TESIS DOCTORAL

UNIVERSIDAD DE SALAMANCA

FACULTAD DE ECONOMÍA Y EMPRESA

DEPARTAMENTO DE ECONOMÍA E HISTORIA ECONÓMICA

ENFOQUE SEMI-NOPARAMÉTRICO PARA LA MEDICIÓN DE VARIABLES POSITIVAS DE COLAS PESADAS EN LOS CAMPOS DE LA ECONOMÍA Y LAS FINANZAS

AUTORA

LINA MARCELA CORTÉS DURÁN

DIRECTORES DE TESIS

JAVIER PEROTE PEÑA

ANDRÉS MORA VALENCIA

Salamanca

2017

Esta tesis está dedicada:

A mi Padre Celestial pues es el soporte y guía en mi caminar.

Al amor de mi vida, mi esposo Eugenio por su sacrificio, esfuerzo, por todo el apoyo brindado y acompañamiento incondicional. Por sus palabras de aliento y ánimo en aquellos momentos de nostalgia y debilidad.

A mis hijos Isabel y Samuel por su comprensión y por perdonar con amor mis ausencias y mis silencios. Los amo con todo mi ser y son mi principal motivación día a día.

A mi amada madre Cielo y mi amado padre Oscar pues son la piedra angular del ser humano y de la profesional que soy hoy en día. Gracias por creer siempre en mí.

A mi adorada hermana Verónica por su apoyo, por ser mi mejor amiga y compañera de vida.

A mis amigos quienes me brindaron su consejo en los momentos difíciles, en especial a Diana, Catalina y a John.

A Santi por estar ahí siempre de manera incondicional y por ser mi hermano, no de sangre, pero si del alma.

AGRADECIMIENTOS

Quiero agradecer a mis directores, los Dres. Andrés Mora Valencia y Javier Perote Peña por haber aceptado dirigir mi tesis. Les agradezco por brindarme su conocimiento y experiencia en la modelización de distribuciones semi-noparamétricas, por la confianza depositada en mí y por su apoyo durante el desarrollo de esta investigación. Al Dr. Javier le quiero agradecer de manera especial, por su disposición incondicional para orientarme. También, por su calidad humana pues es el mayor y más importante aprendizaje que me dejó este proceso. Le agradezco además porque fue una experiencia única y maravillosa el trabajar a su lado.

Agradezco de todo corazón a la Universidad EAFIT porque sin su ayuda no hubiese sido posible realizar este doctorado. Igualmente agradezco al Centro de Investigaciones Económicas y Financieras (CIEF) de la Escuela de Economía y Finanzas por permitirme realizar mi estancia de investigación. Agradezco al profesor Dr. Diego Agudelo por haberme encaminado por el apasionante mundo de la investigación, su acompañamiento y su consejo fueron determinantes para empezar este proyecto de vida académica.

Agradezco a toda mi familia por su apoyo, comprensión, paciencia y acompañamiento.

Por último, a todos aquellos que, de alguna u otra manera, me apoyaron o contribuyeron en el desarrollo de esta tesis.

A todos, ¡mil gracias!

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE APÉNDICES

CAPÍTULO I. Introducción al estudio

I.1. Introducción

En los campos del conocimiento, los seres humanos se han interesado por analizar el comportamiento de diferentes fenómenos con el fin de comprenderlos y anticiparse al futuro (Dekking, Kraaikamp, Lopuhaä, & Meester, 2007, p. 1). En ese sentido, una idea importante es que la ciencia es un medio por el cual se logra el aprendizaje, con una interacción entre la teoría y la práctica (Allen & Lunneborg, 1996). Sin embargo, en ocasiones para alcanzar esta interacción se hacen suposiciones tentativas acerca del mundo real, que se saben son falsas, pero que pueden ser de utilidad. Por ejemplo, el estadístico sabe que en la naturaleza nunca hubo una distribución normal, nunca hubo una línea recta, pero con supuestos normales y lineales a menudo se puede derivar resultados que coinciden, a una aproximación útil, con las que se encuentran en el mundo real (Box, 1976).

Dos campos de gran interés para el análisis de fenómenos son los de la economía y las finanzas. Por ejemplo, gran parte de la modelación financiera supone que los precios siguen un paseo aleatorio lognormal geométrico (o su análogo de tiempo continuo, el movimiento Browniano geométrico) (Merton, 1971; Cox & Ross, 1976; Stein & Stein, 1991). Así que una pregunta natural es si este supuesto es generalmente cierto. Este tipo de modelos hace dos supuestos: primero, los log-retornos se distribuyen normalmente y segundo, los log-retornos son independientes entre sí lo cual ha sido rechazado a partir del estudio empírico de series de retornos de activos financieros (Ruppert, 2010, p. 9; Christoffersen, 2012, p. 9).

Por lo tanto un problema fundamental dentro de los campos de la economía y las finanzas es la estimación de la función de densidad de una variable o vector aleatorio a partir de la información proporcionada por una muestra (Mandelbrot, 2003, p. 6). Un posible enfoque consiste en considerar que la función de densidad que se desea estimar pertenece a una determinada clase de funciones paramétricas, por ejemplo a algunas de las clásicas distribuciones: normal, exponencial, Poisson, etc. (Jondeau, Poon, & Rockinger, 2007, p. 383). Dicha suposición usualmente se basa en informaciones sobre la variable que son

externas a la muestra y cuya validez debe ser comprobada con posterioridad mediante pruebas de bondad de ajuste.

De esa manera, algunas de las variables que se estudian en economía y finanzas se pueden ajustar a partir de estimaciones paramétricas de la densidad. El problema radica en que en algunos fenómenos presentes en esos campos del conocimiento, las colas que exhiben los datos pueden ser muy pesadas (Hsieh, 1988; Loretan & Phillips, 1994; Plerou, Gopikrishnan, Nunes Amaral, Gabaix, & Eugene Stanley, 2000; Cont, 2001) y las funciones de densidad paramétricas pueden no caracterizar de manera adecuada el comportamiento de las variables. Al respecto, estudios realizados por Jurczenko, Maillet, & Negrea (2004), Jondeau & Rockinger (2006), Boudt, Peterson, & Croux (2008) y Gabaix, Lasry, Lions, & Moll (2016) proporcionan evidencia sobre la importancia que tiene representar correctamente las variables con el fin de no sobreestimar o subestimar el valor observado en las colas.

En la modelación de fenómenos económicos y financieros, otra posibilidad alternativa es no predeterminar a priori ninguna función de densidad de probabilidad de la variable. Este enfoque se denomina estimación no paramétrica de la densidad (Jondeau, Poon, & Rockinger, 2007, p. 383; Zhao, 2008) y permite que los datos determinen la forma de la distribución. Una posibilidad intermedia es el uso de aproximaciones semi-noparamétricas (SNP) en el que la forma funcional se parametriza solo en parte y siendo el resto una función desconocida (Ahn & Powell, 1993; Chen, 2007, p. 5552; Jondeau, Poon, & Rockinger, 2007, p. 383).

En este trabajo se considera un enfoque SNP en donde la función desconocida se modeliza a partir de una expansión de series de polinomios ortogonales. Específicamente se analizan expansiones de Gram-Charlier, las cuales se han aplicado en campos muy diversos en los que la precisión en la medición de las colas de la distribución es importante para la correcta medición de la ocurrencia de los valores extremos. Investigaciones en los campos de la economía y las finanzas que han incorporado este tipo de expansiones son las de Mauleón (1997), Mauleon & Perote (2000), Verhoeven & McAleer (2004), Ñíguez & Perote (2016) y Del Brio, Mora-Valencia, & Perote (2017) quienes analizaron el comportamiento de series financieras de alta frecuencia y modelizaron momentos de orden superior. Por su parte,

Corrado & Su (1996) y Rubinstein (1998) utilizaron estas densidades en estudios de opciones financieras. Asimismo, Perote & Del Brio (2003), Christoffersen & Gonçalves (2005) y Del Brio, Mora-Valencia, & Perote (2014a, b) se basaron en este tipo de densidades para mejorar medias del Valor en Riesgo (VaR). Finalmente, Ñíguez, Paya, Peel, & Perote (2012) quienes mostraron que bajo distribuciones SNP, los modelos de equilibrio general son estables, ya que la existencia de utilidad esperada está garantizada.

En particular, para la presente tesis se propone transformaciones logarítmicas de una distribución SNP (log-SNP) que son extensiones de una distribución lognormal que permiten aproximar cualquier distribución empírica mediante la introducción de parámetros adicionales.[1] Esta transformación busca mantener la estructura general y flexible de los parámetros de la "bien conocida" expansión de la serie Gram-Charlier (Sargan, 1975), distribución referida como SNP, pero restringiendo el dominio a valores positivos. El estudio se enfoca en la modelización de fenómenos económicos y financieros, cuyas series exhiben colas superiores pesadas. A lo largo de este estudio, se muestra que, ante datos muy sesgados y con posibles saltos en la cola debido a observaciones positivas extremas, la log-SNP es un proceso de generación de datos más fiable que la lognormal (que está anidado en el modelo).

I.2. Objetivos del estudio

Objetivo general:

Modelizar a partir de trasformaciones logarítmicas de una distribución semi-noparamétrica (log-SNP), fenómenos económicos y financieros, cuyas series exhiben colas superiores pesadas.

[1] En realidad, y hasta donde se sabe en el presente estudio, fue propuesta en primer lugar por Ñíguez, Paya, Peel, & Perote (2012), pero en otro contexto al presentado a lo largo de esta tesis.

Objetivos específicos:

- Medir la productividad investigadora y determinar los cuantiles que delimitan a los investigadores más productivos en cada área del conocimiento mediante la distribución univariante log-SNP.
- Modelizar la distribución del tamaño de la empresa y ajustar los cuantiles superiores de la distribución mediante la distribución univariante log-SNP.
- Desarrollar una expresión para la densidad de la distribución log-SNP multivariante con densidades marginales que se comportan como distribuciones univariantes de la log-SNP.
- Modelizar la función de densidad neutral de riesgo log-SNP usando los precios de opciones sobre petróleo crudo West Texas Intermediate (WTI).

I.3. Estructura del documento

El presente documento está escrito como capítulos independientes y autocontenidos que se dividen como sigue: en el capítulo II se realiza un estudio económico al analizar la productividad en la actividad investigadora partir del método bibliométrico. Este método consiste, principalmente, en cuantificar el número de documentos publicados por un país, institución, grupo de investigación o individuo, así como las citas recibidas por dichos documentos (Broadus, 1987; Borokhovich, Bricker, Brunarski, & Simkins, 1995; Abramo, D'Angelo, & Pugini, 2008; Heberger, Christie, & Alkin, 2010; Finardi, 2013; Bertocchi, Gambardella, Jappelli, Nappi, & Peracchi, 2015). Para ese fin, se utilizan los datos de O'Boyle & Aguinis (2012), quienes cuantifican el desempeño investigador de 140,971 investigadores que han producido 253,634 publicaciones en 18 campos del conocimiento en el periodo entre 2000 y 2009. También, se consideran publicaciones en el campo de las finanzas de 330 instituciones académicas presentadas por Borokhovich, Bricker, Brunarski, & Simkins (1995) para el periodo 1989 y 1993.

En dicho capítulo, se plantea por primera vez el uso de la distribución log-SNP para medir la productividad investigadora. El objetivo es determinar con mayor precisión los cuantiles que delimitan a los investigadores más productivos en cada área del conocimiento, como

medición de la dificultad que conlleva ser un investigador top en cada ámbito. La distribución de la productividad investigadora exhibe colas pesadas dado que habitualmente son unos pocos investigadores los que acumulan la mayor parte de las publicaciones top y sus correspondientes citas (Guerrero-Bote, Zapico-Alonso, Espinosa-Calvo, Gomez-Crisostomo, & Moya-Anegon, 2007; Lancho-Barrantes, Guerrero-Bote, & Moya-Anegón, 2010; Sabharwal, 2013). Los resultados muestran que las mediciones de esta productividad son muy sensibles al campo analizado y a la distribución empleada. En particular, distribuciones como la lognormal parecen infravalorar sistemáticamente la productividad de los investigadores top.

Por su parte, el capítulo III tiene como objetivo modelizar la distribución del tamaño de la empresa. Ese estudio es pertinente bajo la luz de la presente tesis, ya que existe evidencia de que las distribuciones empíricas del tamaño de la empresa en la cola superior, típicamente exhiben una alta asimetría y son series leptocúrticas. Ese fenómeno es el resultado de que un pequeño número de grandes empresas conviven con un gran número de empresas más pequeñas (Stanley, et al., 1995; Axtell, 2001; Bottazzi, Pirino, & Tamagni, 2015). De esta manera, se hace necesario proponer modelos que permitan capturar toda la forma de la distribución del tamaño de la empresa, incluyendo la cola derecha de la distribución (Crosato & Ganugi, 2007).

Para llevar a cabo el capítulo, se propone por primera vez el uso de la distribución log-SNP para modelizar la distribución del tamaño de la empresa a partir de muestra de empresas Estadounidenses en el periodo de 2004 a 2015. Adicionalmente, se toman diferentes niveles de agregación por actividad económica y se contrasta el desempeño de las distribuciones lognormal y log-SNP ajustando cada grupo de empresas. El estudio muestra que la log-SNP provee un mejor ajuste de la distribución del tamaño de la empresa. Adicionalmente, se desarrolla por primera vez una expresión para la densidad de la distribución log-SNP multivariante, esto permite obtener estimaciones más eficientes y analizar conjuntamente el comportamiento de variables altamente correlacionadas.

El capítulo IV contribuye a la literatura sobre la estimación de la función de densidad neutral de riesgo (RND) a partir de la modelización de los precios de opciones sobre petróleo crudo

West Texas Intermediate transadas en el periodo de enero de 2016 a enero de 2017. Ese capítulo busca extraer la RND implícita en los precios de opciones aplicando el modelo tradicional de Black & Scholes (1973) y el modelo SNP propuesto por Backus, Foresi, Li, & Wu (1997). Usar un modelo SNP en el caso de las opciones financieras es pertinente, ya que la evidencia empírica ha mostrado que las colas de la distribución de los rendimientos de los activos resultan ser más pesadas que las de una distribución normal (Fama, 1965; Das & Sundaram, 1999; Dennis & Mayhew, 2002; Nikkinen, 2003; Huang, Yu, Fabozzi, & Fukushima, 2009; Feng & Dang, 2016).

Al respecto, los resultados obtenidos en el capítulo muestran que al comparar el precio promedio observado en el mercado respecto al precio promedio teórico, la especificación lognormal tienden a infravalor sistemáticamente la valoración. En general la modelización permite concluir que incorporar los términos de ajuste de asimetría y de exceso de curtosis producen una precisión significativamente mejorada para obtener los precios de opciones sobre el WTI y ajustarlos a una RND log-SNP.

En la última parte del documento, se presentan las conclusiones obtenidas a través de los estudios llevados a cabo a lo largo de la tesis. Así mismo se hacen algunas recomendaciones que serían útiles para investigaciones futuras teniendo en cuenta el enfoque log-SNP aquí desarrollado.

Referencias

Abramo, G., D'Angelo, A. C., & Pugini, F. (2008). The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology. *Scientometrics, 76(2),* 225–244.

Ahn, H., & Powell, J. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal Of Econometrics, 58(1-2),* 3-29.

Allen, D., & Lunneborg, C. (1996). Modeling experimental and observational data. *Technometrics, 38(3),* 291-291.

Axtell, R. (2001). Zipf distribution of U.S. Firm sizes. *Science, 293*, 1818–1820.

Backus, D., Foresi, S., Li, K., & Wu, L. (1997). Accounting for Biases in Black-Scholes. *Manuscript, The Stern School at New York University*, 1-46.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy, 44(2),* 451-466.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy, 81(3),* 637–659.

Borokhovich, K. A., Bricker, R. J., Brunarski, K. R., & Simkins, B. J. (1995). Finance research productivity and influence. *The Journal of Finance, 50(5),* 1691-1717.

Bottazzi, G., Pirino, D., & Tamagni, F. (2015). Zipf law and the firm size distribution: a critical discussion of popular estimators. *Journal of Evolutionary Economics, 25*(3), 585–610.

Boudt, K., Peterson, B., & Croux, C. (2008). Estimation and decomposition of downside risk for portfolios with non-normal returns. *The Journal Of Risk, 11(2),* 79-103.

Box, G. (1976). Science and statistics. *Journal Of The American Statistical Association, 71(356),* 791-799.

Broadus, R. N. (1987). Toward a definition of 'bibliometrics'. *Scientometrics, 12(5-6),* 373-379.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. Heckman, & E. Leamer, *Handbook of Econometrics, Vol. 6. Part B* (pp. 5549-5632). Amsterdam: Elsevier.

Christoffersen, P. (2012). *Elements of financial risk management (2nd ed.).* Waltham: Academic Press.

Christoffersen, P., & Gonçalves, S. (2005). Estimation risk in financial risk management. *The Journal Of Risk, 7(3)*, 1-28.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance, 1(2)*, 223-236.

Corrado, C. J., & Su, T. (1996). Skewness and kurtosis in S&P 500 index returns implied by option prices. *The Journal of Financial Research, 19(2)*, 175-192.

Cox, J., & Ross, S. (1976). The valuation of options for alternative stochastic processes.,. *Journal Of Financial Economics, 3(1-2)*, 145-166.

Crosato, L., & Ganugi, P. (2007). Statistical regularity of firm size distribution: the Pareto IV and truncated Yule for Italian SCI manufacturing. *Statistical Methods and Applications, 16*(1), 85-115.

