression Model for Response Data in $\text{Sym}^+(m)$

Because the DTs are in a non-linear space, it is theoretically and computationally difficult to develop a formal statistical framework that includes estimation

theory and hypothesis tests where by a set of covariates are used to directly predict DTs as responses. With the recently developed log-Euclidian metric Arsigny, Fillard, Pennec & Ayache (2006), DTs can be transformed from non-linear space into logarithmic matrices in a Euclidian space. Zhu et al. (2009) developed a regression model with the log-transformation of the DTs as the response. The model was based on a semi-parametric method, which avoids the specification of parametric distributions for aleatory log-transformed DTs. Inference processes have been proposed for estimating the regression coefficients and test statistics of this model to contrast linear hypotheses of unknown parameters as well as to test processes based on re-sampling methods to simultaneously evaluate the statistical significance of linear hypotheses throughout large ROIs. The procedure for the laying out of the local intrinsic polynomial regression model (RPLI) for SPD matrices as a response is described below, ver Zhu et al. (2009).

The procedure to estimate $\mu(x) = E[\mathbf{S}|X = x_0]$ in the RPLI model will now be described. Because $\mu(x)$ is on a curved space, it cannot be directly expand to $\mu(x)$ in $\mathbf{x} = x_0$ using a Taylor series. Instead, the Riemannian logarithmic map of $\mu(x)$ in $\mu(x_0)$ on the space $T_{\mu(x)}\text{Sym}^+(m)$ is considered, that is, we are considering $\text{Log}_{\mu(x_0)}\mu(x) \in T_{\mu(x)}\text{Sym}^+(m)$. Because $\text{Log}_{\mu(x_0)}\mu(x)$ occupies a different tangent space for each value of $\mathbf{X}$, it can be transported from the common tangent space $T_{I_m}\text{Sym}^+(m)$ through the parallel transport given by:

$$\Phi_{\mu(x_0)} : T_{\mu(x_0)}\text{Sym}^+(m) \longrightarrow T_{I_m}\text{Sym}^+(m); \\ \text{Log}_{\mu(x_0)}\mu(x) \longrightarrow \Phi_{\mu(x_0)}(\text{Log}_{\mu(x_0)}\mu(x)) = Y(x) \quad (5)$$

Its inverse is given by $\text{Log}_{\mu(x_0)}\mu(x) = \Phi^{-1}_{\mu(x_0)}(Y(x)) \in T_{\mu(x_0)}\text{Sym}^+(m)$.

For $\text{Log}_{\mu(x_0)}\mu(x_0) = O_m \in T_{\mu(x_0)}Sym^+(m)$, because $\Phi_{\mu(x_0)}(O_m) = Y(x_0) = O_m$ and because $Y(x)$ y $Y(x_0)$ are in the same tangent space $T_{I_m}\text{Sym}^+(m)$, a Taylor series expansion can be used for $Y(x)$ in x_0. The following is obtained:

$$\text{Log}_{\mu(x_0)}\mu(x) = \Phi^{-1}_{\mu(x_0)}(Y(x)) \approx \Phi^{-1}_{\mu(x_0)}\left(\sum_{k=1}^{k_0} Y^{(k)}(x_0)(x - x_0)^k\right) \quad (6)$$

with k_0 as a whole and $Y^{(k)}$ as the kth derivative of $Y(x)$ with respect to x divided by por $k!$. Equivalently,

$$\mu(x) = \text{Exp}_{\mu(x_0)}\left(\Phi^{-1}_{\mu(x_0)}(Y(x))\right) = \\ \text{Exp}_{\mu(x_0)}\left(\Phi^{-1}_{\mu(x_0)}\left(\sum_{k=1}^{k_0} Y^{(k)}(x_0)(x - x_0)^k\right)\right) = \mu\left(x, \alpha(x_0), k_0\right) \quad (7)$$

where $\alpha(x_0)$-contains all the parameters in $\{\mu(x_0), Y^{(1)}(x_0), \ldots, Y^{(k)}(x_0)\}$.

