Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 295 a 305

Funciones de varianza y correlación bicuadrática para distribuciones normales

Biweight Variance and Correlation Functions for Normal Distributions

Carlos Eduardo Alonso[a], Jorge Martínez[b]

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

Resumen

En este trabajo se analiza el comportamiento del funcional ϱ asociado al estimador de correlación bicuadrático $-\widehat{\varrho}-$, asumiendo que se observan vectores aleatorios con distribución normal bivariada. Esto, con el objetivo de verificar si este estimador robusto es un estimador insesgado del coeficiente de correlación $-\rho-$.

El trabajo se desarrolló a partir de las propiedades de la función generadora de momentos de una distribución.

De acuerdo con los resultados, $\varrho > \rho$ cuando $\rho < 0$, $\varrho < \rho$ cuando $\rho > 0$, y $\varrho = 0$ cuando $\rho = 0$, e indican que el estimador propuesto $\widehat{\varrho}$ no es un estimador insesgado del coeficiente de correlación.

Lo anterior plantea como reto modificar el estimador $\widehat{\varrho}$ con el objetivo de obtener un estimador robusto insesgado o asintóticamente insesgado del coeficiente de correlación.

***Palabras clave*:** coeficiente de correlación, distribución truncada, estimación robusta, estimador M.

Abstract

In this paper, we have analized the behavior of the functional ϱ, associated to the biweight correlation estimator $-\widehat{\varrho}-$, assuming the sampled population has a bivariate normal distribution. The purpose is to verify if the estimator $\widehat{\varrho}$ is an unbiased estimator of the correlation coefficient ρ.

The results show $\varrho > \rho$ when $\rho < 0$, $\varrho < \rho$ when $\rho > 0$, y $\varrho = 0$ when $\rho = 0$. This results indicate $\widehat{\varrho}$ is not an unbiased estimator of the correlation coefficient.

***Key words*:** Correlation coefficient, M-estimate, Truncated distribution, Robust estimation.

[a]Profesor asistente. E-mail: cealonsom@unal.edu.co
[b]Profesor especial. E-mail: jmartinezc@unal.edu.co

1. Coeficiente de correlación bicuadrático

En la práctica es importante estudiar el desempeño de los procedimientos estadísticos ante incumplimientos de los supuestos, variaciones en los supuestos que son comunes en la cotidianidad de un usuario. Esto con el objetivo de hallar situaciones en las cuales no se recomienda usar la herramienta, y al tiempo plantear herramientas no tan sensibles ante el incumplimiento de los supuestos.

En este sentido, en Wei (2006) y Valcárcel (2007) se ha mostrado que el estimador clásico de la función de autocorrelación (FAC) es altamente sensible ante la presencia de valores extremos, sensibilidad que contrasta con la relevancia de este estimador en el análisis de series de tiempo, porque a partir de los valores estimados de la FAC se puede identificar el modelo, se construyen los estimadores de los parámetros y se analizan los residuales, entre otros. De lo anterior se ha planteado un estimador robusto del coeficiente de correlación, y posteriormente un estimador robusto de la FAC, estimador que se presenta a continuación.

Dada una muestra aleatoria $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, el estimador del coeficiente de correlación propuesto está dado por,

$$\widehat{\varrho}_{xy} = \frac{\delta_{xy}^2}{\delta_{xx}\delta_{yy}} \tag{1}$$

con

$$\delta_{xy}^2 = nk^2[MAD_x][MAD_y]\frac{\sum_{i=1}^{n}\psi(z_{x_i})\psi(z_{y_i})}{\left(\sum_{i=1}^{n}\psi'(z_{x_i})\right)\left(\sum_{i=1}^{n}\psi'(z_{y_i})\right)} \tag{2}$$

donde k es una constante positiva de estandarización[1], $\psi(\cdot)$ y $\psi'(\cdot)$ son la función de bicuadrática planteada por Beaton & Tukey (1974) y su derivada, respectivamente. La función bicuadrática está dada por