Das, S. R., & Sundaram, R. K. (1999). Of smiles and smirks: a term structure perspective. *The Journal of Financial and Quantitative Analysis, 34(2)*, 211-239.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2007). *A modern introduction to probability and statistics: understanding why and how*. London: Springer.

Del Brio, E., Mora-Valencia, A., & Perote, J. (2014a). Semi-nonparametric VaR forecasts for hedge funds during the recent crisis. *Physica A: Statistical Mechanics And Its Applications, 401*, 330-343.

Del Brio, E., Mora-Valencia, A., & Perote, J. (2014b). VaR performance during the subprime and sovereign debt crises: An application to emerging markets. *Emerging Markets Review, 20*, 23-41.

Del Brio, E., Mora-Valencia, A., & Perote, J. (2017). The kidnapping of Europe: high-order moments' transmission between developed and emerging markets. *Emerging Markets Review*, In press.

Dennis, P., & Mayhew, S. (2002). Risk-neutral skewness: evidence from stock options. *Journal of Financial and Quantitative Analysis, 37(3)*, 471–493.

Fama, E. (1965). The behaviour of stock market prices. *Journal of Business, 38(1)*, 34–105.

Feng, P., & Dang, C. (2016). Shape constrained risk-neutral density estimation by support vector regression. *Information Sciences, 333(C)*, 1-9.

Finardi, U. (2013). Correlation between Journal Impact Factor and Citation Performance: An experimental study. *Journal of Informetrics, 7(2)*, 357–370.

Gabaix, X., Lasry, J. M., Lions, P. L., & Moll, B. (2016). The dynamics of inequality. *Econometrica, 84(6)*, 2071–2111.

Guerrero-Bote, V. P., Zapico-Alonso, F., Espinosa-Calvo, M. E., Gomez-Crisostomo, R., & Moya-Anegon, F. (2007). Import–export of knowledge between scientific subject categories: The iceberg hypothesis. *Scientometrics, 71(3)*, 423–441.

Heberger, A. E., Christie, C. A., & Alkin, M. C. (2010). A bibliometric analysis of the academic influences of and on evaluation theorists' published works. *American Journal of Evaluation, 31(1)*, 24-44.

Hsieh, D. (1988). The statistical properties of daily foreign exchange rates: 1974–1983. *Journal Of International Economics, 24(1-2)*, 129-145.

Huang, D., Yu, B., Fabozzi, F., & Fukushima, M. (2009). CAViaR-based forecast for oil price risk. *Energy Economics, 31(4)*, 511-518.

Jondeau, E., & Rockinger, M. (2006). Optimal portfolio allocation under higher moments. *European Financial Management, 12(1)*, 29-55.

Jondeau, E., Poon, S.-H., & Rockinger, M. (2007). *Financial Modeling Under Non-Gaussian Distributions.* London: Springer.

Jurczenko, E., Maillet, B., & Negrea, B. (2004). A note on skewness and kurtosis adjusted option pricing models under the Martingale restriction. *Quantitative Finance, 4(5)*, 479-488.

Lancho-Barrantes, B. S., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). The iceberg hypothesis revisited. *Scientometrics, 85(2)*, 443–461.

Loretan, M., & Phillips, P. (1994). Testing the covariance stationarity of heavy-tailed time series: an overview of the theory with applications to several financial datasets. *Journal Of Empirical Finance, 1(2)*, 211-248.

Mandelbrot, B. (2003). Heavy tails in finance for independent or multifractal price increments. In S. Rachev, *Handbook of heavy tailed distributions in finance (1st ed.)* (pp. 4-32). Amsterdam: Elsevier.

Mauleón, I. (1997). Instability and long memory in conditional variances. *Journal de la Societe de Statistique de Paris, 134(4)*, 67-88.

Mauleon, I., & Perote, J. (2000). Testing densities with financial data: an empirical comparison of the Edgeworth-Sargan density to the Student's t. *The European Journal Of Finance, 6(2)*, 225-239.

Merton, R. (1971). Optimum consumption and portfolio rules in a continuous-time model. *Journal Of Economic Theory, 3(4)*, 373-413.

Nikkinen, J. (2003). Normality tests of option-implied risk-neutral densities: evidence from the small Finnish market. *International Review of Financial Analysis, 12(2)*, 99–116.

Ñíguez, T-M., & Perote, J. (2016). Multivariate moments expansion density: Application of the dynamic equicorrelation model. *Journal Of Banking & Finance, 72(S)*, S216-S232.

Ñíguez, T-M., Paya, I., Peel, D., & Perote, J. (2012). On the stability of the constant relative risk aversion (CRRA) utility under high degrees of uncertainty. *Economics Letters, 115(2)*, 244-248.

O'Boyle, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology, 65(1)*, 79–119.

Perote, J., & Del Brio, E. (2003). Measuring Value-at-Risk under the conditional Edgeworth-Sargan distribution. *Finance Letters, 1(3)*, 23-40.

Plerou, V., Gopikrishnan, P., Nunes Amaral, L., Gabaix, X., & Eugene Stanley, H. (2000). Economic fluctuations and anomalous diffusion. *Physical Review E, 62(3)*, R3023-R3026.

Rubinstein, M. (1998). Edgeworth binomial trees. *Journal of Derivatives, 5(3)*, 20-27.

Ruppert, D. (2010). *Statistics and data analysis for financial engineering.* New York: Springer-Verlag Inc.

Sabharwal, M. (2013). Comparing research productivity across disciplines and career stages. *Journal of Comparative Policy Analysis: Research and Practice, 15(2)*, 141-163.

Sargan, J. (1975). Gram-Charlier approximations applied to t ratios of k-class estimators. *Econometrica, 43(2)*, 327-347.

Stanley, M. H., Buldyrev, S. V., Havlin, S., Mantegna, R. N., Salinger, M. A., & Stanley, H. E. (1995). Zipf plots and the size distribution of firm. *Economics Letters, 49*(4), 453-457.

Stein, E., & Stein, J. (1991). Stock price distributions with stochastic volatility: an analytic approach. *Review Of Financial Studies, 4(4)*, 727-752.

Verhoeven, P., & McAleer, M. (2004). Fat tails and asymmetry in financial volatility models. *Mathematics And Computers In Simulation, 64(3-4)*, 351-361.

Zhao, Z. (2008). Parametric and nonparametric models and methods in financial econometrics. *Statistics Surveys, 2(0)*, 1-42.

CAPÍTULO II. Productividad investigadora: un enfoque semi-noparamétrico[†]

II.1. Introducción

En los últimos años la valoración de la productividad de los investigadores académicos en los diferentes campos del conocimiento, ha estado relacionada con la medición del impacto de los resultados de la producción científica (Abramo, D'Angelo, & Pugini, 2008; Sabharwal, 2013; Campanario, 2015). La motivación de estudiar la productividad radica en el deseo de promover la excelencia académica, y hacer de la investigación dentro de cada país más competitiva a nivel mundial (Frandsen, 2005; Kocher, Luptacik, & Sutter, 2006; Abramo & D'Angelo, 2014).

La calidad de un estudio de investigación se determina por una gran cantidad de variables que van desde características personales del investigador hasta tendencias y políticas nacionales e internacionales (Genest, 1997; Dundar & Lewis, 1998; Williamson & Cable, 2003; Seggie & Griffith, 2009; Duch, et al., 2012; Kaur, Ferrara, Menczer, Flammini, & Radicchi, 2015). Sin embargo, los criterios para evaluar el desempeño de la investigación se combinan principalmente de dos maneras. En primer lugar, se asume como principal método de valoración la revisión por pares, pero ésta se encuentra sujeta a cierto nivel de subjetividad (Abramo, D'Angelo, & Pugini, 2008; Bornmann, 2011; Bertocchi, Gambardella, Jappelli, Nappi, & Peracchi, 2015; Day, 2015).

Alternativamente, otra manera de valorar la actividad científica es, en términos de productividad, a partir de análisis bibliométrico. Este método consiste, principalmente, en cuantificar el número de documentos publicados por un país, institución, grupo de investigación o individuo, así como las citas recibidas por dichos documentos (Broadus, 1987; Borokhovich, Bricker, Brunarski, & Simkins, 1995; Abramo, D'Angelo, & Pugini, 2008; Heberger, Christie, & Alkin, 2010; Finardi, 2013). Las medidas bibliométricas más

[†] Una versión de este capítulo se encuentra publicada bajo el título "The productivity of top researchers: a semi-nonparametric approach" en el journal *Scientometrics*, JCR 2015 (2.084-Q1).

comunes se basan en las publicaciones y en las citas y esta información proviene de diferentes bases de datos como Web of Science (WoS), Scopus, Google Scholar, entre otras. Sin embargo, la heterogeneidad en las políticas de publicación y citación entre los diferentes campos del conocimiento (Kaur, Radicchi, & Menczer, 2013; Ruiz-Castillo & Costas, 2014; Mingers & Leydesdorff, 2015) hace que la comparación directa en términos del número de artículos publicados y citas sea "injusta" (Crespo, Ortuño-Ortín, & Ruiz-Castillo, 2012) y plantean la necesidad de buscar métodos de comparación más apropiados.

Por otra parte, la mayoría de los estudios de productividad investigadora se centran en un solo campo del conocimiento. Por ejemplo, la literatura enfocada en el desempeño de la investigación en economía es abundante (Hodgson & Rothman, 1999; Coupé, 2003; Kocher, Luptacik, & Sutter, 2006; Ellison, 2013). En consecuencia, y teniendo en cuenta los avances científicos existentes en cada campo del conocimiento, se hace relevante estudiar la productividad investigadora no solo desde la medición de los resultados de la producción científica, sino también analizando diferencias entre los campos del conocimiento en los cuales se está investigando (Sabharwal, 2013; Abramo & D'Angelo, 2014; Ruiz-Castillo & Costas, 2014; Bertocchi, Gambardella, Jappelli, Nappi, & Peracchi, 2015).

Adicionalmente, los estudios sobre la productividad investigadora han tenido en cuenta distintas funciones de distribución probabilidad con el fin de identificar patrones, en la relación cuantitativa, entre los autores y sus contribuciones a lo largo de un periodo de tiempo. Estas investigaciones han identificado que indicadores bibliométricos tales como el número de artículos publicados o el número de citas recibidas por autor se caracterizan por exhibir distribuciones de colas pesadas (Lotka, 1926; Price, 1976; Chung & Cox, 1990; Redner, 1998; Albarrán, Juan, Ortuño, & Ruiz-Castillo, 2011; Eom & Fortunato, 2011; Da Silva, Kalil, De Oliveira, & Martinez, 2012; Ruiz-Castillo & Costas, 2014; Campanario, 2015).

Como resultado, los modelos de distribución de probabilidad que más se han aplicado en la literatura sobre productividad investigadora son los que atienden a las siguientes leyes: la ley de Lotka (Lotka, 1926; Nicholls T. P., 1986; Chung & Cox, 1990; Kretschmer & Kretschmer, 2007), ley potencial (Price, 1976; Egghe, 2005; Albarrán, Juan, Ortuño, & Ruiz-Castillo, 2011; Aguinis, O'Boyle, Gonzalez-Mulé, & Joo, 2015) y la ley de Bradford (Garfield, 1980;

Rousseau, 1994; Nicolaisen & Hjørland, 2007; Campanario, 2015). Estas leyes basadas principalmente en funciones de distribución como la exponencial o la de Pareto, han sido controvertidas y han generado un fuerte debate durante más de un siglo. Por ejemplo, Newman (2005) afirmó que pocos procesos del mundo real siguen una ley potencial en todo su rango, y en particular no para los valores más pequeños de la variable que se está midiendo. Martínez-Mekler, et al. (2009) argumentaron que, cuando se usan datos reales, las leyes de potencia sólo se mantienen para un rango intermedio de valores, mientras que las colas de las distribuciones tienden a desviarse de los valores esperados.

Otros estudios tales como los Kumar, Sharma, & Garg (1998), Radicchi, Fortunado, & Castellano (2008), Perc (2010), Eom & Fortunato (2011) y Birkmaier & Wohlrabe (2014) han propuesto la aplicación de la distribución lognormal para estudiar la actividad investigadora. Sin embargo, la evidencia sobre la verdadera distribución de la producción científica y la citación aún no es concluyente (Albarrán, Juan, Ortuño, & Ruiz-Castillo, 2011), lo cual podría ser una consecuencia del uso de distribuciones de uno o dos parámetros.

De hecho, todas las distribuciones propuestas tienen la desventaja de que dependen de muy pocos parámetros para capturar toda la forma de la distribución de la productividad, especialmente en la cola derecha de la distribución. Este hecho podría resultar en mediciones de productividad más imprecisas y las comparaciones de desempeño poco fiables entre diferentes campos del conocimiento. Con el objeto de obtener medidas de productividad investigadora fiables este estudio propone el uso de aproximaciones semi-noparamétricas (SNP) de la distribución de productividad basadas en expansiones de Edgeworth y Gram-Charlier. Estas distribuciones se han aplicado en campos muy diversos en los que la precisión en la medición de las colas de la distribución es importante para la correcta medición de la ocurrencia de los valores extremos (véase Blinnikov & Moessner 1998 o Mauleón & Perote 2000, como ejemplos de aplicaciones a astronomía o finanzas, respectivamente). En este capítulo se plantea por primera vez el uso de las mismas para medir la productividad investigadora y determinar con mayor precisión los cuantiles que delimitan a los investigadores más productivos en cada área del conocimiento como medición de la dificultad que conlleva ser un investigador top en cada ámbito.

Con el fin de mantener la flexibilidad de parámetros de las distribuciones de Gram-Charlier, pero restringiendo el dominio a valores positivos, se proponen transformaciones logarítmicas de una distribución SNP (referida como log-SNP) que son extensiones de una distribución lognormal que permiten aproximar cualquier distribución empírica mediante la introducción de parámetros adicionales. Dado que los indicadores bibliométricos, usualmente, exhiben colas relativamente largas y con multimodalidad (Guerrero-Bote, Zapico-Alonso, Espinosa-Calvo, Gomez-Crisostomo, & Moya-Anegon, 2007; Lancho-Barrantes, Guerrero-Bote, & Moya-Anegón, 2010; Sabharwal, 2013), el presente estudio muestra que, en contraste a la distribución lognormal, la distribución log-SNP provee un mejor ajuste a la hora de caracterizar el desempeño investigador.

Este capítulo se divide de la siguiente forma: la sección II.2 define la distribución que permite caracterizar la productividad investigadora. La sección II.3 presenta los datos a utilizar y la metodología de estudio. La sección II.4 recoge los resultados y las pruebas de robustez, y finalmente la sección II.5 expone las conclusiones del capítulo.

II.2. La distribución de productividad

La caracterización de una variable aleatoria mediante su función de densidad de probabilidad (pdf) y su ajuste a la distribución empírica de una serie se puede realizar mediante distintos enfoques que van desde una perspectiva paramétrica basada en una distribución de frecuencias con forma funcional conocida a un enfoque no paramétrico puro. Una posibilidad intermedia es el uso de aproximaciones SNP en el que la forma funcional se parametriza sólo en parte y siendo el resto una función desconocida (Chen, 2007, p. 5552). En este trabajo se considera un enfoque SNP en donde la función desconocida se modeliza a partir de una expansión de series de polinomios ortogonales. En particular, se analizan expansiones de Edgeworth y Gram-Charlier que se han mostrado como aproximaciones asintóticas válidas a cualquier distribución empírica bajo condiciones de regularidad relativamente débiles (Sargan, 1975; Phillips, 1977). A continuación se define la distribución SNP basada en series de Gram-Charlier y su transformación logarítmica, analizando sus propiedades básicas.

II.2.1. La distribución SNP

Sea $\{P_s(x)\}$, $x \in \mathbb{R}$ y $s \in \mathbb{N}$ una familia de polinomios ortogonales con respecto a una función de densidad $w(x)$ que satisface la siguiente relación[3]

$$\int_{-\infty}^{\infty} P_s(x)P_j(x)w(x)dx = 0, \quad \forall s \neq j,\ s,j = 0,1,2,\ldots \tag{II.1}$$

Dentro de esta familia los polinomios de Hermite (HP) son aquellos que utilizan como peso una función de densidad normal estándar, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$. En particular el polinomio de Hermite de orden *s*, $H_s(x)$, se puede obtener en términos de derivada de orden *s* de la función de densidad la distribución normal estándar tal como se expresa en la ecuación (II.2):

$$H_s(x) = \frac{(-1)^s}{\phi(x)} \frac{d^s\phi(x)}{dx^s}. \tag{II.2}$$

A continuación se muestran los ocho primeros HP:

$$H_0(x) = 1, \tag{II.3}$$
$$H_1(x) = x, \tag{II.4}$$
$$H_2(x) = x^2 - 1, \tag{II.5}$$
$$H_3(x) = x^3 - 3x, \tag{II.6}$$
$$H_4(x) = x^4 - 6x^2 + 3, \tag{II.7}$$
$$H_5(x) = x^5 - 10x^3 + 15x, \tag{II.8}$$
$$H_6(x) = x^6 - 15x^4 + 45x^2 - 15, \tag{II.9}$$
$$H_7(x) = x^7 - 21x^5 + 105x^3 - 105x, \tag{II.10}$$
$$H_8(x) = x^8 - 28x^6 + 210x^4 - 420x^2 + 105. \tag{II.11}$$

Es sencillo comprobar que estos polinomios satisfacen la menciona propiedad de ortogonalidad dado que $\forall s, j = 0, 1, 2, \ldots$

$$\int_{-\infty}^{\infty} H_s(x)H_j(x)\phi(x)dx = \begin{cases} 0, & s \neq j \\ s!, & s = j \end{cases}. \tag{II.12}$$

[3] Diferentes funciones de pesos se pueden usadar $w(x)$, para detalles ver Abramowitz & Stegun, 1972 (pp. 774-775). Por convenio se considera $P_0(x) = 1$.