For a set of vectors in $T_{\mu(x)}\text{Sym}^+(m)$, various Riemannian metrics can be defined. Among these metrics is the log-Euclidian metric, and some of its basic properties will now be reviewed. Notations exp(.) and log(.) are used to represent the exponential and log matrices, respectively; Exp and Log are used to

represent the exponential and logarithmic maps, respectively. The differential of the logarithmic matrix in $\mu(x) \in \text{Sym}^+(m)$ is denoted by $\partial_{\mu(x)}\log.(\mathbf{u})$, which acts on an infinitesimal movement $\mathbf{u} \in T_{\mu(x)}\text{Sym}^+(m)$. The log-Euclidian metric on $\text{Sym}^+(m)$ is defined as:

$$\langle \mathbf{u}, \mathbf{v} \rangle := tr\left[(\partial_{\mu(x)}\log.\mathbf{u})(\partial_{\mu(x)}\log.\mathbf{v})\right] \tag{8}$$

for $\mathbf{u}, \mathbf{v} \in T_{\mu(x)}\text{Sym}^+(m)$.

The geodesic $\gamma_{\mu(x)}(t;\mathbf{u})$-is given by:

$$\gamma_{\mu(x)}(t;\mathbf{u}) := \exp\left[\log(\mu(x)) + t\partial_{\mu(x)}\log.\mathbf{v}\right] \ , \quad \forall t \in \mathbb{R} \tag{9}$$

The differential of the exponential matrix is denoted by $\partial_{\log(\mu(x))}\exp.(\mathbf{A})$, in $\log(\mu(x)) \in \text{Sym}(m) = T_{\mu(x)}\text{Sym}^+(m)$ which acts on an infinitesimal movement $\mathbf{A} \in T_{\log(\mu(x))}\text{Sym}^+(m)$. The exponential and logarithmic Riemannian maps are defined, respectively, as follows: for $\mathbf{S} \in \text{Sym}^+(m)$,

$$\begin{aligned} \text{Exp}_{\mu(x)}(\mathbf{u}) &:= \exp\left[\log(\mu(x)) + \partial_{\mu(x)}\log.(\mathbf{u})\right]; \\ \text{Log}_{\mu(x)}(\mathbf{S}) &:= \partial_{\log(\mu(x))}\exp\left[\log(\mathbf{S}) - \log(\mu(x))\right] \end{aligned} \tag{10}$$

For $\mu(x)$ and $\mathbf{S} \in \text{Sym}^+(m)$, the geodesic distance is given by:

$$d_g^2(\mu(x), \mathbf{S}) := tr\left[(\log\,\mu(x) - \log(\mathbf{S}))^{\otimes 2}\right] \tag{11}$$

with $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ and with $\mathbf{a}$-vector. For two matrices $\mu(x)$ and $\mu(x_0) \in \text{Sym}^+(m)$ and any $\mathbf{u}_{\mu(x_0)} \in T_{\mu(x_0)}\text{Sym}^+(m)$, the parallel transport is defined as follows:

$$\begin{aligned} \Phi_{\mu(x_0)} : T_{\mu(x_0)}\text{Sym}^+(m) &\longrightarrow T_{I_m}\text{Sym}^+(m); \\ \mathbf{u}_{\mu(x_0)} &\longrightarrow \Phi_{\mu(x_0)}(\mathbf{u}_{\mu(x_0)}) := \partial_{\mu(x_0)}\log.\left(U_{\mu(x_0)}\right) \end{aligned}$$

If $\mathbf{u}_{\mu(x_0)} = \text{Log}_{\mu(x_0)}\mu(x) \in T_{\mu(x_0)}\text{Sym}^+(m)$, then

$$Y(x) = \Phi_{\mu(x_0)}\left(\text{Log}_{\mu(x_0)}\mu(x)\right) = \log\,\mu(x) - \log\,\mu(x_0) \tag{12}$$

and $\mu(x) = \exp\left[\log\,\mu(x_0) + Y(x)\right]$.

The residual of $\mathbf{S}$ with respect to $\mu(x)$ is defined as: $\varepsilon_\mu(x) := \log(\mathbf{S}) - \log(\mu(x))$ with $E[\log\mathbf{S}|X = x] = \log\,\mu(x)$. The model RPLI is defined as:

$$\log(\mathbf{S}|x) = \log(\mu(x)) + \varepsilon_\mu(x) \tag{13}$$

with $E[\varepsilon_\mu(x)] = 0$, which indicates that $E[\log\mathbf{S}|X = x] = \log(\mu(x))$.