$$\psi(z) = \begin{cases} z(1-z^2)^2 & \text{para } |z| < 1 \\ \\ 0 & \text{para } |z| \geqq 1 \end{cases} \tag{3}$$

con $z_t = \frac{x_t - Med_x}{kMAD_x}$, $Med_x = mediana\{x_1, x_2, \ldots, x_n\}$ y $MAD_x = mediana|x_t - Med_x|$. Para utilizar el estimador $\widehat{\varrho}_{xy}$ en la estimación de la FAC $-\rho_h-$, para una serie de tiempo estacionaria $y_1, y_2, \ldots, y_T$, la ecuación (2) se transforma en

$$\varphi_h = Tk^2MAD_y^2\frac{\sum_{t=1}^{T-|h|}\psi(z_{y_t})\psi(z_{y_{t+|h|}})}{\left(\sum_{t=1}^{T}\psi'(z_{y_t})\right)\left(\sum_{j=1}^{T-|h|}\psi'(z_{y_{t+|h|}})\right)} \tag{4}$$

[1]La propuesta de Lax (1975) es $k = 9$, valor planteado desde las propiedades de una distribución $N(\mu, \sigma^2)$, donde $MAD \approx \frac{2}{3}\sigma$; así, si $k = 3d$, se tiene que $kMAD \approx d \times \sigma$, es decir al construir el estimador no se tienen en cuenta las observaciones a más de d desviaciones estándar de la media.

A partir de (4) el estimador de ρ_h está dado por

$$\widehat{\rho}_h = \frac{\varphi_h}{\varphi_0} \tag{5}$$

2. Resultados

Uno de los resultados intermedios del trabajo es la generalización del concepto de covarianza, resultado que se muestra porque a partir de este se desarrolla el funcional asociado a $\widehat{\varrho}_{xy}$.

2.1. Generalización de la covarianza del coeficiente de correlación

Definición 1 (ψ-*Covarianza*)**.** Dadas dos variables aleatorias X y Y, con función distribución conjunta y segundos ψ-momentos finitos ($E\left[\psi^2(X)\right] < \infty$ y $E\left[\psi^2(Y)\right] < \infty$), la ψ-covarianza entre X y Y se define como:

$$\gamma_{\psi_{XY}} = \frac{E\left[\psi(Z_x)\psi(Z_y)\right]}{E\left[\psi'(Z_x)\right]E\left[\psi'(Z_y)\right]} \tag{6}$$

Un caso particular de esta definición se obtiene haciendo $\psi(x) = x$, $\psi'(x) = \frac{\partial\psi(x)}{\partial x} = 1$, $Z_x = X - EX$ y $Z_y = Y - EY$, de donde $\gamma_{\psi_{XY}} = E\left[(X - EX)(Y - EY)\right]$, que es la definición de covarianza clásica[2].

Definición 2 (ψ-*Correlación*)**.** Dadas dos variables aleatorias X y Y, con $\gamma_{\psi_{XX}} < \infty$ y $\gamma_{\psi_{YY}} < \infty$, la ψ-correlación entre X y Y se define como:

$$\varrho_{\psi_{XY}} = \frac{\gamma_{\psi_{XY}}}{\left(\gamma_{\psi_{XX}}\gamma_{\psi_{YY}}\right)^{\frac{1}{2}}} \tag{7}$$

De lo desarrollado, es claro que el coeficiente de correlación de Pearson es un caso particular de esta definición.

2.2. Funcional asociado a $\widehat{\varrho}_{xy}$

Se define

$$\begin{aligned}\widehat{\tau}^2_{xy} &= n\frac{\sum_{i=1}^{n}\psi(z_{x_i})\psi(z_{y_i})}{\left(\sum_{i=1}^{n}\psi'(z_{x_i})\right)\left(\sum_{i=1}^{n}\psi'(z_{y_i})\right)} \\ &= \frac{\int\psi(z_{x_i})\psi(z_{y_i})dF_n}{\left(\int\psi'(z_{x_i})dF_n\right)\left(\int\psi'(z_{y_i})dF_n\right)} = T(F_n)\end{aligned} \tag{8}$$

[2]Esto se plantea inicialmente para polinomios.