Los HP constituyen además la base de las series de Edgeworth y Gram-Charlier (tipo A) que permiten, bajo ciertas condiciones de regularidad (Cramér, 1925), expresar cualquier pdf, $f(x)$, en términos de una una serie infinita (Wallace, 1958) de la siguiente forma[4]

$$f(x) = \sum_{s=0}^{\infty} \delta_s H_s(x)\phi(x), \text{ donde } \delta_s = \frac{1}{s!}\int_{-\infty}^{\infty} H_s(x)\phi(x)dx. \quad \text{(II.13)}$$

Incluso, gracias a la ortogonalidad de los HP, la truncación de las series en un determinado orden *n* de la expansión permite definir una familia de distribuciones SNP, $g(x;\boldsymbol{d})$, donde $\boldsymbol{d} = (d_1, \dots, d_n)' \in \mathbb{R}^n$ denota el vector de parámetros.[5]

$$g(x;\boldsymbol{d}) = [1 + \sum_{s=1}^{n} d_s H_s(x)]\phi(x) \underset{n\to\infty}{\longrightarrow} f\ (x) \quad \text{(II.14)}$$

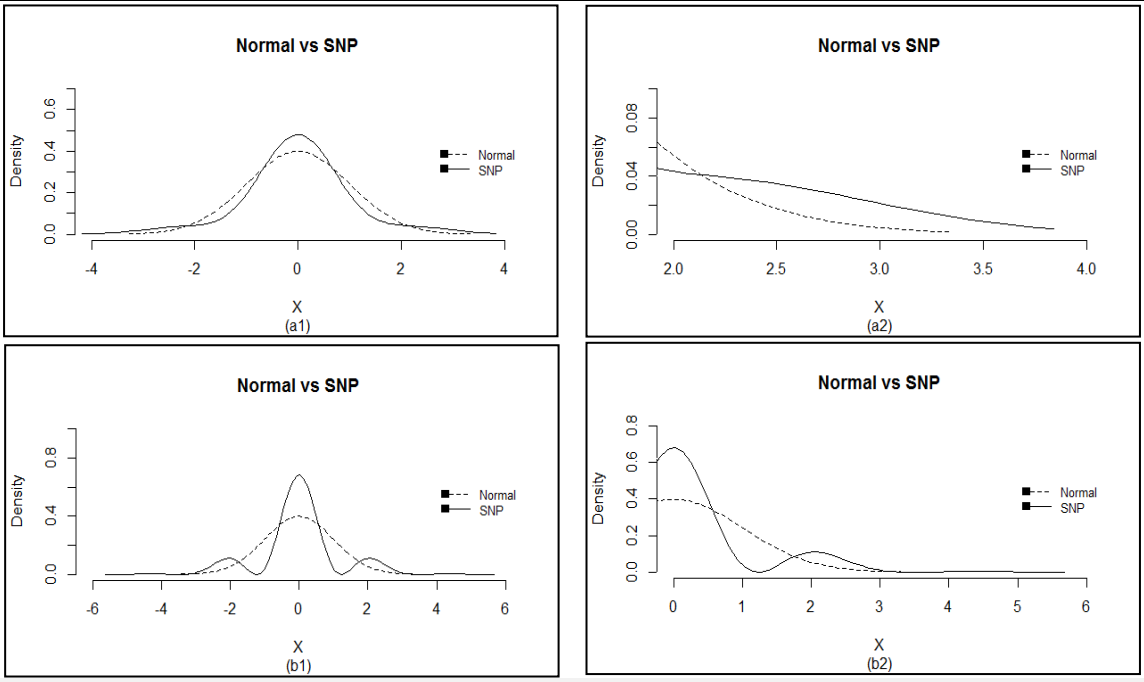
Sin embargo la distribución SNP definida en la ecuación (II.14) sólo es una función de densidad para el subconjunto de valores de $\boldsymbol{d}$ que garanticen $g(x;\boldsymbol{d}) \geq 0$. Para solucionar este problema se han propuesto distintos tipos de restricciones o transformaciones de positividad (Gallant & Nychka, 1987), si bien éstas suponen introducir una complejidad innecesaria para aplicaciones empíricas que implementan algoritmos de máxima verosimilitud (dado que en el óptimo éstos conducen a estimaciones que garantizan la positividad).

La gran ventaja de esta distribución SNP frente a otras especificaciones paramétricas radica precisamente en las mejoras en los ajustes a los que conduce, dada su gran flexibilidad paramétrica que permite ajustar la localización y la escala con diferentes parámetros que los utilizados para la asimetría, leptocurtosis e incluso momentos de mayor orden. La Figura II.1 ilustra la forma de la distribución SNP (representada con 1,000 observaciones simuladas) en comparación con una distribución normal. Por razones de comparación, en ambos casos se consideran los mismos parámetros de ubicación y escala, $\mu = 0$ y $\sigma = 1$, pero se introducen parámetros adicionales (pares) en la SNP. En particular, los Paneles (a1) y (a2) incorporan $d_2 = 0.1$ y $d_4 = 0.1$ y los Paneles (b1) y (b2) $d_2 = 0.1$, $d_4 = 0.1$, $d_6 = 0.001$ y $d_8 = 0.005$. Obsérvese también que los paneles (a1) y (b1) representan todo el dominio pero los Paneles (a2) y (b2) sólo un detalle de las colas de distribución derecha. Es evidente a partir

[4] Para más detalles sobre las series de Edgeworth y Gram-Charlier véase Kendall & Stuart (1977, pp. 167-172).
[5] Hay que anotar que dado un orden de truncación la distribución resultante es puramente paramétrica, pero con carácter general el orden de truncación puede variar para admitir una aproximación más precisa a una distribución dada. Sin pérdida de generalidad se asume $d_0 = 1$.

de estas imágenes que la SNP no sólo captura la leptocurtosis sino que también presenta colas ondulantes y pesadas que pueden adaptar el patrón de probabilidad de cualquier proceso de generación de datos.

Figura II.1 Pdf de la distribución Normal versus SNP



Las figuras comparan la forma de ambas distribuciones Normal (línea discontinua) y SNP (línea sólida) con parámetros de escala y localización, μ=0 y σ=1, y parámetros adicionales para la última. Particularmente, los Paneles (a1) y (a2) incorporan parámetros $d_2 = 0.1$ y $d_4 = 0.1$ y los Paneles (b1) y (b2) consideran $d_2 = 0.1$, $d_4 = 0.1$, $d_6 = 0.001$ y $d_8 = 0.005$. Los Paneles (a1) y (b1) representa el total del dominio y (a2) y (b2) un detalle de las colas derechas. Las figuras fueron simuladas a través de 1,000 repeticiones.

Además este mayor número de parámetros no conlleva una mayor dificultad ni en términos teóricos ni empíricos. Por ejemplo, los momentos centrales pueden obtenerse fácilmente como funciones lineales de los parámetros de la distribución (véase el Apéndice II.A). Obsérvese que el momento par/impar de orden *n* depende sólo de los *n* primeros parámetros pares/impares. Esto permite obtener fácilmente valores iniciales para los algoritmos de optimización mediante la aplicación del método de momentos (MM). También se puede obtener una fórmula cerrada para la función de distribución acumulada (cdf) de la distribución SNP en función de la cdf de la distribución normal, tal como se muestra en la

ecuación (II.15) (véase la prueba en el Apéndice II.B). Esto permite de forma sencilla el cálculo de las probabilidades y los cuantiles de la distribución SNP.

$$G_x(a) = \int_{-\infty}^{a} g(x;\boldsymbol{d})dx = \int_{-\infty}^{a} \phi(x)dx - \phi(a)\sum_{s=1}^{n} d_s H_{s-1}(a). \quad \text{(II.15)}$$

II.2.2. La distribución log-SNP

Ñíguez, Paya, Peel, & Perote (2012) definen una variable $z > 0$ como log-SNP (estándar) si la variable $x = \log(z)$ tienen como pdf la distribución SNP presentada en la ecuación (II.14). La distribución resultante hereda todas las buenas propiedades de la distribución SNP y, especialmente, su flexibilidad para recoger los valores extremos de la distribución, pero la densidad se define en $\mathbb{R}^+$, lo cual se requiere para ajustar los datos de productividad. En este estudio se da un paso más y se define una distribución log-SNP de forma similar, pero sobre una transformación lineal $y = \sigma x + \mu$.

Definición II.2.2.1: Se dice que la variable $z > 0$ se distribuye log-SNP con parámetros de localización $\mu \in \mathbb{R}$, escala $\sigma^2 \in \mathbb{R}$ y forma $\boldsymbol{d} = (d_1, \dots, d_n)' \in \mathbb{R}^n$ si su pdf puede expresarse como

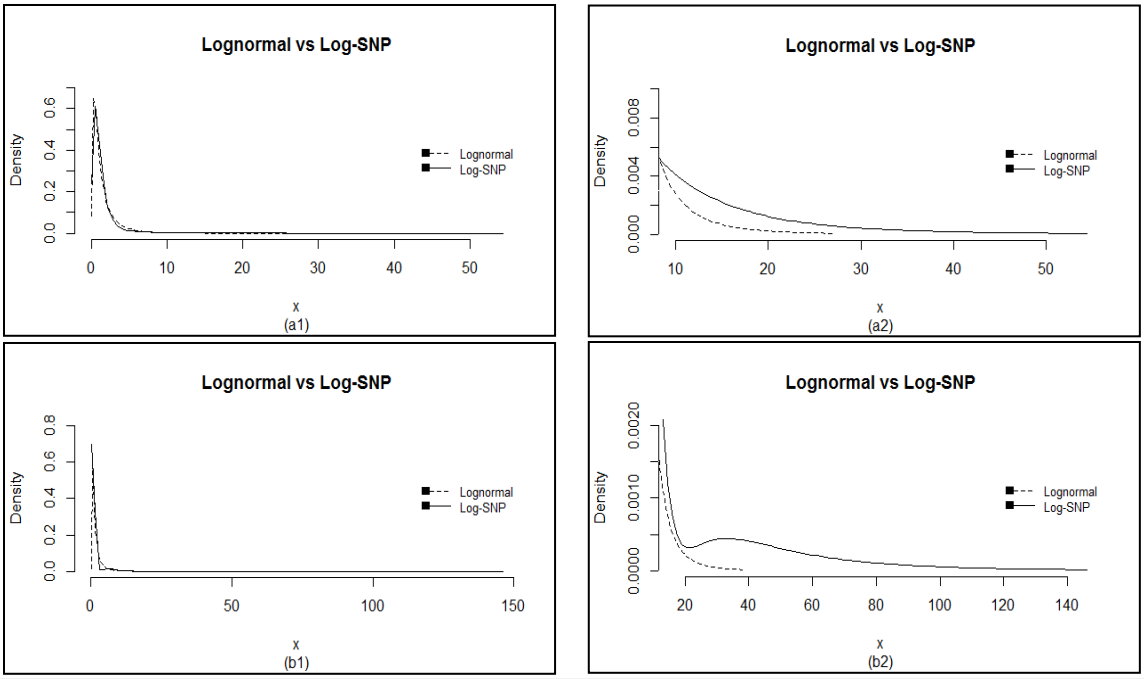
$$h(z;\ \mu, \sigma^2, \boldsymbol{d}) = \left[1 + \sum_{s=1}^{n} d_s H_s\left(\frac{\log(z)-\mu}{\sigma}\right)\right]\left(\frac{1}{z\sigma\sqrt{2\pi}} e^{\frac{-(\log(z)-\mu)^2}{2\sigma^2}}\right). \quad \text{(II.16)}$$

Definida de esta manera, la distribución lognormal es un caso particular de la log-SNP (para $d_s = 0,\ \forall s$) lo que permite contrastar la mejora en los ajustes de ésta última frente a la lognormal mediante un contraste de restricciones lineales como el de razón de verosimilitudes. El presente estudio muestra que efectivamente la flexibilidad paramétrica de la log-SNP permite mejoras de ajuste significativas para medir distribuciones de productividad al ser muy flexible para representar diferentes formas (incluyendo saltos en la masa probabilística y colas pesadas) a partir de la incorporación de parámetros adicionales a los de una distribución paramétrica tradicional como la lognormal. Estos parámetros están directamente relacionados con los momentos[6] de distribución y constituyen grados

[6] Los momentos de la distribución log-SNP se pueden derivar directamente como $E_h[z^t] = e^{\mu t + \frac{1}{2}t^2\sigma^2}[1 + \sum_{s=1}^{n} d_s(\sigma t)^s]$ (ver Ñíguez, Paya, Peel, & Perote 2013).

adicionales de libertad para los procedimientos de estimación. Por ejemplo, si sólo se consideran los parámetros d_s para *s* par, la asimetría depende sólo del parámetro σ, y cuanto mayor sea la expansión, la cola de distribución será más pesada (y posiblemente más ondulante).

Figura II.2 Pdf de la distribución Lognormal versus Log-SNP



Las Figuras comparan la forma de las distribuciones Lognormal (línea discontinua) y Log-SNP (línea sólida) con parámetros de localización y de escala, μ=0 y σ=1, y parámetros adicionales para la última. Particularmente, Los Paneles (a1) y (a2) incorporan parámetros $d_2 = 0.12$ y $d_4 = 0.11$ y los Paneles (b1) y (b2) consideran $d_2 = 0.28$, $d_4 = 0.44$, $d_6 = 0.07$ y $d_8 = 0.009$. Los Paneles (a1) y (b1) representa el total del dominio y (a2) y (b2) un detalle de las colas derechas. Las figuras fueron simuladas a través de 1,000 repeticiones.

La Figura II.2 presenta una ilustración (1,000 repeticiones simuladas) de la forma de la distribución log-SNP en comparación con la lognormal, ambas con los mismos parámetros de localización y de escala, es decir, $\mu = 0$ y $\sigma = 1$. Los Paneles (a1) y (a2) representan una log-SNP con parámetros adicionales $d_2 = 0.12$ y $d_4 = 0.11$ y los Paneles (b1) y (b2) incorporan parámetros $d_2 = 0.28$, $d_4 = 0.44$, $d_6 = 0.07$ y $d_8 = 0.009$. Con el fin de

enfatizar en el comportamiento de los valores extremos (positivos) los Paneles (a2) y (b2) muestran una ampliación de las colas derechas de la distribución.

De este ejemplo está claro que la log-SNP permite tener más flexibilidad para capturar las colas pesadas (y onduladas) y, lo que es aún más importante, si se trata de obtener la forma de la distribución así como la cola pesada con un solo parámetro (σ), posiblemente se van a obtener estimaciones sesgadas y resultados engañosos.

II.3. Datos y metodología

II.3.1. Datos

Para probar si una distribución lognormal o una log-SNP ajusta mejor a la distribución de desempeño de 140,971 investigadores que han producido 253,634 publicaciones en 18 campos del conocimiento, se utilizaron los datos de O'Boyle & Aguinis (2012).[7] Estos autores clasificaron los campos del conocimiento a partir del Journal Citation Reports (JCR), que proporciona los factores de impacto (IFs) en distintas áreas del conocimiento clasificadas dentro de las categorías de "ciencias" y "ciencias sociales". Como es bien sabido, hay varios subcampos incluidos dentro de una categoría JCR, pero ellos identificaron a autores en todos los subcampos tal que los autores que publican en más de un área tuvieran todas sus publicaciones incluidas.

Los autores usaron los factores de impacto del JCR en 2007 para identificar los cinco principales journals dentro de cada uno de los campos del conocimiento. Ellos eligieron revistas específicas de cada campo para evitar que la búsqueda fuera contaminada por autores de otras ciencias. Además los autores utilizaron el programa Publish o Perish (Harzing, 2008), que se basa en Google Scholar, para identificar a todos los autores que habían publicado al menos un artículo en una de estas revistas entre enero de 2000 y junio de 2009.[8] De esta manera, como medida de productividad investigadora, se tomó el número de artículos

[7] La autora agradece a Herman Aguinis y Ernest O'Boyle por permitir el uso de su base de datos sobre productividad académica recopilada en O'Boyle & Aguinis (2012).
[8] Para detalles sobre la forma de depuración de los datos véase O'Boyle & Aguinis (2012), p.86.

publicados por un autor en cada uno de los campos del conocimiento durante el período de observación de 9.5 años.

Una limitación del uso de esta medida de productividad es la autoría múltiple de los manuscritos, ya que ésta puede sesgar los resultados a favor de algunos autores y, comparativamente, en áreas donde un alto número de autores por artículo es comúnmente aceptado por la comunidad investigadora (Ruiz-Castillo & Costas, 2014; Mingers & Leydesdorff, 2015). La literatura sobre producción científica, sin embargo, no discrimina por autoría múltiple. Este procedimiento se conoce como "recuento completo", es decir, el hecho de que un artículo es igualmente valioso para todos sus autores sin importar el número de autores y su contribución marginal al manuscrito (Nicholls P. T., 1989; Ruiz-Castillo & Costas, 2014).

Por otra parte, otro aspecto importante al analizar la productividad de la investigación es la relación entre cantidad y calidad (Kaur, Ferrara, Menczer, Flammini, & Radicchi, 2015). En el presente estudio, así como en O'Boyle & Aguinis (2012), la calidad de las mejores investigaciones se impone por el hecho de que sólo se consideran publicaciones en los 5 principales journals en todos los campos del conocimiento. Entonces las comparaciones se hacen solamente en términos de cantidad de manuscritos para una probabilidad dada en la distribución "verdadera" (es decir, un cuantil), como proxies de nivel de dificultad teniendo en cuenta las diferentes prácticas entre las áreas.