4. The PLS Regression Model

Suppose we have n DTs, denoted by $\mathbf{T}_i : i = 1, 2, \ldots, n$, obtained from a voxel correspondent with a normalized and especially re-oriented DTI from n subjects. The log-transformation of T_k is then obtained, which is denoted by

$$\mathbf{L}_{T,i} = (L^i_{T_{(1,1)}}, L^i_{T_{(1,2)}}, L^i_{T_{(1,3)}}, L^i_{T_{(2,2)}}, L^i_{T_{(2,3)}}, L^i_{T_{(3,3)}})^T \tag{14}$$

where $L^i_{T_{(j,k)}}$-denotes the (j,k)-element of the logarithm matrix of $\mathbf{T}_k$. For each individual, a set of covariates of interest is observed as well.

In studies of medical images, many demographic or clinical measurements are normally observed for different patients considered in a certain study. The amount of available information is abundant, and there may be problems of linear dependences between the covariates of interest, which generates the problem of multicollinearity. In addition, available data to analyze the information are scarce. For the log-transformed DTs, a linear model is considered, which is given by:

$$\underset{1\times 6}{\mathbf{L}_{T,i}} = \underset{1\times pp\times 6}{\mathbf{x}_i\ \boldsymbol{\beta}} + \underset{1\times 6}{\boldsymbol{\varepsilon}_i}\ , \quad i = 1, 2, \ldots, n \tag{15}$$

or

$$\underset{n\times 6}{\mathbf{L}_T} = \underset{n\times pp\times 6}{\mathbf{X}\ \mathbf{B}} + \underset{n\times 6}{\boldsymbol{\varepsilon}} \tag{16}$$

with $E[\boldsymbol{\varepsilon}|\mathbf{x}] = \mathbf{0}_{n\times p}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}|\mathbf{x}) = \Sigma_{np\times np}$ and where $\mathbf{X}$, $\mathbf{Y}$=$\mathbf{L}$, $\mathbf{B}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\Sigma}$, are matrices representing the covariates, responses, regression coefficients, the model errors and covariance of $\boldsymbol{\varepsilon}|\mathbf{x}$.

Compared to the general lineal model, the model, based on the conditional mean and covariance in equation (16)does not assume any distributional suppositions for the image measurements.

If $\boldsymbol{\theta}_{(6p+21)\times 1}$ is the vector of unknown parameters contained in $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, then to estimate $\boldsymbol{\theta}$, the objective function given by:

$$l_n(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n}\left(\log|\boldsymbol{\Sigma}| + (L_{T,i} - \boldsymbol{\beta}\mathbf{x}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{L}_{T,i} - \boldsymbol{\beta}\mathbf{x}_i)\right) \tag{17}$$

is maximized using the iterative algorithm proposed by Li et al. (2009).

The regression model (16) has been adjusted using existing algorithms for PC and PLS regression, following the steps described in Section 2.2 and taking into account the log-transformations on the original data to transfer them to a Euclidian space.

4.1. Evaluation of the PLS Regression Model with Simulated Data

The behavior of the PLS regression model is evaluated with sets of simulated data, and predicted results are compared with those obtained using the PC technique in the case of a design matrix of full range.

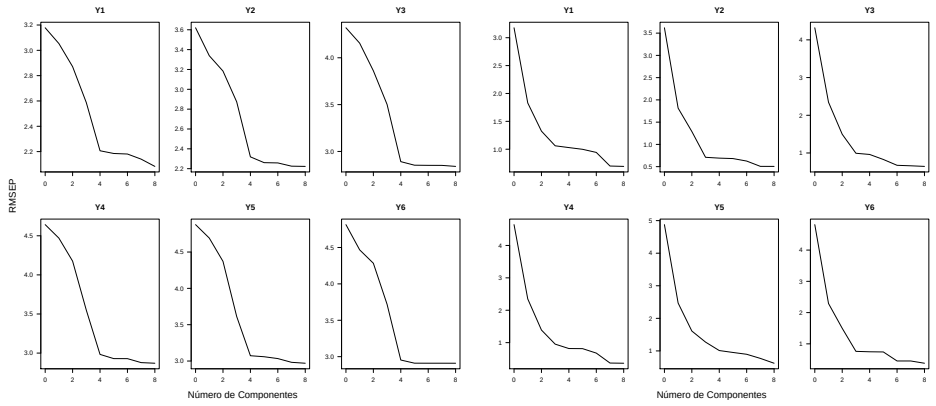
The settings considered to simulate the data are the following. First, a sample of SPD matrices with a size of $n = 20$ with $k = 15$ covariates was generated from a multivariate normal distribution with a mean of zero and a covariance structure given by $\boldsymbol{\Sigma} = 0.6\mathbf{I}_6$. Then, the sample size was increased to $n = 30$, and the number of covariates was increased from $k = 15$ to $k = 40$, with a covariance structure given by $\boldsymbol{\Sigma} = 0.3\mathbf{I}_6 + 0.61_6\mathbf{1}_6^T$, with $\mathbf{1}_6$, a vector of ones. In both settings, the values for the coefficients of beta were used in the matrix given by