donde F_n es la función de distribución muestral. Asociado a $\widehat{\tau}^2_{xy}$ se tiene el funcional

$$\tau^2_{xy} = \frac{\int \psi(z_{x_i})\psi(z_{y_i})dF}{\left(\int \psi'(z_{x_i})dF\right)\left(\int \psi'(z_{y_i})dF\right)} = T(F) \tag{9}$$

donde F es la función de distribución poblacional. Unido a lo anterior se puede mostrar que

$$\widehat{\varrho}_{xy} = \frac{\widehat{\tau}^2_{xy}}{\widehat{\tau}_{xx}\widehat{\tau}_{yy}} \tag{10}$$

A partir de las ecuaciones (9) y (10), se tiene que el funcional asociado al estimador planteado en (1) es el coeficiente de ψ-*Correlación* presentado en la definición 2, ecuación (7), con $\psi(\cdot)$ la función bicuadrática definida en (3).

El objetivo de este trabajo es analizar el comportamiento de este funcional, es decir

$$\varrho_{\psi_{XY}} = \frac{\gamma_{\psi_{XY}}}{\gamma_{\psi_{XX}}\gamma_{\psi_{YY}}} \tag{11}$$

asumiendo que el vector (X, Y) tiene distribución $N(0, 0, 1, 1, \rho)$. En lo que sigue se mencionarán $\gamma_{\psi_{XX}}$ y $\gamma_{\psi_{XY}}$ como funciones de varianza y covarianza bicuadrática respectivamente. En este mismo sentido, $\varrho_{\psi_{XY}}$ se llama en adelante coeficiente de correlación bicuadrático[3].

2.3. Varianza bicuadrática

Si se asume que el vector (X, Y) tiene distribución $N(0, 0, 1, 1, \rho)$, las variables aleatorias X y Y tienen distribución univariada $N(0, 1)$, y

$$Med_X = Med_Y = 0 \qquad \text{y} \qquad MAD_X = MAD_Y = 0,67448975$$

resultados que conllevan a que las variables estandarizadas estén dadas por $Z_x = \frac{X}{l}$ y $Z_y = \frac{Y}{l}$, $l = k \times 0.67448975$, k constante de estandarización definida en (2)[4].

Si se define la variable aleatoria $M = I_{\{X \leq l\}}X$ la varianza bicuadrática de X, se puede escribir en términos de M como

$$\gamma_{\psi_{XX}} = \frac{\frac{E_2}{l^2} - 4\frac{E_4}{l^4} + 6\frac{E_6}{l^6} - 4\frac{E_8}{l^8} + \frac{E_{10}}{l^{10}}}{\left(1 - 6\frac{E_2}{l^2} + 5\frac{E_4}{l^4}\right)^2} \tag{12}$$

donde $E_r = E(M^r)$. La función de densidad de M está dada por $f_M(m) = \frac{\phi(m)}{c_1} I_{\{|m|<l\}}$, $f_M(\cdot)$ es la función de densidad resultante de truncar a dos colas la función de densidad asociada a una distribución normal estándar $\phi(\cdot)$, y $c_1 = P(-l < X < l)$.

[3] Esto indica una distribución normal bivariada con parámetros $E(X) = E(Y) = 0$, $V(X) = V(Y) = 1$ y $Corr(X, Y) = \rho$.

[4] Los valores de k con los cuales se realizó este trabajo son $k = 3, 6, 9$.

Si se nota la derivada de orden r de la función generatriz de momentos de la variable M, por $m^{(r)}(t) = \frac{\partial E(e^{tM})}{\partial t^r}$, esta derivada cumple con la siguiente recurrencia

$$\begin{aligned} m^{(r)}(t) =& (r-1)m^{(r-2)}(t) + tm^{(r-1)}(t) \\ &+ (-l)^{r-1}\frac{e^{-\frac{1}{2}l^2}\left[e^{-lt} + (-1)^k e^{lt}\right]}{c_1(2\pi)^{\frac{1}{2}}} \quad \text{para} \quad r \geq 2 \end{aligned} \tag{13}$$

Esta ecuación permite obtener el valor de los momentos de la variable aleatoria $\mathbb{M}$; no es difícil observar que los momentos de orden impar son cero.