La Tabla II.1 incorpora las estadísticas descriptivas para la muestra seleccionada en el presente estudio. En esta tabla también se registra el indicador Median Impact Factor (MIF) de los cinco principales journals en cada uno de los campos del conocimiento seleccionados, basado en el JCR del año 2007[9] para cada una de las categorías analizadas clasificadas en ciencias y ciencias sociales. Se proporciona este índice de citas para obtener una visión más amplia de cada uno de los campos seleccionados y, en particular, su correlación con la producción científica.

[9] En el presente estudio se toma el JCR del año 2007 para que sea consistente con O'Boyle & Aguinis (2012), ya que ese fue el año utilizado por los autores para la selección de los cinco principales journals dentro de cada campo del conocimiento

Tabla II.1 Estadística descriptiva de la producción científica

Campo del conocimiento	N	N (Posición ordinal)	Media (1)	Media (2)	% autores con publicaciones menores o iguales a la media (1)	% autores con publicaciones superiores a la media (2)	Std	Skew	K	Max	Max (posición ordinal)	Median Impact Factor	Median Impact Factor (posición ordinal)	Edición JCR
Agronomy	8,923	7	1.42	2.84	77.23	8.13	1.16	6.36	72.68	26	13	2.36	12	Science
Anthropology	5,755	9	1.87	3.55	66.06	10.50	1.95	4.49	34.52	30	8	2.31	14	Social Sciences
Clinical psychology	10,418	6	1.89	4.88	64.80	6.08	2.38	10.80	267.22	93	2	4.68	3	Social Sciences
Dentistry	12,345	3	2.26	5.92	79.18	6.98	2.98	6.54	74.62	66	4	3.37	6	Science
Dermatology	30,531	1	2.25	4.24	61.45	9.23	3.38	8.01	113.19	93	2	3.50	5	Science
Ecology	5,730	10	1.71	3.18	67.61	8.53	1.68	7.88	148.90	50	6	4.82	2	Science
Economics	3,048	13	1.62	3.20	71.75	6.63	1.67	7.14	82.10	27	11	3.69	4	Social Sciences
Educational psychology	3,032	14	1.70	3.03	65.60	7.26	1.55	5.41	52.04	27	11	2.35	13	Social Sciences
Ethics	1,073	18	1.65	3.21	70.74	6.43	1.78	6.82	71.24	26	13	1.31	16	Social Sciences
Ethnic studies	2,003	16	1.48	3.02	76.49	4.34	1.38	5.99	50.89	17	15	0.89	17	Social Sciences
Finance	3,019	15	2.14	5.40	77.97	6.69	2.52	4.69	33.93	28	9	2.99	8	Social Sciences
Forestry	12,211	4	1.82	3.27	63.80	9.77	1.80	5.66	68.58	46	7	2.14	15	Science
Genetics	16,574	2	1.71	3.12	83.49	3.56	2.18	26.42	1240.47	120	1	18.30	1	Science
History	6,708	8	1.54	2.56	65.28	12.00	0.97	3.33	25.11	14	17	0.85	18	Social Sciences
Law	1,350	17	1.55	2.93	71.48	11.48	1.24	3.88	24.07	13	18	3.09	7	Social Sciences
Linguistics	3,600	12	1.73	3.32	68.69	8.50	1.78	5.98	59.06	28	9	2.37	11	Social Sciences
Mathematics	3,972	11	1.45	2.61	72.18	8.06	1.02	4.86	41.42	15	16	2.56	10	Science
Statistics	10,679	5	2.08	5.52	81.04	5.48	2.52	6.22	67.39	54	5	2.97	9	Science

Esta tabla incorpora las estadísticas descriptivas de las publicaciones en las 5 mejores revistas para 18 campos del conocimiento pertenecientes a las categorías JCR de las ciencias y las ciencias sociales entre los años 2000 y 2009. N=número de investigadores por campo del conocimiento, Media (1)=media de publicaciones de investigadores en toda la muestra; Media (2)=media de las publicaciones de los investigadores con un número de artículos por encima de la Media (1), Std=desviación estandar, Skew=coeficiente de asimetría, K=coeficiente de exceso de curtosis, Max=máxima puntuación, MIF=Median Impact Factor (5 mejores revistas en 2007).

Como se observa, a través de los 18 campos del conocimiento analizados, el número de investigadores tiene un mínimo de 1,073 para el campo del conocimiento Ethics y un máximo de 30,531 para Dermatology. Para cada campo, se calculan dos valores medios sobre la producción científica: La Media (1) es la media de publicaciones para cada campo y para toda la muestra de investigadores; La Media (2) es la media de las publicaciones para cada campo, pero sólo para los investigadores con un número de artículos por encima de la Media (1). La Media (1) varía de 1.42 a 2.26 y la Media (2) de 2.56 a 5.92. Además, se calcula el porcentaje de autores con un número de publicaciones menores o iguales a la Media (1) y el porcentaje de autores con publicaciones superiores a la Media (2).

El la tabla se muestra que, en promedio, el 71,38% de todos los investigadores tienen una productividad por debajo de la Media (1), mientras que los investigadores con una productividad por encima de la Media (2) representan el 7,76%. Estos resultados respaldan los hallazgos de Ruiz-Castillo & Costas (2014) sobre la asimetría de las distribuciones de productividad por campo del conocimiento, ya que una gran proporción de investigadoress tienen una productividad media inferior y sólo un pequeño porcentaje de ellos es responsable de la mayoría de las publicaciones.

Con respecto a las otras estadísticas de la Tabla II.1, la desviación tiene un rango entre 0.97 y 3.38 publicaciones. Al analizar la asimetría y el exceso de curtosis de la distribución de productividad, se observa que todos los campos presentan asimetría positiva y leptocurtosis, siendo el campo del conocimiento Genetics el más asimétrico y leptocúrtico de la muestra. Por su parte, el número máximo de artículos por investigador varía entre 13 (Law) y 120 (Genetics) en función del campo considerado. Por otra parte, también se encuentran grandes diferencias en el indicador MIF que varía de de 0.85 (History) hasta 18.30 (Genetics). Además, el MIF se encuentra relacionado con el número máximo de artículos por investigador. Así, Genetics tiene la primera posición ordinal en el MIF y en el número máximo de publicaciones por investigador, mientras que History ocupa la posición ordinal 18 en el indicador y la 17 en las publicaciones.

En conjunto, los resultados confirman la existencia de amplias diferencias en la producción científica en términos de número de artículos por autor entre los diferentes campos del conocimiento, lo cual es consistente con otros estudios, por ejemplo, Abramo & D'Angelo

(2014) o Mingers & Leydesdorff (2015). A continuación se propone una nueva metodología basada en la distribución log-SNP para estudiar cómo estas diferencias afectan la distribución de la productividad, especialmente cuando se mide la productividad de los mejores investigadores.

II.3.2. Metodología

Esta sección presenta la metodología aplicada para caracterizar la productividad investigadora en cada área del conocimiento a partir de la distribución log-SNP. Se ofrecen detalles sobre la estimación máximo verosímil (ML) de los parámetros de su pdf y la valoración de su bondad de ajuste respecto a la distribución lognormal. La pdf de la distribución log-SNP se estima secuencialmente hasta un orden de truncación de $n = 8$.

Sea z_i el número de artículos publicados por un autor en uno de los campos del conocimiento seleccionados, la función de log-verosimilitud[10] para una observación distribuida log-SNP($\mu, \sigma^2, \boldsymbol{d}$) truncada en el octavo momento se especifica como:

$$logL(\mu, \sigma^2, \boldsymbol{d}|z_i) = -\frac{1}{2}\log(2\pi\sigma^2 z_i^2) - \frac{1}{2}\left(\frac{\log(z_i)-\mu}{\sigma}\right)^2 + \log\left[1 + \sum_{s=1}^{8} d_s H_s\left(\frac{\log(z_i)-\mu}{\sigma}\right)\right]. \quad \text{(II.17)}$$

La estimación secuencial comienza con la más simple densidad anidada, la lognormal, y recursivamente se agregan los parámetros d_s cuyos valores iniciales se seleccionan consistentemente con los momentos muestrales. La inclusión de nuevos parámetros en la distribución de productividad se realiza en función de criterios de precisión de los ajustes de la log-verosimilitud (logL) y el Criterio de Información de Akaike (AIC) y los contrastes de restricciones lineales proporcionados por la razón de verosimilitud (LR). Atendiendo a estos criterios se seleccionó *n*=8 como el orden óptimo de truncación óptimo y se incluyeron únicamente los parámetros pares d_2, d_4, d_6 y d_8.

[10] El código para la implementación del algoritmo de estimación de máxima verosimilitud en el paquete R está disponible bajo petición.

II.4. Resultados

La Tabla II.2 presenta la estimación ML para cada uno de los campos seleccionados. El Panel A muestra los parámetros estimados para la distribución lognormal y el Panel B muestra los parámetros estimados para la distribución log-SNP. En el Panel C se encuentra el estadístico LR para el contraste de distribución log-SNP frente a lognormal. Los resultados de la estimación muestran que todos los modelos capturan de manera adecuada la media y la desviación estándar de cada uno de los campos, capturada por los parámetros μ y σ respectivamente. Como se observa, los p-valores indican que estos parámetros son altamente significativos para ambas distribuciones. Es de destacar que el parámetro σ, que también captura la asimetría de la lognormal y la log-SNP siempre que no se incluyan parámetros impares, permanece muy estable para todas las distribuciones de productividad.

Esta evidencia es consistente con Ruiz-Castillo & Costas (2014) quienes encontraron que "a pesar de las grandes diferencias en las prácticas de producción y citación entre campos, la forma de las distribuciones de productividad es muy similar en todos los campos". Sin embargo, como se muestra en el Panel B, para la distribución log-SNP, los parámetros d_s también resultan altamente significativos para la mayoría de los campos del conocimiento. Al analizar el AIC (que penaliza la inclusión de parámetros adicionales) para las dos distribuciones, se encuentra que este criterio resulta consistentemente inferior para la distribución log-SNP, lo que sugiere que la modelización a partir de esta distribución resulta claramente superior. Además, los estadísticos LR incluidos en el Panel C, concluyen que, para todos los campos seleccionados, la incorporación de los parámetros d_s mejoran la razón de verosimilitudes del modelo.[11]

[11] Nótese que no se incluyen los parámetros d_s para s impar, después de haber probado que no eran significativamente diferentes de cero. Este hecho refuerza el hecho de que, para estas series de datos, el parámetro σ captura todas las características relevantes sobre la asimetría. Cabe destacar que este hecho no contradice el hecho de que los parámetros d_s para s pares son muy significativos, lo que significa que las distribuciones de productividad tienen colas muy pesadas y, por lo tanto, requieren parámetros diferentes para proporcionar medidas precisas de la "probabilidad de ser un investigador top" en cada campo.

Tabla II.2 Estimaciones para la distribución de la productividad bajo lognormal y log-SNP

Campo del conocimiento	Panel A Lognormal				Panel B Log-SNP								Panel C LR
	μ	σ	logL	AIC	μ	σ	d_2	d_4	d_6	d_8	logL	AIC	
Agronomy	0.2143	0.4368	-3359.52	6723.04	0.1182	0.4771	-0.0786	0.1448	0.0252	0.0042	-1890.49	3792.98	2938.07
	(<.0001)	(<.0001)			(0.000)	(<.0001)	(0.000)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Anthropology	0.3753	0.6024	-3089.70	6183.40	0.1693	0.5438	0.1912	0.2733	0.0408	0.0050	-2259.29	4530.58	1660.83
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Clinical psychology	0.3791	0.5994	-5501.26	11006.52	0.1689	0.5556	0.1535	0.2611	0.0444	0.0055	-4236.31	8484.63	2529.90
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Dentistry	0.4934	0.6763	-6598.224	13200.45	0.2959	0.6913	0.0194	0.1481	0.0157	0.0027	-5740.93	11493.86	1714.58
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(0.1765)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Dermatology	0.4553	0.6914	-18154.32	36312.64	0.8375	0.4294	1.1923	0.3812	0.1092	0.0179	-7262.16	14536.32	21784.32
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Ecology	0.3335	0.5445	-2736.83	5477.66	0.1653	0.5435	0.0499	0.1708	0.0174	0.0037	-2027.75	4067.50	1418.16
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(0.0023)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Economics	0.2887	0.5198	-1450.68	2905.37	0.1418	0.5133	0.0538	0.2073	0.0277	0.0041	-935.65	1883.29	1030.08
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(0.0819)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Educational psychology	0.3404	0.5320	-1356.60	2717.21	0.1764	0.5367	0.0381	0.1614	0.0194	0.0034	-1108.26	2228.51	496.70
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(0.0900)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Ethics	0.2952	0.5262	-516.72	1037.45	0.1556	0.5301	0.0282	0.2231	0.0351	0.0048	-338.55	689.11	356.34
	(<.0001)	(<.0001)			(0.0028)	(<.0001)	(0.4423)	(<.0001)	(0.0017)	(<.0001)			(<.0001)
Ethnic studies	0.2287	0.4647	-849.22	1702.44	0.1290	0.5045	-0.0854	0.1877	0.0347	0.0050	-511.39	1034.78	675.66
	(<.0001)	(<.0001)			(0.0038)	(<.0001)	(0.0011)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Finance	0.4560	0.6688	-1692.96	3389.92	0.1693	0.5763	0.2975	0.2992	0.0484	0.0060	-1390.41	2792.82	605.10
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)

continúa

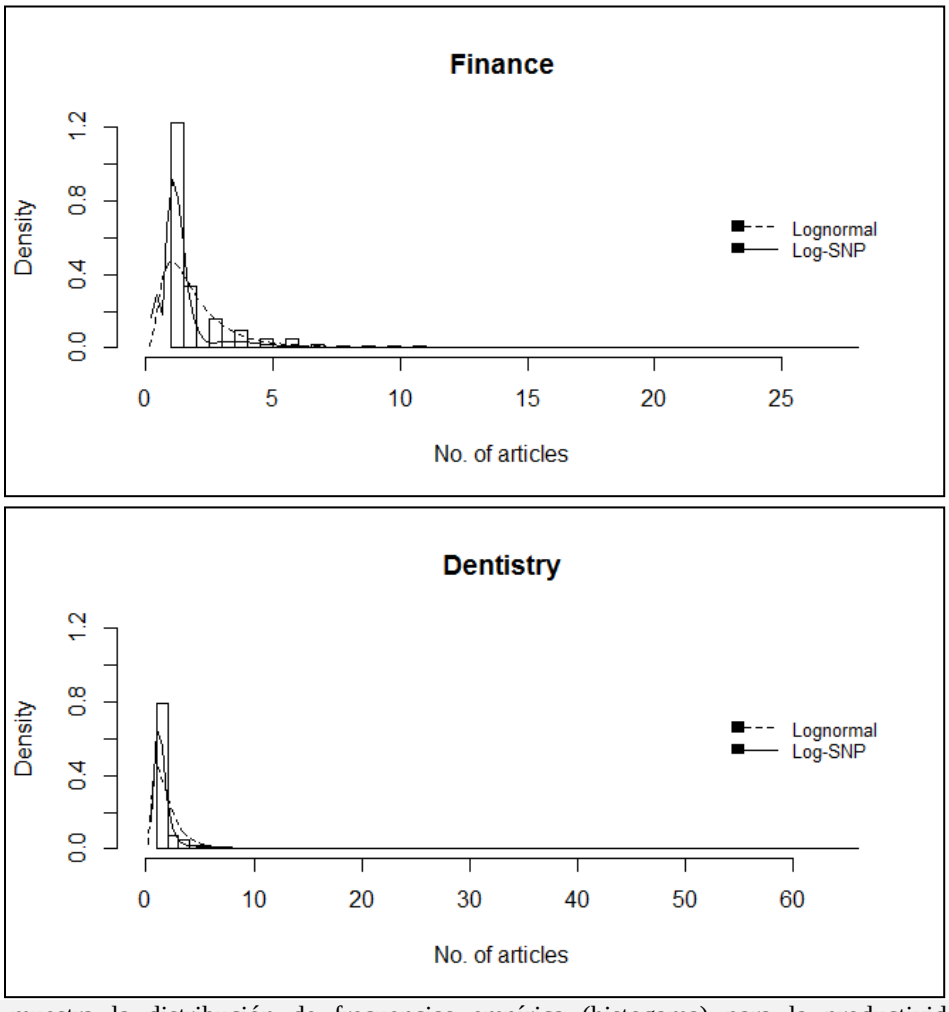
Tabla II.2 Continuación

Campo del conocimiento	Panel A Lognormal				Panel B Log-SNP								Panel C LR
	μ	σ	logL	AIC	μ	σ	d_2	d_4	d_6	d_8	logL	AIC	
Forestry	0.3785	0.5755	-5958.31	11920.63	0.1797	0.5490	0.1149	0.1942	0.0232	0.0037	-4879.51	9771.02	2157.61
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Genetics	0.3338	0.5350	-7617.37	15238.74	0.1720	0.5379	0.0399	0.1748	0.0224	0.0037	-6015.27	12042.54	3204.20
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
History	0.3080	0.4570	-2198.69	4401.39	0.1984	0.5112	-0.0776	0.0627	-0.0004	0.0013	-2095.15	4202.29	207.10
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.8251)	(<.0001)			(<.0001)
Law	0.2788	0.4908	-578.29	1160.59	0.1507	0.4953	0.0244	0.1747	0.0163	0.0027	-389.59	791.18	377.40
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(0.6560)	(<.0001)	´(0.0272)	(<.0001)			(<.0001)
Linguistics	0.3307	0.5556	-1801.66	3607.31	0.1558	0.5395	0.0844	0.2007	0.0246	0.0042	-1270.77	2553.54	1061.77
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Mathematics	0.2458	0.4342	-1346.20	2696.39	0.1652	0.4945	-0.1013	0.1159	0.0071	0.0019	-971.81	1955.62	748.77
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0210)	(<.0001)			(<.0001)
Statistics	0.4510	0.6390	-5553.69	11111.38	0.2429	0.6251	0.0779	0.1858	0.0253	0.0036	-4758.50	1590.38	1590.38
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)

Esta tabla incorpora la estimación ML para cada uno de los campos seleccionados. El Panel A muestra los parámetros estimados para la distribución lognormal. El Panel B muestra los parámetros estimados para la distribución log-SNP. El Panel C muestra la razón de verosimilitudes aplicada entre ambas distribuciones. μ y σ son los parámetros de localización y escala y d_2, d_4, d_6 y d_8 los parámetros de peso de los polinomios de Hermite. logL=log-verosimilitud, AIC=Criterio de Información Akaike, LR=razón de verosimilitudes para el contraste de log-SNP frente a lognormal. P-valores en paréntesis. El estudio corresponde a 18 campos de conocimiento pertenecientes a las categorías JCR de las ciencias y las ciencias sociales entre los años 2000 y 2009.