$p \times 6$, $\beta_k = [1 + 0.1 \times (k-1)]^T$. The exponential of $\mathbf{\Sigma}$ was calculated to ensure its positive definiteness. Results obtained in each scenario are expounded below.

For the first setting, shown in Table 1, the percentages of variance explained by each of the latent components through PC and PLS regression demonstrate that PC explains more of the variability of $\mathbf{X}$ than PLS regression, which is a typical result. In Table 2, the PLS components explain a higher percentage of the variability of $\mathbf{Y}$ than the PC components; with two components, more than 80% of the variability in $\mathbf{Y}$ and approximately 20% of the variability in $\mathbf{X}$ is explained. Figure 1 shows the graphs of the square root of the prediction middle quadratic error (RMSEP) against the number of components used in the cross validation (CV). Here, it can be observed that in PC, approximately four components would be needed to explain a majority of the variability in the data. However, in PLS regression, three components are needed in most cases. In general, few repetition are shown through this illustration of the repetition results obtained by each method, when compared with the simulation. Figure 2 shows the graphs of the predicted data with the observed responses. A greater precision in the adjustment can be observed when PLS regression is used. For the second setting, Table 3 shows the percentages of variance explained by each of the latent components using PC and PLS regression. Again, PC explains more of the variability of $\mathbf{X}$ than PLS regression. Table 4 shows that the PLS components explain a greater percentage of the variability of $\mathbf{Y}$ than the PLS components. In five components, more than 60% of the variability in $\mathbf{Y}$ and approximately 35% of the variability in $\mathbf{X}$ is explained. Figure 3 shows the graphs for the RMSEP against the number of components. It can be observed that in PC, approximately 7 components would be needed to explain most of the variability of the data, while in PLS regression, five components are needed in most cases. Figure 4 shows the graphs of the predicted data along with the observed values of the responses; a greater precision in the adjustment can be observed when PLS regression is used.

Table 1: Percentages of variance explained by each component.

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8
PC	17.57	15.55	13.59	12.46	11.16	9.16	6.81	4.64
PLS	14.27	9.93	10.16	13.45	12.60	5.75	4.46	7.07



(a) PC regression (b) PLS regression

FIGURE 1: RMSEP versus number of components by PC regression and PLS regression

TABLE 2: Percentages of variance explained cumulated of $\mathbf{X}$ and $\mathbf{Y}$ for the components by PC and PLS regression.

		Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8
PC	X	17.57	33.11	46.70	59.16	70.32	79.48	86.28	90.93
PLS	X	14.27	24.20	34.37	47.82	60.43	66.17	70.64	77.70
PC	Y1	7.69	18.38	33.74	51.79	52.72	52.91	54.64	57.04
PLS	Y1	66.85	82.64	88.85	89.51	90.13	91.21	95.17	95.26
PC	Y2	14.95	22.65	36.98	58.96	60.99	61.09	62.21	62.29
PLS	Y2	74.87	87.30	96.16	96.35	96.46	97.01	98.04	98.05
PC	Y3	7.45	20.12	34.30	55.21	56.38	56.43	56.44	56.77
PLS	Y3	70.51	88.00	94.72	95.05	96.33	97.57	97.67	97.78
PC	Y4	7.30	19.10	41.57	58.71	60.19	60.20	61.57	61.78
PLS	Y4	74.39	91.05	95.78	96.90	96.92	97.87	99.36	99.39
PC	Y5	7.44	19.65	45.13	60.30	60.66	61.30	62.61	62.93
PLS	Y5	74.38	89.10	93.22	95.70	96.19	96.62	97.51	98.38
PC	Y6	13.89	20.83	40.31	62.35	63.45	63.46	63.46	63.47
PLS	Y6	77.35	90.32	97.51	97.60	97.63	99.12	99.12	99.38