2.3.1. Varianza bicuadrática

A partir de la ecuación (13) se obtienen los momentos de la variable M, y a partir de estos, usando la ecuación (12), se hallan valores de la varianza bicuadrática para $k = 9, 6, 3$, resultados que se presentan en la tabla 1.

TABLA 1: Valores de la varianza bicuadrática para una distribución $N(0,1)$.

	Valor de k		
	3	6	9
$\gamma_{\psi_{XX}}$	0,491377330220	0,066819506434	0,027637604055

2.4. Coeficiente de correlación bicuadrático

Si se define el vector aleatorio $\mathbb{M} = (M_1, M_2)^T$, con $\mathbb{M} = I_{\{|X|<l,|Y|<l\}}\mathbb{X}$, $\mathbb{X} = (X, Y)$ vector aleatorio con distribución $N(0, 0, 1, 1, \rho)$, la función de covarianza bicuadrática de $\mathbb{X}$ en función del vector $\mathbb{M}$ está dada por,

$$\gamma_{\psi_{XY}} = \frac{\frac{E_{1,1}}{l^2} - 4\frac{E_{3,1}}{l^4} + 2\frac{E_{5,1}}{l^6} + 4\frac{E_{3,3}}{l^6} - 4\frac{E_{5,3}}{l^8} + \frac{E_{5,5}}{l^{10}}}{\left[1 - 6\frac{E_2}{l^2} + 5\frac{E_4}{l^4}\right]^2} \tag{14}$$

con $E_{r,h} = E(M_1^r M_2^h)$, $E_r = E(M_1^r) = E(M_2^r)$, para r y h enteros[5]. Análogo a lo realizado para el caso de la varianza bicuadrática, la distribución del vector aleatorio $\mathbb{M}$ es resultado de truncar una distribución normal bivariada; de lo anterior la función de densidad conjunta de $\mathbb{M}$ está dada por $f_{\mathbb{M}}(m_1, m_2) = \frac{\phi(m_1,m_2)}{c_2} I_{\{|X|<l,|Y|<l\}}$, donde $\phi(\cdot,\cdot)$ es la función de densidad conjunta de una distribución $N(0, 0, 1, 1, \rho)$, y $c_2 = P(|X| < l, |Y| < l)$.

[5] $E(M_1^r) = E(M_2^r)$, dado que X y Y tienen la misma distribución.

2.4.1. Covarianza bicuadrática - caso $\rho = 0$

Si se supone $\rho = 0$, de la definición del vector $\mathbb{M} = (M_1, M_2)$, se sigue que las variables aleatorias M_1 y M_2 son independientes. Este resultado conlleva a que la ecuación (14) se transforme en

$$\gamma_{\psi_{XY}/\rho=0} = \frac{\frac{E^2(M_1)}{l^2} - 4\frac{E(M_1^3)E(M_1)}{l^4} + 2\frac{E(M_1^5)E(M_1)}{l^6}}{\left[1 - 6\frac{E_2}{l^2} + 5\frac{E_4}{l^4}\right]^2} + \frac{4\frac{E^2(M_1^3)}{l^6} - 4\frac{E(M_1^5)E(M_1^3)}{l^8} + \frac{E^2(M_1^5)}{l^{10}}}{\left[1 - 6\frac{E_2}{l^2} + 5\frac{E_4}{l^4}\right]^2} = 0 \tag{15}$$

El denominador es distinto de cero, y en el numerador sólo se tienen momentos de orden impar. Dado que la función de densidad de M_1 es simétrica alrededor de cero, los valores esperados en el numerador son cero, de donde se tiene $\gamma_{\psi_{XY}/\rho=0} = 0$.

2.4.2. Covarianza bicuadrática - caso $\rho \neq 0$

A partir de la ecuación (14), el camino por seguir para el caso $\rho \neq 0$, es calcular los momentos conjuntos del vector $\mathbb{M}$, tarea que se realiza usando la función generatriz. El valor de los momentos univariados de M_1 y M_2, ya se desarrollaron en la sección 2.3.