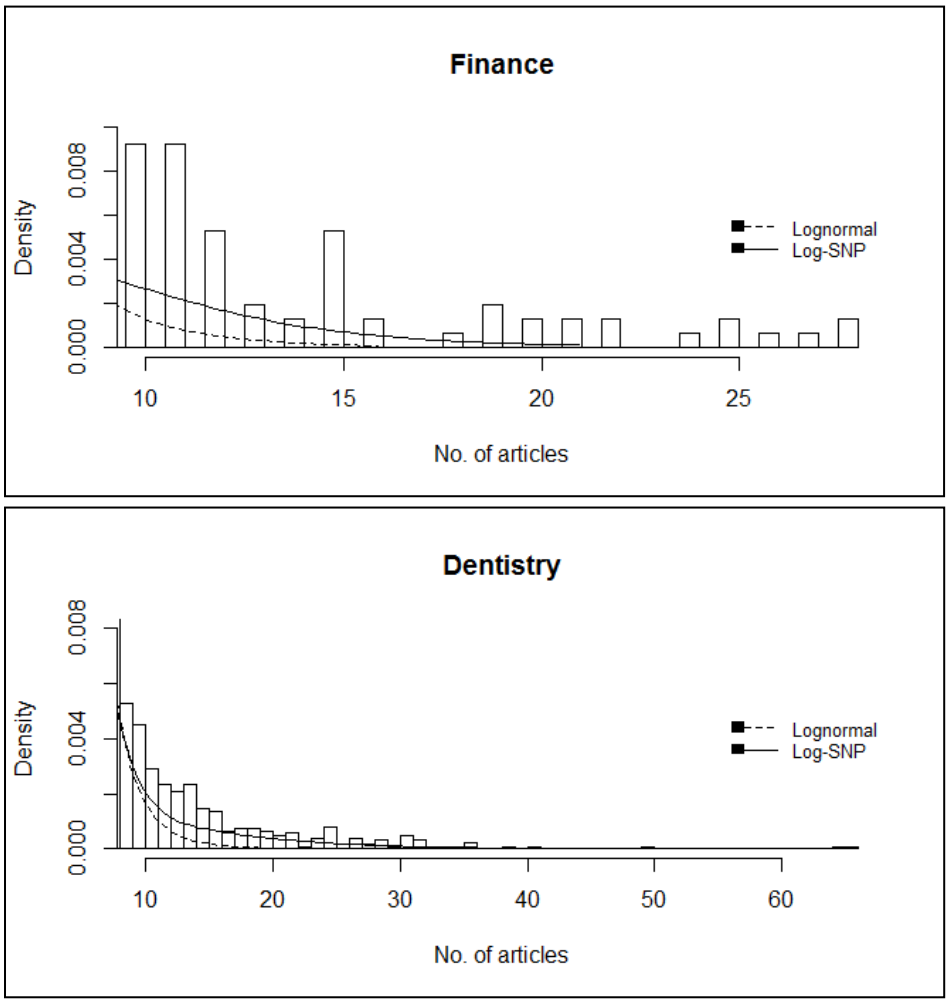
Un ejemplo de la calidad que se obtiene al ajustar dos campos seleccionados aleatoriamente, Finance y Dentistry, se recoge en la Figura II.3. En esta figura se muestran el histograma empírico y las pdf estimadas bajo una especificación lognormal y la log-SNP. En ambos casos la distribución log-SNP recoge más adecuadamente no sólo los valores en torno a la media sino, especialmente, los valores extremos. En la Figura II.4 se muestra un detalle de las colas derechas de la distribución que incluyen a los investigadores de alta productividad en ambas áreas. Como se observa, la especificación log-SNP permite caracterizar mejor la actividad investigadora.

Figura II.3 Pdf para la productividad investigadora en Finance y Dentistry



La figura muestra la distribución de frecuencias empírica (histogama) para la productividad de los investigadores que publicaron en las 5 mejores revistas (en términos JCR-2007) de Finance y Dentistry durante el periodo 2000-2009. Sobrepuestas se encuentran las pdf estimadas bajo una especificación lognormal y log-SNP.

Figura II.4 Pdf para la productividad investigadora en Finance y Dentistry (cola derecha)



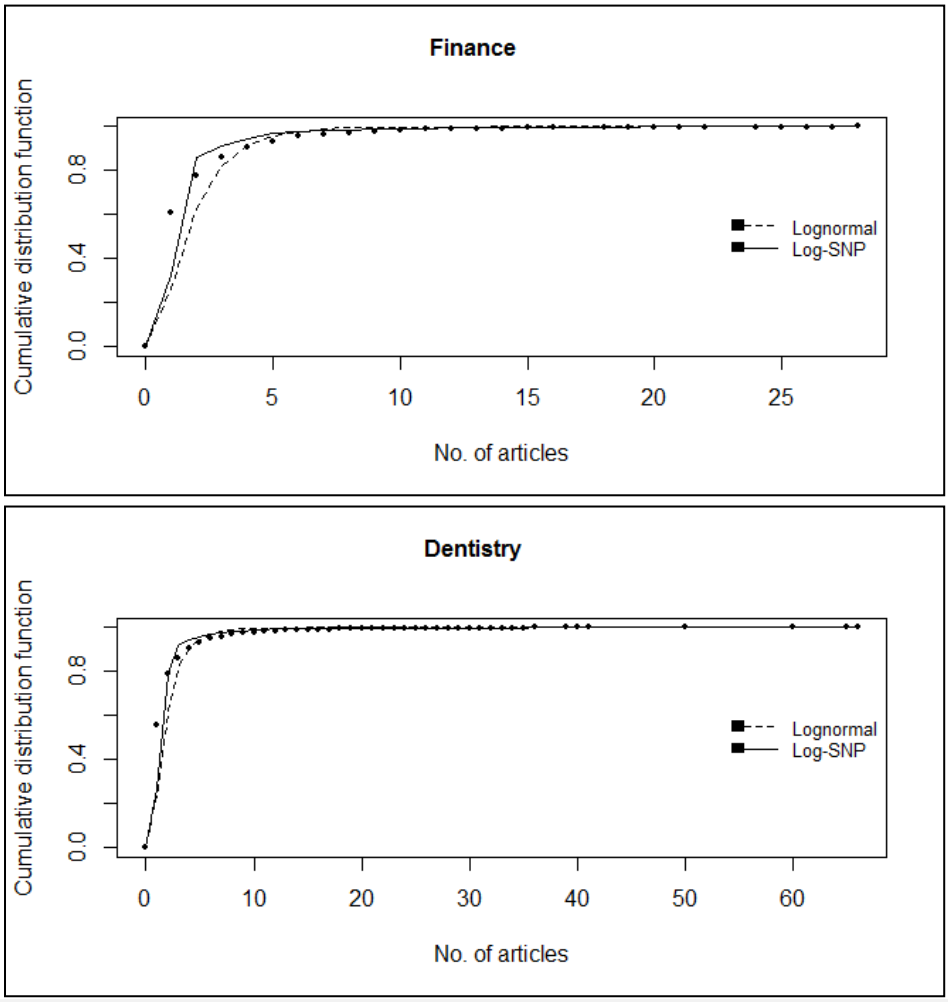
La figura muestra la cola derecha para la distribución de frecuencias empírica (histogama) para la productividad de los investigadores que publicaron en las 5 mejores revistas (en términos JCR-2007) de Finance y Dentistry durante el periodo 2000-2009. Sobrepuestas se encuentran las pdf estimadas bajo una especificación lognormal y log-SNP.

Alternativamente, la Figura II.5 muestra la comparación de los ajustes de la distribución para Finance y Dentristry en términos de la cdf empírica versus la teórica para ambas distribuciones, la log-SNP y la lognormal. Esta última parece infravalorar la probabilidad acumulada (especialmente para Dentistry) en comparación con la log-SNP.

La Figura II.4 muestra como la distribución lognormal subestima la productividad investigadora especialmente para los valores más extremos (bajo la distribución lognormal un investigador debe publicar menos artículos para ser considerado top). La Tabla II.3

ilustra[12] estos efectos para las distintas áreas del conocimiento calculando los cuantiles empíricos y estimados bajo lognormal y log-SNP para niveles de confianza del 5%, 1%, 0.1% y 0.05%.

Figura II.5 Cdf para la productividad investigadora en Finance y Dentistry



La figura muestra la función de distribución acumulada (cdf) empírica para la productividad de los investigadores que publicaron en las 5 mejores revistas (en términos JCR-2007) de Finance y Dentistry durante el periodo 2000-2009. Sobrepuestas se encuentran las pdf estimadas bajo una especificación lognormal y log-SNP.

[12] Para obtener los cuantiles de la distribución log-SNP se usa la cdf presentada en la ecuación (II.15) y el Método de la Transformada Inversa (ITM).

Obsérvese que, una vez estimadas las distribuciones de productividad, la definición de un investigador top en cada campo requiere sólo del cálculo del cuantil correspondiente para una probabilidad dada. Estos cuantiles representan los límites del desempeño en términos de número de artículos (independientemente del número de autores), siempre que la calidad se garantice al considerar sólo las publicaciones en los 5 primeros journals en todos los campos del conocimiento. Además, estos cuantiles son bastante comparables entre diferentes áreas.

La interpretación de los valores de la Tabla II.3 muestra claramente la mayor precisión de la distribución log-SNP en los ajustes en las colas y los errores inducidos en la estimación de la productividad de los investigadores top mediante el uso de distribuciones paramétricas tradicionales como la lognormal.

Por ejemplo, para el campo del conocimiento Agronomy se observa empíricamente que para pertenecer al top 0.05% de los investigadores que publican un mayor número de artículos en las mejores revistas se requieren empíricamente 15 publicaciones mientras que ese límite es mucho más laxo si se asume que la distribución es lognormal (6) que si se utiliza una log-SNP (12). Estos resultados son consistentes con las investigaciones de Kumar, Sharma, & Garg (1998), Perc (2010) y Eom & Fortunato, (2011) quienes al aplicar la distribución lognormal a indicadores bibliométricos encontraron que ésta se queda corta a la hora de modelizar series con colas muy largas.

Tabla II.3 Número de artículos observados empíricamente versus esperados teoricamente bajo lognormal y log-SNP

Campo del conocimiento	N	No. de artículos top observados				Número de artículos esperados							
						Lognormal Top				Log-SNP Top			
		5%	1%	0.1%	0.05%	5%	1%	0.1%	0.05%	5%	1%	0.1%	0.05%
Agronomy	8,923	3	7	13	15	3	4	5	6	3	4	10	12
Anthropology	5,755	5	10	19	22	4	6	10	11	4	9	16	17
Clinical psychology	10,418	5	11	27	35	4	6	10	11	4	9	17	19
Dentistry	12,345	7	15	32	36	5	8	14	16	5	11	29	34
Dermatology	30,531	7	16	40	50	5	8	14	16	7	14	20	22
Ecology	5,730	4	8	17	20	4	5	8	9	4	7	14	16
Economics	3,048	4	8	25	26	4	5	7	8	3	7	13	14
Educational psychology	3,032	4	8	18	18	4	5	8	9	4	7	14	16
Ethics	1,073	4	9	24	25	4	5	7	8	3	8	14	16
Ethnic studies	2,003	3	8	16	16	3	4	6	6	3	5	12	14
Finance	3,019	6	13	26	28	5	8	13	15	5	11	19	21
Forestry	12,211	5	9	18	22	4	6	9	10	4	8	15	17
Genetics	16,574	4	8	18	23	4	5	8	9	4	7	14	16
History	6,708	3	5	8	12	3	4	6	7	3	5	8	11
Law	1,350	4	7	13	13	3	5	7	7	3	6	11	12
Linguistics	3,600	5	9	22	23	4	6	8	9	4	7	14	16
Mathematics	3,972	3	6	13	14	3	4	5	6	3	5	10	11
Statistics	10,679	6	13	26	35	5	7	12	13	5	10	22	26

Esta tabla compara el número de artículos observados empíricamente en cada uno de los campos versus los esperados teóricamente bajo las distribuciones lognormal y log-SNP. N=número de investigadores por campo del conocimiento. Los valores 5%, 1%, 0.1% y 0.05% son cuantiles de las distribuciones. El estudio corresponde a 18 campos del conocimiento pertenecientes a las categorías JCR de las ciencias y las ciencias sociales entre los años 2000 y 2009.

II.4.1. Otros resultados

Este capítulo propone una nueva metodología para calcular la productividad investigadora de los mejores investigadores a través de los cuantiles de una nueva y general distribución denominada log-SNP. La aplicación principal compara estas medidas con las de la distribución lognormal con una muestra de producción científica en 18 campos (arbitrariamente elegidos), encontrando un desempeño superior de la distribución log-SNP. Sin embargo, el enfoque del trabajo se hace sobre la técnica más que sobre los resultados particulares.

Con el fin de justificar que los resultados presentados a lo largo de este estudio son generales, se ha replicado el estudio con los datos de productividad proporcionados en Borokhovich, Bricker, Brunarski, & Simkins (1995), los cuales se refieren a las instituciones académicas (en el campo de las finanzas) en lugar de los investigadores individuales. En particular, los datos incorporan el número de artículos publicados entre 1989 y 1993 en un conjunto de 16 journals de finanzas de autores afiliados a diferentes instituciones en el momento de la publicación. Los journals de finanzas (excluyendo bienes raíces y seguros) fueron seleccionados de los que figuran en el Heck's Finance Literature Index de 1993. Sólo se incluyeron artículos y notas en la muestra. El número de publicaciones atribuidas a cada institución académica se ajustó al número de autores. Por ejemplo, para publicaciones con dos autores afiliados a diferentes instituciones, cada institución recibió crédito por 0.5 artículo. Cualquier proporción de un artículo que no fuera atribuible a un autor afiliado a una institución académica ubicada en los Estados Unidos o Canadá fue eliminada del estudio. Un total de 330 instituciones fueron incluidas en esta muestra.

Tabla II.4 Resultados. Productividad institucional de la investigación bajo lognormal y log-SNP

Panel A: Estimaciones de la distribución de la productividad bajo lognormal y log-SNP

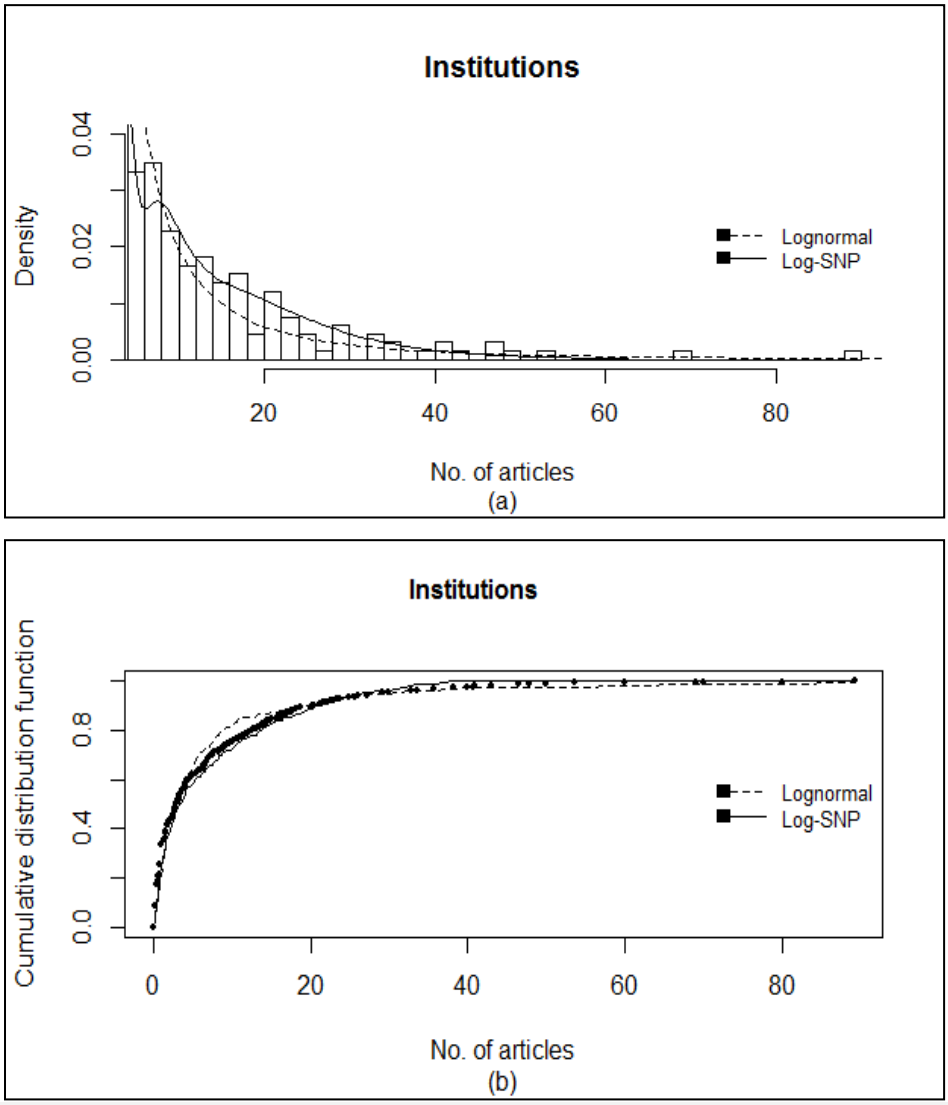
Productividad institucional	Lognormal				Log-SNP								LR
	μ	σ	logL	AIC	μ	σ	d_2	d_4	d_6	d_8	logL	AIC	
Instituciones académicas	1.102	1.4319	-950.38	1904.76	1.1574	0.61	2.2593	1.1595	0.1979	0.011	-920.19	1852.38	2157.61
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0018)			(<.0001)

Panel B: Número de artículos empíricamente observados frente a los teóricamente esperados en la lognormal y log-SNP

Productividad institucional	N	No. de artículos observados				Número de artículos esperados							
						Lognormal				Log-SNP			
		5%	1%	0.10%	0.05%	5%	1%	0.10%	0.05%	5%	1%	0.10%	0.05%
Instituciones académicas	330	30	50	83	86	32	85	252	335	29	45	70	78

La tabla incorpora los resultados de la productividad investigadora de las instituciones académicas que publicaron en un conjunto de 16 journals financieros entre los años 1989 y 1993. El Panel A presenta las estimaciones de ML para los parámetros de las distribuciones lognormal y log-SNP y la razón de verosimilitud para probar las diferencias entre ellas. μ y σ son los parámetros de localización y escala, respectivamente, y d_2, d_4, d_6 y d_8 los parámetros de peso de los polinomios de Hermite. logL=log-verosimilitud, AIC=Criterio de Información Akaike, LR=razón de verosimilitudes para el contraste de log-SNP frente a lognormal. P-valores en paréntesis. El Panel B compara el número de artículos observados empíricamente frente a los esperados teóricamente bajo las distribuciones lognormal y log-SNP. N=número de instituciones académicas. Los valores 5%, 1%, 0.1% y 0.05% son cuantiles de las distribuciones.