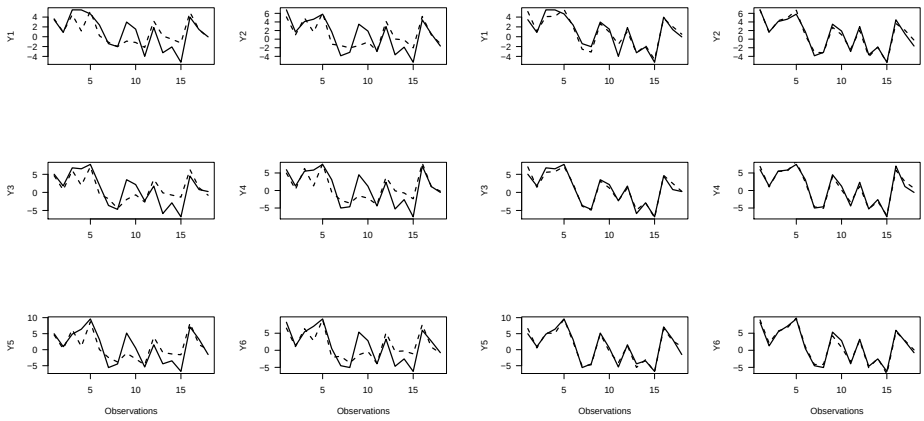
TABLE 3: Percentages of variance explained by each component, 2.

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10
PC	12.81	9.33	8.74	7.42	7.22	6.39	6.33	5.12	4.97	4.44
PLS	10.63	8.65	6.39	5.21	3.85	5.34	4.88	5.36	5.32	5.00

5. Conclusions and Recommendations

A PLS linear regression model is proposed in this article to study the relationship between a large set of covariates of interest in a Euclidian space with a set of response variables in a symmetric Riemannian space. The theory of exponential and Riemannian maps has been used to transform data from a non-Euclidian space into a Euclidian space of symmetrical matrices, where the methodology has been developed. Results indicate support for the proposed methodology as compared to

a technique using regression by major components, as has been observed in classic situations of data analysis in euclidean spaces with matrices of covariates presenting high multicollinearity, or in problems with a low number of observations and many covariates. In future works, we will investigate more realistic models, such as non-linear PLS models for the types of SPD matrix data discussed in this study and other types of manifold-valued data, such as data obtained by geometric representations of objects via medial axial representation (m-rep), orthogonal rotation groups, and other methods. The illustration presented in this article for simulated data favorably sheds light on the results that can be obtained by applying these types of models to real data.

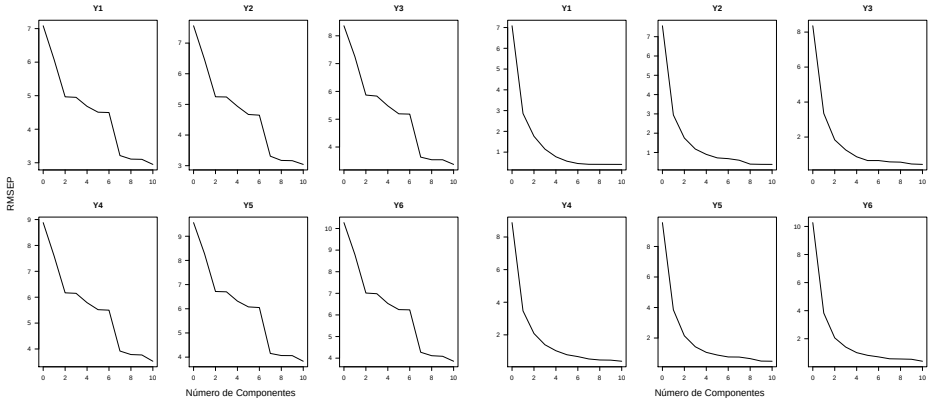


(a) PC regression (b) PLS regression

FIGURE 2: Predicted values with the observables values by PC regression and PLS regression. Solid lines: Observed, dashed lines: Predicted.