El trabajo con distribuciones normales truncadas no es nuevo. Pearson inicialmente trabajó en los años de 1930 sobre estas distribuciones, con el propósito de generar algunas tablas (tomado de Rosenbaum 1961, p. 405); posteriormente trabajaron sobre este tipo de distribuciones Cohen (1955), Singh (1960), Rosenbaum (1961), Tallis (1961), Finney (1962) y Khatri & Jaiswal (1963). Si se nota $\mathbf{m} = (m_1, m_2)^T$ y $\mathbf{t} = (t_1, t_2)^T$, la función generatriz del vector $\mathbf{M}$ está dada por

$$G_{\mathbb{M}}(\mathbf{t}) = E\left(e^{\mathbf{t}^T\mathbb{M}}\right) = \frac{e^{\frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}}}{c_2}\int_{-l}^{l}\int_{-l}^{l}\frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)}e^{-\frac{1}{2}\left[(\mathbf{m}-\Sigma\mathbf{t})^T\Sigma^{-1}(\mathbf{m}-\Sigma\mathbf{t})\right]}dm_1dm_2$$

donde $|\cdot|$ indica determinante[6]. Para hacer más corta la escritura, las derivadas de la función generatriz se notan como $D^{(h,r)} = \frac{\partial^{r+h}G_{\mathbb{M}}(\mathbf{t})}{\partial t_2^r \partial t_1^h}$. Los momentos conjuntos de orden r y h de $\mathbb{M}$ (r y h enteros nonegativos), alrededor al origen, están dados por

$$E(M_1^h M_2^r) = D^{(h,r)}\Big|_{t_1=0,t_2=0} \tag{16}$$

[6]El valor de acotamiento $l = k \times 0,67448975$ usado en el cálculo de varianza bicuadrática, sección 2.2.

A partir de que la distribución normal cumple con las condiciones de regularidad (ver Bickel & Docksum 1977, p. 378), se tiene que

$$D^{(1,0)} = G_{\mathbb{M}}(\mathbf{t})(t_1 + \rho t_2) - \frac{\beta_1(l,t_1,t_2) - \beta_2(-l,t_1,t_2) - \rho\left[\beta_3(l,t_2,t_1) - \beta_4(-l,t_2,t_1)\right]}{c_2(2\pi)^{\frac{1}{2}}} \tag{17}$$

con

$$\beta_j(v,u,w) = e^{-\frac{v^2}{2(1-\rho^2)}+vu+\frac{\left[\rho v+(1-\rho^2)w\right]^2}{2(1-\rho^2)}} P(-|v| < \xi_j < |v|) \quad j = 1,2,3,4$$

donde ξ_1, ξ_2, ξ_3 y ξ_4 son variables aleatorias cuyas distribuciones se presentan a continuación:

$$\xi_1 \sim N\left(\rho l + (1-\rho^2)t_2; 1-\rho^2\right), \qquad \xi_2 \sim N\left(-\rho l + (1-\rho^2)t_2; 1-\rho^2\right)$$
$$\xi_3 \sim N\left(\rho l + (1-\rho^2)t_1; 1-\rho^2\right) \quad \text{y} \quad \xi_4 \sim N\left(-\rho l + (1-\rho^2)t_1; 1-\rho^2\right)$$

y para $h \geq 2$ y $r \geq 1$ se tiene

$$D^{(h,r)} = r\rho D^{(h-1,r-1)} + (h-1)D^{(h-2,r)} + D^{(h-1,r)}(t_1 + \rho t_2) - \frac{\left[l^{h-1}(\beta_1 + (-1)^h\beta_2)^{(0,r)} + \rho l^r(\beta_3 - (-1)^r\beta_4)^{(h-1,0)}\right]}{c_2(2\pi)^{\frac{1}{2}}} \tag{18}$$

donde

$$\frac{\partial^{r+s}(\beta_1 \pm \beta_2)}{\partial t_2^r \partial t_1^s} = (\beta_1 \pm \beta_2)^{(s,r)} \qquad \text{y} \qquad \frac{\partial^{m+n}(\beta_3 \pm \beta_4)}{\partial t_2^m \partial t_1^n} = (\beta_3 \pm \beta_4)^{(n,m)}$$