Figura II.6 Pdf y cdf de la productividad investigadora institucional



La figura muestra: (a) la cola derecha para la distribución de frecuencias empírica (histogama) para la productividad investigadora de las instituciones académicas que publicaron en un conjunto de 16 journals financieros entre los años 1989 y 1993. Sobrepuestas se encuentran las pdf estimadas bajo una especificación lognormal (línea discontinua) y log-SNP (línea sólida). (b) La función de distribución acumulada empírica para la misma muestra. Sobrepuestas se encuentran las cdfs de la lognormal (línea discontinua) y la log-SNP (línea sólida).

La Tabla II.4 presenta los resultados de esta nueva estimación. El Panel A muestra las estimaciones de ML para la productividad investigadora de las instituciones académicas en los principales journals de finanzas. Los resultados son consistentes con los obtenidos previamente para los investigadores en diferentes campos del conocimiento, es decir, de

acuerdo con la prueba LR la distribución log-SNP supera la lognormal y por lo tanto los parámetros log-SNP son altamente significativos. El panel B compara el número de artículos empíricamente observados con los teóricamente esperados bajo ambas especificaciones revelando el desempeño superior de la distribución log-SNP. En este caso, parece que la lognormal sobreestima las colas de distribución, en particular para bajos niveles de confianza. Este resultado corrobora la evidencia de que el uso de distribuciones rígidas implica resultados engañosos porque son incapaces de ajustarse a diferentes características de la distribución (particularmente valores extremos) con un solo (o dos) parámetro(s).

La Figura II.6 ilustra la valoración anterior que muestra el mejor ajuste de la distribución log-SNP en términos de la pdf de las colas derechas (Figura (a)) y la cdf (Figura (b)) en comparación con las distribuciones empíricas.

II.5. Conclusiones

El análisis bibliométrico ha mostrado ser un método valioso para la evaluación de la producción científica y tiene un impacto creciente. Sin embargo, la literatura indica que en algunos casos, las distribuciones habitualmente utilizadas para las mediciones de productividad son incapaces de representar el comportamiento de los investigadores top. Esto se debe a que su productividad parece regirse por una distribución que presenta colas muy pesadas, lo que pone de manifiesto la necesidad de proponer otras distribuciones y metodologías adecuadas para tales propósitos.

Esta investigación analiza la productividad investigadora en 18 campos del conocimiento pertenecientes a las categorías JCR de las ciencias y las ciencias sociales entre los años 2000 y 2009. Los resultados muestran que el nivel de productividad, medido por el número de publicaciones por autor, depende del campo del conocimiento estudiado, lo cual es consistente con la evidencia previa. En especial, los campos que pertenecen a la categoría de las ciencias cuentan con un mayor número de publicaciones por autor. Adicionalmente, se observa que el indicador MIF se encuentra altamente correlacionado con el número máximo de artículos por investigador. Es decir, a mayor número de artículos publicados en revistas top por cada investigador (habitualmente los más citados), mayor MIF por campo del conocimiento.

En este trabajo se propone una novedosa metodología basada en la distribución log-SNP para la medición de la distribución de la productividad científica de los investigadores top en las distintas áreas del conocimiento. Dicha distribución anida a la lognormal incluyendo nuevos parámetros capaces de recoger mejor el comportamiento de la cola de alta productividad y permitiendo contrastar las deficiencias de la distribución lognormal en esta dirección. El presente estudio muestra que la log-SNP provee un mejor ajuste de la distribución de desempeño de los investigadores y de las instituciones académicas y cuantifica las diferencias en las medidas de la productividad de los mejores investigadores unidas a la hipótesis distribucional.

Este estudio argumenta que la log-SNP es un proceso de generación de datos preciso para la productividad de los mejores investigadores, por lo que este proceso es más fiable que la lognormal (que está anidado en la log-SNP), ya que la log-SNP es más flexible cuando los datos son muy sesgados y hay posibles saltos en la cola debido a observaciones extremas. Por lo tanto, se ofrece una metodología interesante para medir la productividad científica que se puede utilizar cuando el desempeño de autores, instituciones o campos tienen que ser comparados o agregados para implementar políticas basadas en el desempeño académico.

Referencias

Abramo, G., & D'Angelo, C. A. (2014). Assessing national strengths and weaknesses in research fields. *Journal of Informetrics, 8(3),* 766–775.

Abramo, G., D'Angelo, A. C., & Pugini, F. (2008). The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology. *Scientometrics, 76(2),* 225–244.

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover Publications.

Aguinis, H., O'Boyle, E., Gonzalez-Mulé, E., & Joo, H. (2015). Cumulative advantage: Conductors and insulators of heavy-tailed productivity distributions and productivity tars. *Personnel Psychology*, http://dx.doi.org/10.1111/peps.12095 (in press).

Albarrán, P., Juan, A. C., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics, 88(2),* 385-397.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy, 44(2),* 451-466.

Birkmaier, D., & Wohlrabe, K. (2014). The Matthew effect in economics reconsidered. *Journal of Informetrics, 8(4),* 880–889.

Blinnikov, S., & Moessner, R. (1998). Expansions for nearly Gaussian distributions. *Astronomy and astrophysics Supplement Series, 130(1),* 193–205.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45(1),* 199–245.

Borokhovich, K. A., Bricker, R. J., Brunarski, K. R., & Simkins, B. J. (1995). Finance research productivity and influence. *The Journal of Finance, 50(5),* 1691-1717.

Broadus, R. N. (1987). Toward a definition of 'bibliometrics'. *Scientometrics, 12(5-6),* 373-379.

Campanario, J. M. (2015). Providing impact: The distribution of JCR journals according to references they contribute to the 2-year and 5-year journal impact factors. *Journal of Informetrics, 9(2),* 398–407.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. Heckman, & E. Leamer, *Handbook of Econometrics, Vol. 6. Part B* (pp. 5549-5632). Amsterdam: Elsevier.

Chung, K. H., & Cox, R. A. (1990). Patterns of productivity in the finance literature: a study of the bibliometric distributions. *The Journal of Finance, 45(1) 1, 301-309*, 301-309.

Coupé, T. (2003). Revealed performances. Worldwide rankings of economists and economics departments. *Journal of the European Economic Association, 1(6)*, 1309–1345.

Cramér, H. (1925). On some classes of series used in mathematical statistics. *Sixth Scandinavian Congress of Mathematicians*, (págs. 399-425). Copenhagen.

Crespo, J. A., Ortuño-Ortín, I., & Ruiz-Castillo, J. (2012). The citation merit of scientific publications. *PLoS ONE 7(11)*, e49156.

Da Silva, R., Kalil, F., De Oliveira, J. M., & Martinez, A. S. (2012). Universality in bibliometrics. *Physica A: Statistical Mechanics and its Applications, 391(5)*, 2119-2128.

Day, T. E. (2015). The big consequences of small biases: A simulation of peer review. *Research Policy, 44(6)*, 1266–1270.

Del Brio, E. B., & Perote, J. (2012). Gram–Charlier densities: Maximum likelihood versus the method of moments. *Insurance: Mathematics and Economics, 51(3)*, 531-537.

Duch, J., Zeng, X. T., Sales-Pardo, M., Radicchi, F., Otis , S., Woodruff, T. K., & Nunes Amaral, L. A. (2012). The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS ONE 7(12)*, e51332.

Dundar, H., & Lewis, D. (1998). Determinants of research productivity in higher education. *Research in Higher Education, 39(6)*, 607-631.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics.* Kidlington, UK: Elsevier Academic Press.

Ellison, G. (2013). How does the market use citation data? the hirsch index in economics. *American Economic Journal: Applied Economics, 5(3)*, 63-90.

Eom, Y. H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE, 6(9)*, e24926.

Finardi, U. (2013). Correlation between Journal Impact Factor and Citation Performance: An experimental study. *Journal of Informetrics, 7(2)*, 357–370.

Frandsen, T. F. (2005). Geographical concentration. The case of economics journals. *Scientometrics, 63(1)*, 69-85.

Gallant, A. R., & Nychka, D. W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica, 55(2),* 363–390.

Garfield, E. (1980). Bradford's Law and related statistical pattern. *Essays of an Information Scientist, 4(19),* 476-483.

Genest, C. (1997). Statistics on statistics: measuring research productivity by journal publications between 1985 and 1995. *The Canadian Journal of Statistics, 25(4),* 427-443.

Guerrero-Bote, V. P., Zapico-Alonso, F., Espinosa-Calvo, M. E., Gomez-Crisostomo, R., & Moya-Anegon, F. (2007). Import–export of knowledge between scientific subject categories: The iceberg hypothesis. *Scientometrics, 71(3),* 423–441.

Harzing, A. (2008). Publish or Perish: A citation analysis software program. *Available from http://www.harzing.com/resources.htm.*

Heberger, A. E., Christie, C. A., & Alkin, M. C. (2010). A bibliometric analysis of the academic influences of and on evaluation theorists' published works. *American Journal of Evaluation, 31(1),* 24-44.

Hodgson, G. M., & Rothman, H. (1999). The editors and authors of economics journals: A case of institutional oligopoly? *The Economic Journal, 109(453)* , 165–186.

Kaur, J., Ferrara, E., Menczer, F., Flammini, A., & Radicchi, F. (2015). Quality versus quantity in scientific impact. *Journal of Informetrics, 9(4),* 800-808.

Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics, 7(4),* 924–932.

Kendall, M., & Stuart, A. (1977). *The Advanced Theory of Statistics, Vol. I, 4th ed.* London: C. Griffin.

Kocher, M. G., Luptacik, M., & Sutter, M. (2006). Measuring productivity of research in economics: A cross-country study using DEA. *Socio-Economic Planning Sciences, 40(4),* 314-332.

Kretschmer, H., & Kretschmer, T. (2007). Lotka's distribution and distribution of co-author pairs' frequencies. *Journal of Informetrics, 1(4),* 308–337.

Kumar, S., Sharma, P., & Garg, K. C. (1998). Lotka's law and institutional productivity. *Information Processing & Management, 34(6),* 775–783.

Lancho-Barrantes, B. S., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). The iceberg hypothesis revisited. *Scientometrics, 85(2),* 443–461.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science, 16(12),* 317-323.

Martínez-Mekler, G., Martínez, R. A., del Río, M. B., Mansilla, R., Miramontes, P., & Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE, 4*(3), e4791.

Mauleón, I., & Perote, J. (2000). Testing densities with financial data: an empirical comparison of the Edgeworth-Sargan density to the Student's t. *European Journal of Finance, 6(2),* 225-239.

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research, 246(1),* 1-19.

Newman, M. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323-351.

Nicholls, P. T. (1989). Bibliometric modelling processes and the empirical validity of Lotka's law. *Journal of the American Society for Information Science, 40(6),* 379–385.

Nicholls, T. P. (1986). Empirical validation of Lotka's law. *Information Processing & Management, 22(5),* 417–419.

Nicolaisen, J., & Hjørland, B. (2007). Practical potentials of Bradford's law: a critical examination of the received view. *Journal of Documentation, 63(3),* 359 - 377.

Ñíguez, T-M., Paya, I., Peel, D., & Perote, J. (2012). On the stability of the constant relative risk aversion (CRRA) utility under high degrees of uncertainty. *Economics Letters, 115(2),* 244-248.

Ñíguez, T-M., Paya, I., Peel, D., & Perote, J. (2013). Higher-order moments in the theory of diversification and portfolio composition. *Economics Working Paper Series 2013/003. Lancaster University.*

O'Boyle, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology, 65(1),* 79–119.

Perc, M. (2010). Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. *Journal of Informetrics, 4(2),* 358–364.

Phillips, P. B. (1977). A general theorem in the theory of asymptotic expansions as approximations to the finite sample distributions of econometric estimators. *Econometrica, 45(6),* 1517–1534.

Price, D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27(5),* 292–306.

Radicchi, F., Fortunado, S., & Castellano, C. (2008). Universality of citation distribution: Towards an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America, 105(45)*, 17268–17272.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems, 4(2)*, 131-134.

Rousseau, R. (1994). Bradford curves. *Information Processing and Management, 30(2)*, 267–277.

Ruiz-Castillo, J., & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics, 8(4)*, 917–934.

Sabharwal, M. (2013). Comparing research productivity across disciplines and career stages. *Journal of Comparative Policy Analysis: Research and Practice, 15(2)*, 141-163.

Sargan, D. (1975). Gram-Charlier approximation applied t ratios or k-class estimatiors. *Econometrica, 43(2)*, 327-346.

Seggie, S. H., & Griffith, D. A. (2009). What does it take to get promoted in marketing academia? Understanding exceptional publication productivity in the leading marketing journals. *Journal of Marketing, 73(1)*, 122-132.

Wallace, D. L. (1958). Asymptotic approximations to distributions. *Annals of Mathematical Statistics, 29(3)*, 635-654.

Williamson, I. O., & Cable, D. M. (2003). Predicting early career research productivity: The case of management faculty. *Journal of Organizational Behavior, 24(1)*, 25-44.

Apéndice II.A

Este apéndice recoge los primeros ocho parámetros d_s en términos de los momentos de la distribución. Para más información véase Del Brio & Perote (2012)

$$d_1 = \mu_{1,} \tag{II.A.1}$$

$$d_2 = \tfrac{1}{2}(\mu_2 - 1), \tag{II.A.2}$$

$$d_3 = \tfrac{1}{6}(\mu_3 - 3\mu_1), \tag{II.A.3}$$

$$d_4 = \tfrac{1}{24}(\mu_4 - 6\mu_2 + 3), \tag{II.A.4}$$

$$d_5 = \tfrac{1}{120}(\mu_5 - 10\mu_3 + 15\mu_1), \tag{II.A.5}$$

$$d_6 = \tfrac{1}{720}(\mu_6 - 15\mu_4 + 45\mu_2 - 15), \tag{II.A.6}$$

$$d_7 = \tfrac{1}{5040}(\mu_7 - 21\mu_5 + 105\mu_3 - 105\mu_1), \tag{II.A.7}$$

$$d_8 = \tfrac{1}{40320}(\mu_8 - 28\mu_6 + 210\mu_4 - 420\mu_2 + 105). \tag{II.A.8}$$

Apéndice II.B

Prueba 1. La función de distribución acumulada (cdf) de la distribución SNP se puede obtener como:

$$G_x(a) = \int_{-\infty}^{a} g(x;\boldsymbol{d})dx = \int_{-\infty}^{a} \phi(x)dx + \sum_{s=1}^{n} d_s \int_{-\infty}^{a} H_s(x)\phi(x)dx$$

$$= \int_{-\infty}^{a} \phi(x)dx - \sum_{s=1}^{n} d_s H_{s-1}(x)\phi(x)\Big|_{-\infty}^{a}$$

$$= \textstyle\int_{-\infty}^{a} \phi(x)dx - \phi(a)\sum_{s=1}^{n} d_s H_{s-1}(a) \tag{II.B.1}$$

Dado que $\lim_{x\to\pm\infty} H_s(x)\phi(x) = 0 \ \ \forall s \geq 1$, se obtiene

$$\int H_s(x)\phi(x)dx = \int (-1)^s \frac{d^s\phi(x)}{dx^s}dx_t = (-1)^s \frac{d^{s-1}\phi(x)}{dx^{s-1}}$$

$$= (-1)^s(-1)^{s-1}H_{s-1}(x)\phi(x) = -H_{s-1}(x)\phi(x).\square$$

CAPÍTULO III. Medición de la distribución del tamaño de la empresa con densidades semi-noparamétricas [*]

III.1. Introducción

El estudio de la distribución del tamaño de la empresa ha despertado gran interés tanto entre investigadores del campo de la física como de la economía (Stanley, et al., 1995; Hart & Oulton, 1997; Gupta, Campanha, de Aguiar, Queiroz, & Raheja, 2007; Hernández-Pérez, 2010). La relevancia de este tema radica en el hecho de que a partir del conocimiento sobre la forma de la distribución del tamaño de la empresa, los investigadores y los responsables de formular políticas pueden obtener información sobre el grado de concentración industrial, los ciclos económicos y para implementar leyes de competencia (Simon & Bonini, 1958; Barba Navaretti, Castellani, & Pieri, 2014; Heinrich & Dai, 2016).