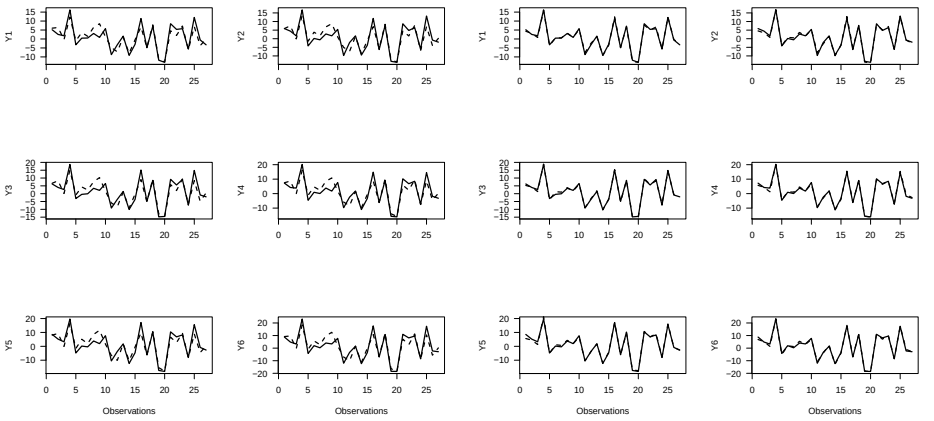
TABLE 4: Percentages of variance explained cumulated of $\mathbf{X}$ and $\mathbf{Y}$ for the components by PC and PLS regression, 2.

		Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10
PC	X	12.81	22.14	30.88	38.30	45.52	51.90	58.23	63.35	68.33	72.77
PLS	X	10.63	19.28	25.67	30.88	34.73	40.07	44.95	50.32	55.64	60.64
PC	Y1	26.52	50.85	51.17	56.31	59.51	59.69	79.40	80.74	80.83	82.65
PLS	Y1	83.70	93.81	97.39	98.80	99.37	99.59	99.66	99.66	99.66	99.67
PC	Y2	26.97	51.87	51.99	57.41	61.87	62.25	80.86	82.42	82.52	83.83
PLS	Y2	84.85	94.65	97.57	98.58	99.09	99.19	99.37	99.72	99.74	99.74
PC	Y3	24.82	50.72	51.34	57.02	61.40	61.56	81.08	82.05	82.05	83.70
PLS	Y3	83.92	95.16	97.72	98.91	99.38	99.38	99.52	99.54	99.70	99.73
PC	Y4	27.00	51.74	52.05	57.50	61.39	61.65	80.51	81.84	81.99	84.23
PLS	Y4	84.74	94.50	97.54	98.67	99.23	99.44	99.66	99.73	99.74	99.81
PC	Y5	25.11	50.70	50.90	56.36	59.61	59.97	81.14	81.93	81.96	83.96
PLS	Y5	83.80	94.97	97.77	98.77	99.14	99.37	99.38	99.54	99.74	99.75
PC	Y6	26.75	53.38	53.80	59.58	63.02	63.15	82.70	83.96	84.18	85.90
PLS	Y6	86.10	95.97	98.12	99.03	99.37	99.53	99.69	99.71	99.73	99.85



(a) PC regression (b) PLS regression

FIGURE 3: RMSEP versus number of components by PC regression and PLS regression, 2.



(a) PC regression (b) PLS regression

FIGURE 4: Predicted values with the observables values by PC regression and PLS regression, 2. Solid lines: Observed, dashed lines: Predicted.

[Recibido: junio de 2012 — Aceptado: mayo de 2013]

References

Arsigny, V., Fillard, P., Pennec, X. & Ayache, N. (2006), 'Log-euclidean metrics for fast and simple calculus on diffusion tensors', *Magnetic Resonance in Medicine,* **56**, 411–421.

Barmpoutis, A., Vemuri, B. C., Shepherd, T. M. & Forder, J. R. (2007), 'Tensor splines for interpolation and approximation of DT-MRI with applications to segmentation of isolated rat hippocampi', *IEEE Transations on Medical Imaging,* **26**, 1537–1546.

Batchelor, P., Moakher, M., Atkinson, D., Calamante, F. & Connelly, A. (2005), 'A rigorous framework for diffusion tensor calculus', *Magnetic Resonance in Medicine,* **53**, 221–225.

Fingelkurts, A. A. & Kahkonen, S. (2005), 'Functional connectivity in the brain - is it an elusive concepts?', *Neuroscience and Biobehavioral Reviews,* **28**, 827–836.