2.4.3. Valores de $\varrho_{\psi_{XY}}$

A partir de los desarrollos mostrados en la sección 2.4.2, se obtienen los valores de los momentos conjuntos del vector $\mathbb{M}$. Una vez calculados estos, se consiguieron los valores de la covarianza bicuadrática utilizando la ecuación (14), valores que hacen posible calcular el coeficiente de correlación bicuadrático mediante la ecuación (11). Los valores de la varianza bicuadrática ya habían sido obtenidos (ver sección 2.3.1).

Los resultados muestran que el valor de $\varrho_{\varphi_{XY}}(\rho) \to \rho$ cuando k crece[7] (ver figuras 1, 2 y tabla 2). Para $k = 9$ las diferencias entre $\varrho_{\varphi_{XY}}$ y ρ son de tal magnitud que las líneas se superponen, razón por la cual se muestra una ampliación de la misma gráfica en el cuadrante (0,4; 0,7) (ver figura 2).

[7]La línea delgada indica la identidad, es decir $\rho = \rho$, los puntos el valor de ϱ; lo ideal es que $\varrho \approx \rho$, es decir que las dos líneas coincidan.

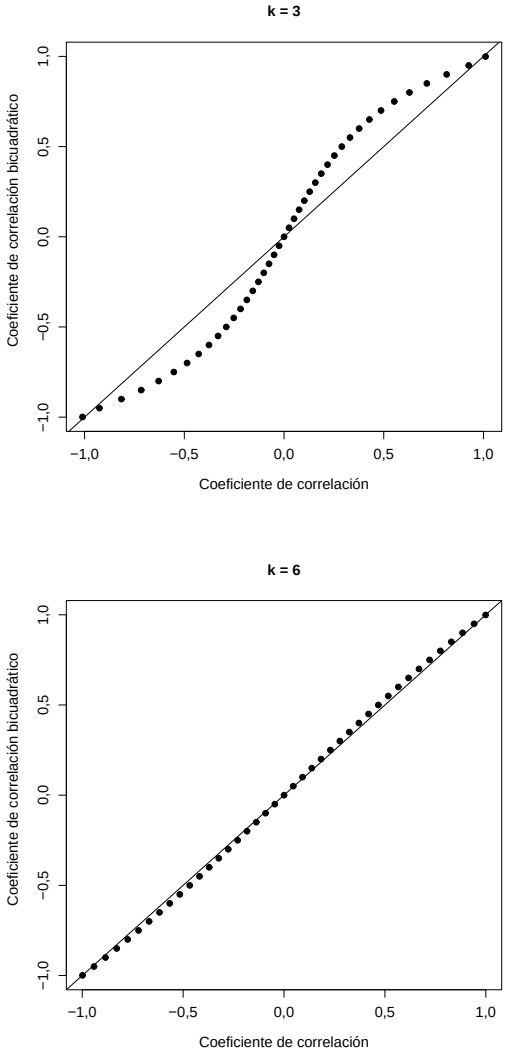

FIGURA 1: Valores de ϱ para $k = 3$ y $k = 6$.