Como estudio pionero sobre la distribución del tamaño de la empresa, Gibrat (1931) obtuvo que el tamaño de la empresa podría ser descrito por la distribución lognormal. Desde entonces varias aplicaciones han validado el uso de esa misma distribución (Stanley, et al., 1995; Voit, 2001). Sin embargo, diferentes tipos de distribuciones se han propuesto. Algunas investigaciones empíricas han mostrado que la distribución del tamaño puede ajustarse a partir de una distribución de Pareto o Power-law (Kaizoji, Iyetomi, & Ikeda, 2006; Coad, 2010) o también puede aproximarse partir de la Zipf's law (Axtell, 2001).

Así, una línea de la literatura empírica se ha encargado de examinar la aplicación de las distribuciones lognormales y tipo Pareto o Power-law usando datos del tamaño de la empresa de corte transversal (Cabral & Mata, 2003; Marsili, 2006; Goddard, Liu, Donal, & Wilson, 2014). Sin embargo, existe evidencia de que en ocasiones se obtiene una pobre aproximación a las distribuciones empíricas del tamaño de la empresa en la cola superior, que típicamente exhiben mayor asimetría, ya que un pequeño número de grandes empresas conviven con un

[*] Una versión de este capítulo se encuentra publicada bajo el título "Measuring firm size distribution with semi-nonparametric densities" en el journal *Physica A: Statistical Mechanics and its Applications*, JCR 2015 (1.785-Q2).

gran número de empresas más pequeñas (Stanley, et al., 1995; Axtell, 2001; Bottazzi, Pirino, & Tamagni, 2015).

Por ejemplo, en su investigación Stanley, et al. (1995) encontraron que la distribución del tamaño para una serie de empresas que cotizaban en bolsa en Estados Unidos se ajustaba bien a la distribución lognormal, con la excepción de la cola superior. En ese caso, la distribución lognormal sobrestimó el tamaño de las grandes empresas. Por su parte, Goddard, Liu, Donal, & Wilson (2014) examinaron la distribución del tamaño de la empresa de los bancos y las cooperativas de crédito a partir de la Zipf's law. Ese estudio la Zipf's law se rechazó como modelo de ajuste para la distribución del tamaño de la empresa en la cola superior.

Las divergencias obtenidas al aplicar este tipo de distribuciones al tamaño de la empresa, han llevado a los estudiosos sobre el tema a una discusión en cuanto a la estabilidad de un modelo único de probabilidad de tamaño de la empresa a través del tiempo, de industrias y países (Simon & Bonini, 1958; Dosi, Marsili, Orsenigo, & Salvatore, 1995; Cabral & Mata, 2003; Marsili, 2006; Crosato & Ganugi, 2007). Estas discrepancias probablemente se deban a que al analizar las distribuciones tradicionalmente usadas para modelizar datos con colas muy pesadas, tienen como inconveniente que éstas dependen de muy pocos parámetros para capturar toda la forma de la distribución del tamaño de la empresa incluyendo la cola derecha de la distribución (Crosato & Ganugi, 2007). Al respecto, Newman (2005) y Martínez-Mekler, et al. (2009) afirman que pocos procesos del mundo real siguen una distribución tipo Pareto o Power-law en todo su rango y, en particular, este tipo de distribuciones no ajustan los valores más pequeños de la variable que se mide.

Además, el punto de partida común bajo la hipótesis de la Zipf's law es suponer que la distribución del tamaño de la empresa está bien descrita por una distribución de Pareto o Power-law por encima de un cierto umbral mínimo (di Giovanni, Levchenko, & Rancière. , 2011; Goddard, Liu, Donal, & Wilson, 2014; Bottazzi, Pirino, & Tamagni, 2015; Pascoal, Augusto, & Monteiro, 2016). De esa manera, si se desea estudiar el crecimiento de las empresas más pequeñas en comparación con el de las empresas más grandes, no se puede

utilizar una distribución tipo Pareto o Power-law porque las pequeñas empresas se encuentran en la cola superior por debajo del valor umbral (Hart & Oulton, 1997; Cirillo & Hüsler, 2009).

Con el objeto de modelizar la distribución del tamaño de la empresa, este capítulo propone el uso de aproximaciones semi-noparamétricas (SNP) basadas en expansiones de Edgeworth y Gram-Charlier. Estas distribuciones se han aplicado en campos muy diversos en los que la precisión en la medición de las colas de la distribución es importante para la correcta medición de la ocurrencia de los valores extremos (véase Kuhs, 1988; Blinnikov & Moessner, 1998; Mauleón & Perote, 2000 o Cortés, Mora-Valencia, & Perote, 2016, como ejemplos de aplicaciones en termodinámica, astronomía, finanzas y cienciometría, respectivamente).

En este capítulo se plantea por primera vez el uso de las mismas para modelizar la distribución del tamaño de la empresa y en particular se proponen transformaciones logarítmicas de una distribución SNP (log-SNP) que son extensiones de una distribución lognormal que permiten aproximar cualquier distribución empírica mediante la introducción de parámetros adicionales. Con esta transformación se busca mantener la flexibilidad de los parámetros de las distribuciones de Gram-Charlier, pero restringiendo el dominio a valores positivos. En este estudio se muestra que, en comparación con la distribución lognormal, la distribución log-SNP proporciona un mejor ajuste al modelizar la distribución del tamaño de la empresa usando diferentes niveles de agregación industrial. También se muestra que distribución log-SNP permite obtener un mejor ajuste en los cuantiles superiores sin tener que imponer un umbral mínimo. Esto es importante ya que conocer el comportamiento de las empresas más grandes y que tienen mayor peso en el mercado es fundamental para analizar la evolución de la economía (Cirillo & Hüsler, 2009). Adicionalmente, se desarrolla por primera vez una expresión para la densidad de la distribución log-SNP multivariante cuyas densidades marginales se comportan como distribuciones univariantes de la log-SNP. La ventaja de desarrollar un marco multivariante se basa en que se obtienen estimaciones más eficientes que permiten analizar conjuntamente el comportamiento de variables altamente correlacionadas.

Este capítulo se divide de la siguiente forma: la sección III.2 define la distribución log-SNP univariante y desarrolla la expresión para la distribución multivariante. La sección III.3

presenta los datos a utilizar e incorpora los resultados del enfoque univariante y multivariante, y finalmente la sección III.4 resume las principales conclusiones del capítulo.

III.2. La distribución log-SNP

Esta sección define la función de densidad de probabilidad (pdf) de la distribución log-SNP y proporciona una extensión directa de esta distribución al caso multivariante. Como esta distribución es una transformación logarítmica de la denominada distribución de Gram-Charlier (o SNP) comenzamos definiendo esta clase de densidades y revisando algunas de sus principales propiedades.

Definición III.2.1: *La densidad Gram-Charlier de una variable aleatoria x_i es una clase general de densidades del tipo*

$$f(x_i; \boldsymbol{d_i}) = \phi(x_i) \sum_{s=0}^{n} d_{is} H_s(x_i) = \phi(x_i) p_i(x_i),\ x_i \in \mathbb{R}, \qquad \text{(III.1)}$$

donde $\phi(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}{x_i}^2}$ es la pdf normal estándar, $H_s(x) = \frac{(-1)^s}{\phi(x_i)} \frac{d^s \phi(x_i)}{d{x_i}^s}$ es el polinomio de Hermite de orden s[th 14], $\boldsymbol{d_i} = (d_{i1}, \dots, d_{in})' \in \mathbb{R}^n$ y n el orden de la expansión.

La condición $d_{i0} = 1$ es suficiente para garantizar que la función (III.1) integre uno, pero la no-negatividad no está garantizada para todos los $\boldsymbol{d_i} \in \mathbb{R}^n$, luego restricciones de positividad deben tenerse en cuenta para asegurar una familia bien definida de densidades.[15] Esta densidad anida la normal estándar (cuando $\boldsymbol{d_i}$ =**0**) y presenta dos ventajas principales con respecto a otras pdfs utilizadas para ajustar las distribuciones empíricas: *(i)* Incorpora suficientes grados de libertad para capturar cualquier momento de la distribución con una estructura paramétrica flexible; *(ii)* La expansión asintótica (cuando $n \to \infty$) captura la

[14] Los primeros polinomios de Hermite son $H_0(x_i) = 1$, $H_1(x_i) = x_i$, $H_2(x_i) = x_i^2 - 1$, $H_3(x) = x_i^3 - 3x_i$, $H_4(x_i) = x_i^4 - 6x_i^2 + 3$.

[15] Véase Jondeau & Rockinger (2001) para una descripción de la región de positividad en términos del sesgo y la curtosis. Alternativamente, la densidad Gram-Charlier se puede definir como $f^*(x_i; \boldsymbol{d}) = \phi(x_i) p_i(x_i)^2$, aunque a costa de una creciente complejidad. Sin embargo, la formulación positiva puede representarse en términos de una expansión mayor del tipo definido en (III.1) – ver Leon, Mencía, & Sentana (2009) – y los algoritmos de estimación de máxima verosimilitud convergen necesariamente en valores que garantizan una pdf bien definida.

verdadera distribución – véase Jarrow & Rudd (1982). Aún más, y a pesar de su aparente complejidad, la pdf Gram-Charlier es muy manejable debido a la ortogonalidad de los polinomios de Hermite, los cuales satisfacen, entre otras propiedades,

$$\int_{-\infty}^{\infty} H_s(x)H_j(x)\phi(x)dx = \begin{cases} 0, & s \neq j \\ s!, & s = j \end{cases}. \quad \text{(III.2)}$$

Por ejemplo, la función de distribución acumulada (cdf) y la función generadora de momentos (mgf) se pueden calcular (respectivamente) como

$$\int_{-\infty}^{a} f(x_i, \boldsymbol{d_i})dx_i = \int_{-\infty}^{a} \phi(x_i)dx_i - \phi(a)\sum_{s=1}^{n} d_{is}H_{s-1}(a) \quad \text{(III.3)}$$

y

$$\int_{-\infty}^{\infty} e^{tx_i} f(x_i, \boldsymbol{d_i})dx_i = e^{t^2/2}\sum_{s=0}^{n} d_{is}t^s. \quad \text{(III.4)}$$

Por lo tanto, se puede comprobar fácilmente que el momento par (impar) de orden *k* depende de los *k* primeros parámetros pares (impares). Por ejemplo si d_1=0 entonces la densidad tiene media cero y d_3 recoge la asimetría y si d_2=0 entonces la densidad tiene varianza unitaria y d_4 captura el exceso de curtosis.

Proposición III.2.1: *Sea z_i distribuida log-SNP con parámetros de localización, escala y forma $\mu_i \in \mathbb{R}$, ${\sigma_i}^2 \in \mathbb{R}^+$ y $\boldsymbol{d_i} = (d_{i1}, \dots, d_{in})' \in \mathbb{R}^n$, respectivamente. Entonces su pdf puede expresarse como*

$$h(z_i;\ \mu_i, {\sigma_i}^2, \boldsymbol{d_i}) = \left[1 + \sum_{s=1}^{n} d_{is}H_s\left(\frac{\ln(z_i)-\mu_i}{\sigma_i}\right)\right]\left(\frac{1}{z_i\sigma_i\sqrt{2\pi}}e^{-\frac{(\ln(z_i)-\mu_i)^2}{2{\sigma_i}^2}}\right), \quad z_i \in \mathbb{R}^+. \quad \text{(III.5)}$$

Por lo tanto, la log-SNP es la transformación exponencial de una variable distribuida Gram-Charlier – es decir $z_i = \exp(x_i)$, donde x_i se distribuye de acuerdo con la pdf (III.1) –, por lo tanto, la distribución lognormal es un caso particular (para $\boldsymbol{d_i}$ =**0**). La densidad resultante presenta la misma flexibilidad de parámetros que la Gram-Charlier, pero se define sólo en el eje real positivo. Las propiedades de esta distribución se pueden obtener fácilmente de las de la Gram-Charlier – véase Ñíguez, Paya, Peel, & Perote (2012) y Ñíguez, Paya, Peel, & Perote (2013) para más detalles. Particularmente, los momentos centrales se pueden obtener directamente de la mgf de la distribución de Gram-Charlier – ecuación (III.4) – como

$$E[z_i{}^t] = e^{\mu_i t + \frac{1}{2}t^2 {\sigma_i}^2}[1 + \sum_{s=1}^{n} d_{is}(\sigma_i t)^s]. \quad \text{(III.6)}$$

Tanto la Gram-Charlier como la log-SNP pueden extenderse al caso multivariante de diferentes maneras. En este trabajo se define la pdf log-SNP multivariante en términos de la densidad multivariante de Edgeworth-Sargan definida por Perote (2004). En lo que sigue y sin pérdida de generalidad se describe el caso bivariante, el cual se aplica en la siguiente sección.

Proposición III.2.2: *Sea* $\boldsymbol{Z} = (z_1, z_2)' \in \mathbb{R}^{2+}$ *un vector aleatorio distribuido como log-SNP bivariante con media* $\boldsymbol{\mu} = (\mu_1, \mu_2)' \in \mathbb{R}^2$ *y matriz de varianzas-covarianzas (definida positiva)* $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, $\sigma_i > 0$, i=1,2, *y* $|\rho| < 1$. *Por lo tanto, la densidad conjunta de* $\boldsymbol{Z}$ *se describe por*

$$H(\boldsymbol{Z};\ \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{D}) = F(\boldsymbol{Z};\ \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{z_1 z_2 \sigma_1 \sigma_2}\phi\left(\frac{\ln(z_1)-\mu_1}{\sigma_1}\right)\phi\left(\frac{\ln(z_2)-\mu_2}{\sigma_2}\right)\left[p_1\left(\frac{\ln(z_1)-\mu_1}{\sigma_1}\right) + p_2\left(\frac{\ln(z_2)-\mu_2}{\sigma_2}\right)\right],$$

$$z_i \in \mathbb{R}^+, i = 1,2, \boldsymbol{D} = (\boldsymbol{d_1} \quad \boldsymbol{d_2})' \in \mathbb{R}^2, \quad \text{(III.7)}$$

donde $F(\boldsymbol{Z};\ \boldsymbol{\mu}, \boldsymbol{\Sigma})$ *es una distribución lognormal multivariante – Aitchison & Brown (1957)–*

$$F(\boldsymbol{Z};\ \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi x_1 x_2 \sigma_1 \sigma_2 \sqrt{1-\rho}} exp\left(-\frac{q}{2}\right), \quad \text{(III.8)}$$

$$q = \frac{1}{1-\rho^2}\left[\left(\frac{\ln(z_1)-\mu_1}{\sigma_1}\right)^2 + \left(\frac{\ln(z_2)-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{\ln(z_1)-\mu_1}{\sigma_1}\right)\left(\frac{\ln(z_2)-\mu_2}{\sigma_2}\right)\right], \quad \text{(III.9)}$$

$\phi(x_i)$ *la normal estándar y* $p_i(x_i) = \sum_{s=0}^{n} d_{is} H_s\left(\frac{\ln(z_i)-\mu_i}{\sigma_i}\right)$, i=1,2.

Las densidades marginales de esta log-SNP multivariante se distribuyen como log-SNP univariantes –ecuación (III.5) – y por lo tanto las aplicaciones empíricas son muy manejables ya que los parámetros estimados de las marginales pueden utilizarse como valores iniciales para los procedimientos de estimación conjunta por máxima verosimilitud (ML). También, la relación entre los momentos (de la muestra) y los parámetros de densidad pueden explotarse para este propósito.

III.3. Distribución del tamaño de la empresa

III.3.1. Descripción de los datos y estadísticas

Este estudio se lleva a cabo sobre un conjunto de empresas estadounidenses durante el período comprendido entre 2004 y 2015. El análisis se basa en información financiera del conjunto de empresas disponibles en la base de datos de Thomson Reuters Datastream. Esta es una base de datos global y recopila series temporales sobre las cuentas de empresas que cotizan en bolsa.

Sobre la información disponible se aplicaron los siguientes criterios de selección para obtener la muestra final utilizada para la estimación: *(i)* Se excluyeron de la muestra las empresas que no tenían información contable durante el período de revisión; *(ii)* Sólo se tomaron en cuenta las empresas activas (activos totales y resultados operativos positivos) en este período; *(iii)* Las empresas sobre las cuales no se tuviera disponible su código SIC (Código de Clasificación Industrial), también se excluyeron. Dados estos criterios, se alcanzó una muestra total de N=2349 empresas por año.[16]

Adicionalmente, con el fin de analizar la distribución del tamaño de la empresa por grupos de acuerdo con la actividad económica, las empresas se dividieron así: Manufacturing (códigos SIC 20-39), Non-manufacturing (códigos SIC 10-14, 15-17, 40-49, 50-51, 52-59 y 70-89), Finance, Insurance & Real Estate (códigos SIC 60-67) y Economy-wide, que corresponde a la unión de los tres grupos anteriores. Finalmente, la variable utilizada para medir el tamaño de la empresa fue el valor de las ventas medido en dólares (USD).