Fletcher, P. T. & Joshi, S. (2007), 'Riemannian geometry for the statistical analysis of diffusion tensor data', *Signal Processing,* **87**, 250–262.

Grenander, U. & Miller, M. I. (1998), 'Computational anatomy: An emerging discipline', *Quarterly of Applied Mathematics,* **56**, 617–694.

Kim, P. T. & Richards, D. S. (2010), 'Deconvolution density estimation on spaces of positive definite symmetric matrices', *IMS Lecture Notes Monograph Series. A Festschrift of Tom Hettmansperger* .

Lepore, N., Brun, C. A., Chou, Y., Chiang, M., Dutton, R. A., Hayashi, K. M., Luders, E., Lopez, O. L., Aizenstein, H. J., Toga, A. W., Becker, J. T. & Thompson, P. M. (2008), 'Generalized tensor-based morphometry of HIV/AIDS using multivariate statistics on deformation tensors', *IEEE Transactions in Medical Imaging,* **27**, 129–141.

Li, Y., Zhu, H., Chen, Y., Ibrahim, J. G., An, H., Lin, W., Hall, C. & Shen, D. (2009), RADTI: Regression analysis of diffusion tensor images, *in* E. Samei & J. Hsieh, eds, 'Progress in Biomedical Optics and Imaging - Proceedings of SPIE', Vol. 7258.

Massy, W. F. (1965), 'Principal components regression in exploratory statistical research', *Journal of the American Statistical Association,* **64**, 234–246.

Pennec, X., Fillard, P. & Ayache, N. (2006), 'A Riemannian framework for tensor computing', *International Journal of Computer Vision,* **66**, 41–66.

Schwartzman, A. (2006), Random ellipsoids and false discovery rates: Statistics for diffusion tensor imaging data, PhD thesis, Stanford University.

Wold, H. (1975), 'Soft modeling by latent variables; the non-linear iterative partial least squares approach', *Perspectives in Probability and Statistics,* pp. 1–2.

Wold, S., Albano, C., Dunn, W.J., I., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W. & Sjöström, M. (1984), Multivariate data analysis in chemistry, *in* B. Kowalski, ed., 'Chemometrics', Vol. 138 of *NATO ASI Series*, Springer Netherlands, pp. 17–95.

Yuan, Y., Zhu, H., Lin, W. & Marron, J. S. (2012), 'Local polynomial regression for symmetric positive-definite matrices', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(4), 697–719.

Zhu, H. T., Chen, Y. S., Ibrahim, J. G., Li, Y. M. & Lin, W. L. (2009), 'Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging', *Journal of the American Statistical Association* **104**, 1203–1212.

La Revista Colombiana de Estadística agradece a las personas que no integran los Comités Editorial y Científico, por su colaboración en el volumen 35 (2012).

Andre Khuri, Ph.D.
Andrea Cardini, Ph.D.
Arijit Chaudhuri, Ph.D.
Barry Arnold, Ph.D.
Broderick O. Oluyede, Ph.D.
Claudia Neves, Ph.D.
Ching-Wen Chen, Ph.D.
David Cox, Ph.D.
Diana Bilková, Ph.D.
Eloísa Díaz-Francés, Ph.D.
Ewa Gudowska-Nowak, Ph.D.
Francisco Torres Avilés, Ph.D.
Federico Marquez, Ph.D.
G. Kumar Pandiyan, Ph.D.
Gilles Bourgault, Ph.D.
Gwo Dong Lin, Ph.D.
Hector Nuñez, Ph.D.
Hector Zapata, Ph.D.
Hugo Salinas, Ph.D.
Hongtu Zhu, Ph.D.
Ivana Mala, CSc.
John Hinde, Ph.D.
Jordan Stoyanov, Ph.D.
Kalliopi Mylona, Ph.D.
Kiyoshi Taniguchi, Ph.D.
Luis Guillermo Díaz, M.Sc.
Margaret A. Oliver, Ph.D.
Mauricio Sadinle, M.Sc.
Michael Smithson, Ph.D.
Miguel Bermejo, Ph.D.
Osvaldo Venegas, Ph.D.
Silvia Ferrari, Ph.D.
Simos Meintanis, Ph.D.
Simon Price, Ph.D.
Rajesh Singh, Ph.D.
Reza Modarres, Ph.D.
Rosemary Bailey, Ph.D.
Tahani Maturi, Ph.D.
Ying Yuan, Ph.D.