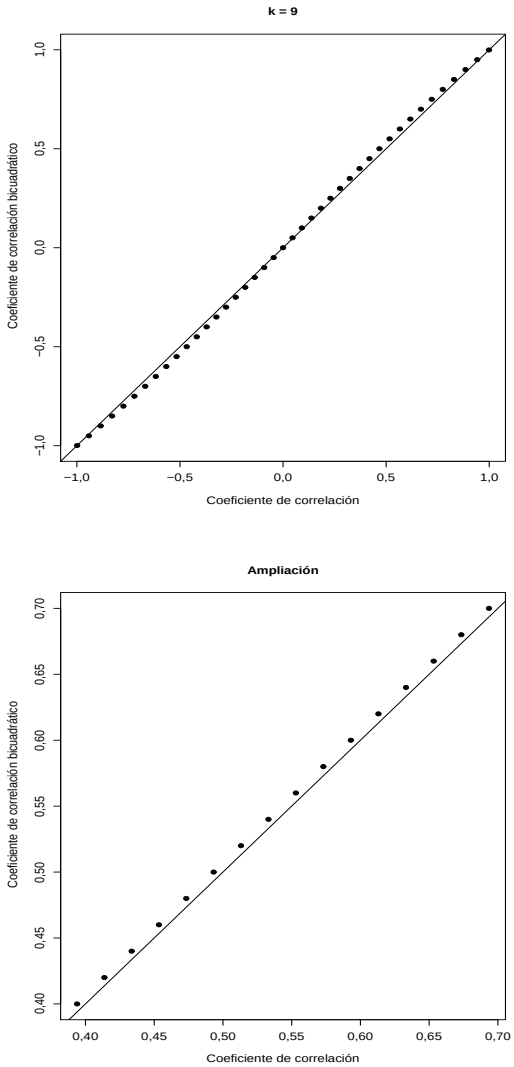

FIGURA 2: Valores de ϱ para $k = 9$.

Tabla 2: Valores de ϱ de acuerdo con los valores de ρ y k.

	Valor de k		
Valor de ρ	3	6	9
0,001	0,000497	0,000914	0,000982
0,100	0,049966	0,091461	0,098207
0,200	0,101565	0,183422	0,196523
0,300	0,156766	0,276384	0,295054
0,400	0,218223	0,370858	0,393911
0,500	0,289691	0,467366	0,493201
0,600	0,376515	0,566446	0,593034
0,700	0,486181	0,668655	0,693520
0,800	0,628749	0,774574	0,794769
0,900	0,815515	0,884808	0,896891
0,999	1,010647	0,998821	0,998964

3. Conclusiones

Asumiendo que el estimador del correlación bicuadrático presenta un comportamiento análogo al comportamiento del funcional aquí estudiado, los resultados sugieren que el estimador bicuadrático subestima el valor ρ cuando $\rho > 0$, y sobreestima su valor cuando $\rho < 0$.

[Recibido: marzo de 2010 — Aceptado: octubre de 2010]

Referencias

Beaton, A. & Tukey, J. (1974), 'The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data', *Technometrics* **16**(2), 147–185.

Bickel, P. & Docksum, K. (1977), *Mathematical Statistics, Basic Ideas and Selected Topics*, Holden-day Inc., San Francisco.

Cohen, C. (1955), 'Restriction and Selection in Samples from Bivariate Normal Distributions', *Journal of the American Statistical Association* **50**(271), 884–893.

Finney, D. (1962), 'Cumulants of Truncated Multi-Normal Distributions', *Journal of the Royal Statistical Society, serie B* **24**(2), 535–536.

Khatri, C. & Jaiswal, M. (1963), 'Estimation of Parameters of a Truncated Bivariate Normal Distribution', *Journal of the American Statistical Association* **58**(302), 519–526.

Lax, D. (1975), An Interim Report of a Monte Carlo Study of Robust Estimators of Widthers, Technical report, Department of Statistics, Princeton University.

Rosenbaum, S. (1961), 'Moments of a Truncated Bivariate Normal Distribution', *Journal of the Royal Statistical Society* **23**(2), 405–408.

Singh, N. (1960), 'Estimation of Parameters of a Multivariate Normal Population from Truncated and Censured Samples', *Journal of the Royal Statistical Society, serie B* **22**(2), 307–311.

Tallis, G. (1961), 'The Moment Generating Function of the Truncated Multinormal Distribution', *Journal of the Royal Statistical Society, serie B* **23**(1), 223–229.

Valcárcel, H. (2007), Propuesta de una función de autocorrelación con base en la función bicuadrática, Trabajo de grado, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá.

Wei, W. (2006), *Time Series Analysis: Univariate and Multivariate Methods*, second edn, Addison Wesley, Boston.