La Tabla III.1 muestra las estadísticas descriptivas para cada uno de los cuatro grupos de industrias organizados en el presente estudio. En la tabla se observa el comportamiento temporal de los momentos de la distribución, calculadas hasta el cuarto orden.

[16] Nótese que en esta investigación no se controlan cambios en la forma de la distribución de las empresas por entrada y salida de empresas, ni por fusiones y adquisiciones (M&A). Al respecto, Cefis, Marsili, & Schenk (2009) muestran que las M&A no afectan el tamaño de la distribución cuando consideran la población entera de empresas. Ese resultado se puede deber al efecto generado por la entrada y salida de empresas y que contrarresta el efecto de las M&A. Sin embargo, esto se presenta como una limitación en el presente trabajo.

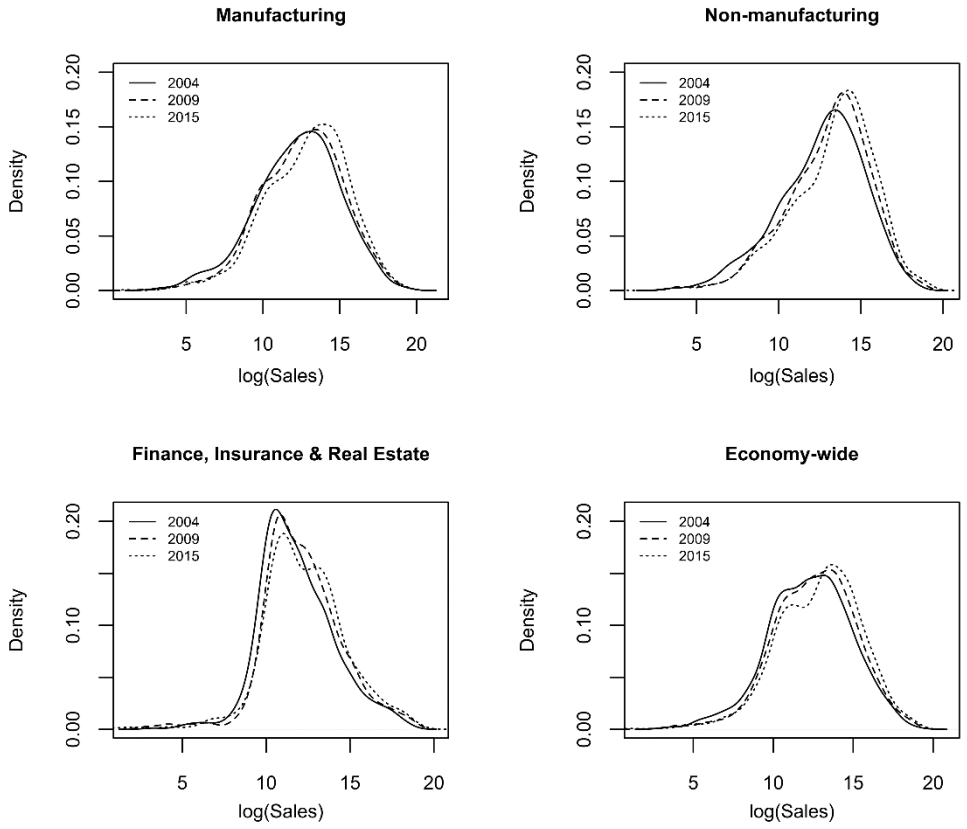
Tabla III.1 Estadísticas descriptivas

Industria	Estadístico	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Manufacturing (N=947)	Min $x10^6$	0.019	0.023	0.007	0.008	0.046	0.008	0.008	0.008	0.002	0.004	0.030	0.003
	Media $x10^9$	3.38	3.80	4.06	4.35	4.79	3.94	4.44	5.04	4.99	4.96	4.99	4.55
	Max $x10^9$	263.99	328.21	335.09	358.60	425.07	275.56	341.58	433.53	420.71	390.25	364.76	236.81
	Desv. $x10^{10}$	1.45	1.71	1.76	1.86	2.15	1.55	1.81	2.18	2.06	1.99	1.93	1.61
	Sesgo	10.90	11.68	11.39	11.49	12.16	9.85	10.71	11.83	11.93	11.18	10.56	8.92
	Curtosis	152.15	175.94	169.26	173.18	191.72	130.63	156.48	188.62	198.25	172.76	154.41	103.90
Non-manufacturing (N=784)	Min $x10^6$	0.025	0.035	0.046	0.021	0.034	0.048	0.050	0.007	0.005	0.002	0.009	0.001
	Media $x10^9$	2.76	3.08	3.37	3.70	4.04	3.73	4.01	4.45	4.69	4.91	5.20	5.12
	Max $x10^9$	96.29	91.13	91.42	118.93	124.03	123.02	124.28	126.72	127.43	128.75	139.37	179.05
	Desv. $x10^{10}$	0.79	0.84	0.92	1.06	1.14	1.14	1.18	1.28	1.37	1.40	1.49	1.60
	Sesgo	6.32	6.00	6.03	6.47	6.27	6.79	6.42	6.06	6.05	5.89	6.00	6.81
	Curtosis	51.25	45.00	44.41	50.61	47.36	53.89	48.49	43.18	42.36	40.15	41.87	53.91
Finance, Insurance & Real Estate (N=618)	Min $x10^6$	0.026	0.021	0.013	0.011	0.011	0.017	0.008	0.009	0.012	0.003	0.016	0.003
	Media $x10^9$	2.11	2.44	2.88	3.19	2.91	3.09	3.39	3.41	3.53	3.54	3.61	3.72
	Max $x10^9$	108.28	120.28	146.56	159.23	112.45	143.27	155.70	146.70	160.84	179.54	194.17	209.85
	Desv. $x10^{10}$	0.90	1.04	1.27	1.41	1.18	1.32	1.50	1.46	1.48	1.47	1.50	1.57
	Sesgo	7.54	7.32	7.31	7.29	6.62	7.01	6.98	6.70	6.58	7.03	7.30	7.72
	Curtosis	68.07	62.53	60.62	59.55	48.53	54.74	53.50	49.57	48.72	58.65	65.28	74.26
Economy-wide (N=2349)	Min $x10^6$	0.019	0.021	0.007	0.008	0.011	0.008	0.008	0.007	0.002	0.002	0.009	0.001
	Media $x10^9$	2.84	3.20	3.52	3.83	4.04	3.65	4.02	4.41	4.51	4.57	4.70	4.52
	Max $x10^9$	263.99	328.21	335.09	358.60	425.07	275.56	341.58	433.53	420.71	390.25	364.76	236.81
	Desv. $x10^{10}$	1.13	1.30	1.40	1.51	1.64	1.37	1.54	1.74	1.70	1.68	1.68	1.60
	Sesgo	11.16	12.13	11.18	10.89	12.59	8.81	9.62	11.25	10.64	9.87	9.13	7.89
	Curtosis	182.20	219.63	185.53	175.95	241.23	108.94	137.87	199.19	183.50	152.72	125.97	79.39

La tabla muestra las estadísticas descriptivas para cada uno de los grupos por actividad económica económicos. N= número de empresas. Los valores presentados corresponden al valor de las ventas en dólares (USD). Min=valor mínimo de las ventas, Max=valor mínimo de las ventas, Media=valor medio de las ventas, Desv=desviación estándar del valor de las ventas, Sesgo y Curtosis corresponden al coeficiente de asimetría y al exceso de curtosis respectivamente.

En particular, los momentos centrales tercero y cuarto proporcionan información útil sobre la forma de la distribución, adicional a la media y la desviación estándar. En general, la distribución del tamaño de la empresa muestra asimetría positiva, con una presencia muy alta de pequeñas empresas. El exceso de curtosis positivo también muestra que la cola superior de la distribución es más pesada respecto a la observada en una distribución lognormal.

Figura III.1 Densidad empírica del logaritmo de las ventas



La figura muestra la densidad del logaritmo de la variable ventas (log(Sales)) resultante de una suavización del histograma correspondiente. Para cada uno de los cuatro grupos de industrias, los años 2004 (línea sólida), 2009 (línea discontinua) y 2015 (línea punteada) se seleccionaron aleatoriamente.

En el largo plazo, para los grupos de empresas Non-manufacturing y Finance, Insurance & Real Estate, los cuatro momentos muestran un ligero aumento. Sin embargo, en la industria

Manufacturing la tendencia se invierte a partir del año 2013. En general, si se observa Economy-wide en el período de tiempo del estudio, la forma de la distribución del tamaño de la empresa se vuelve menos dispersa alrededor de la media, menos sesgada hacia las pequeñas empresas y menos pesada en las colas.

Las gráficas de densidad, del logaritmo de la variable ventas, resultante de una suavización del histograma correspondiente se presentan en la Figura III.1. Con el fin de obtener una mejor visualización de las densidades, los años 2004, 2009 y 2015 fueron seleccionados aleatoriamente. Para cada uno de los cuatro grupos de industrias se puede observar el comportamiento arriba descrito. Adicionalmente, se muestra que las distribuciones del tamaño de la empresa tienen una forma diferente a la lognormal y además son bimodales o incluso multimodales, tal como en los resultados obtenidos por Marsili (2006) y Bottazzi, Cefis, Dosi, & Secchi (2007).

III.3.2. Resultados y discusión

Las Tablas III.2 a III.5 presentan la estimación ML para cada uno de los cuatro grupos de industrias seleccionados. El Panel A muestra los parámetros estimados para la distribución lognormal y el Panel B muestra los parámetros estimados para la distribución log-SNP. En el Panel C se encuentra el estadístico de razón de verosimilitudes (LR) para el contraste de log-SNP frente a lognormal.

Los resultados de la estimación muestran que todos los modelos capturan de manera adecuada la media y la desviación estándar de cada uno de los grupos de industrias. Estos estadísticos se representan por los parámetros de localización μ y de escala σ respectivamente. Como se observa, los p-valores indican que estos parámetros son altamente significativos para ambas distribuciones. Sin embargo, como se muestra en el Panel B, para la distribución log-SNP, los parámetros d_s también resultan altamente significativos para la mayoría de los años y de las industrias. Al analizar el Criterio de Información Akaike (AIC, que penaliza la inclusión de parámetros adicionales) para las dos distribuciones, se encuentra que este criterio resulta ser consistentemente inferior para la distribución log-SNP, lo que sugiere que la modelización a partir de esta distribución resulta claramente superior.

Tabla III.2 Estimaciones para la distribución del tamaño de la empresa bajo lognormal y log-SNP, Industria Manufacturing

Año	Panel A lognormal					Panel B log-SNP									Panel C LR
	μ	σ	logL	AIC	KS test	μ	σ	d_1	d_2	d_3	d_4	logL	AIC	KS test	
2004	5.3598	2.7296	-7370.44	14744.87	(0.0231)	3.9696	2.1032	0.6610	0.5607	0.1308	0.0673	-7357.01	14726.02	(0.6906)	26.86
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0030)	(0.0023)			Not rejected*	(<.0001)
2005	5.4834	2.7026	-7478.08	14960.15	(0.1427)	3.9566	2.1056	0.7251	0.5866	0.1631	0.0690	-7466.14	14944.28	(0.4308)	23.87
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0003)	(0.0011)			Not rejected*	(<.0001)
2006	5.6136	2.6747	-7591.56	15187.13	(0.1754)	3.9653	2.1235	0.7763	0.5946	0.1842	0.0633	-7580.17	15172.34	(0.2831)	22.79
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0022)			Not rejected*	(0.0001)
2007	5.7088	2.6693	-7679.77	15363.55	(0.2584)	3.9904	2.1016	0.8177	0.6409	0.2095	0.0789	-7666.27	15344.55	(0.8677)	27.00
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2008	5.7688	2.6581	-7732.62	15469.24	(0.3981)	4.1036	2.0983	0.7936	0.6173	0.2024	0.0735	-7721.31	15454.63	(0.8022)	22.61
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(0.0002)
2009	5.6448	2.6342	-7606.57	15217.14	(0.1584)	3.8858	2.1308	0.8255	0.6049	0.2011	0.0713	-7595.21	15202.42	(0.1939)	22.72
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(0.0001)
2010	5.7446	2.6509	-7707.10	15418.19	(0.2138)	3.7281	2.2052	0.9144	0.6406	0.2235	0.0708	-7693.91	15399.83	(0.9740)	26.37
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2011	5.8257	2.7323	-7812.53	15629.07	(0.3981)	3.6698	2.2339	0.9651	0.7137	0.2389	0.0836	-7790.51	15593.02	(0.8962)	44.05
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2012	5.8615	2.7386	-7848.61	15701.22	(0.1283)	3.7389	2.2005	0.9646	0.7397	0.2500	0.0922	-7824.27	15660.53	(0.2353)	48.69
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2013	5.8823	2.7532	-7873.40	15750.80	(0.0010)	3.6045	2.2394	1.0172	0.7731	0.2708	0.0927	-7845.51	15703.02	(0.6133)	55.78
	(<.0001)	(<.0001)			Rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2014	5.9594	2.6853	-7922.73	15849.45	(0.0131)	3.9886	2.0708	0.9517	0.7937	0.2806	0.1059	-7897.59	15807.17	(0.8962)	50.28
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2015	5.9591	2.6700	-7917.02	15838.05	(0.0820)	4.0037	2.0577	0.9503	0.7933	0.2729	0.1065	-7888.91	15789.83	(0.9740)	56.22
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)

Esta tabla recoge la estimación ML para la industria Manufacturing. El Panel A muestra los parámetros estimados para la distribución lognormal. El Panel B muestra los parámetros estimados para la distribución log-SNP. El Panel C muestra la razón de verosimilitudes aplicada entre ambas distribuciones. μ y σ son los parámetros de localización y escala y d_s los parámetros de forma. logL=log-verosimilitud, AIC=Criterio de Información Akaike, LR=razón de verosimilitudes, KS test= test de Kolmogorov-Smirnov. P-valores en paréntesis. Not rejected* indica un mejor ajuste en el KS test.

Tabla III.3 Estimaciones para la distribución del tamaño de la empresa bajo lognormal y log-SNP, Industria Non-manufacturing

Año	Panel A lognormal					Panel B log-SNP									Panel C LR
	μ	σ	logL	AIC	KS test	μ	σ	d_1	d_2	d_3	d_4	logL	AIC	KS test	
2004	5.6920	2.5976	-6323.36	12650.72	(0.2340)	4.1361	1.9310	0.8058	0.7294	0.1884	0.0725	-6297.28	12606.57	(0.9171)	52.16
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0003)	(0.0055)			Not rejected*	(<.0001)
2005	5.8579	2.5606	-6442.18	12888.37	(0.1058)	4.4427	1.8864	0.7502	0.7027	0.1726	0.0616	-6416.19	12844.38	(0.8203)	51.99
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0073)	(0.0202)			Not rejected*	(<.0001)
2006	5.9917	2.5499	-6543.85	13091.70	(0.0317)	4.5243	1.8624	0.7880	0.7478	0.1902	0.0711	-6514.36	13040.71	(0.8886)	58.98
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0016)	(0.0054)			Not rejected*	(0.0001)
2007	6.1032	2.5241	-6623.24	13250.48	(0.0274)	4.3527	1.8995	0.9216	0.8075	0.2545	0.0873	-6593.65	13199.29	(0.9754)	59.19
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2008	6.2082	2.5122	-6701.89	13407.78	(0.0032)	4.1176	2.0048	1.0428	0.8289	0.2931	0.0881	-6671.73	13355.46	(0.6569)	60.32
	(<.0001)	(<.0001)			Rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(0.0002)
2009	6.1111	2.4789	-6615.24	13234.49	(0.0149)	3.8505	2.0423	1.1069	0.8493	0.3120	0.0858	-6586.47	13184.94	(0.6144)	57.55
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(0.0001)
2010	6.1958	2.4821	-6682.65	13369.30	(0.2106)	3.8191	2.0698	1.1482	0.8782	0.3224	0.0946	-6652.11	13316.22	(0.8561)	61.09
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2011	6.2690	2.5358	-6756.89	13517.79	(0.0203)	3.9474	2.0741	1.1194	0.8738	0.3087	0.0905	-6724.04	13460.09	(0.7817)	65.70
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2012	6.3150	2.5448	-6795.72	13595.44	(0.0108)	3.9355	2.0671	1.1511	0.9204	0.3346	0.1047	-6760.12	13532.24	(0.2866)	71.20
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2013	6.3541	2.5790	-6836.84	13677.67	(0.0027)	4.0593	2.0753	1.1058	0.8835	0.3096	0.0943	-6800.73	13613.45	(0.9859)	72.22
	(<.0001)	(<.0001)			Rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2014	6.4151	2.5903	-6888.11	13780.23	(0.0824)	3.9697	2.0955	1.1670	0.9449	0.3549	0.1093	-6846.59	13705.18	(0.7412)	83.05
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2015	6.3842	2.5857	-6862.43	13728.86	(0.0127)	3.9279	2.0970	1.1713	0.9461	0.3578	0.1037	-6820.71	13653.42	(0.2340)	83.44
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)

Esta tabla recoge la estimación ML para la industria Non-manufacturing. El Panel A muestra los parámetros estimados para la distribución lognormal. El Panel B muestra los parámetros estimados para la distribución log-SNP. El Panel C muestra la razón de verosimilitudes aplicada entre ambas distribuciones. μ y σ son los parámetros de localización y escala y d_s los parámetros de forma. logL=log-verosimilitud, AIC=Criterio de Información Akaike, LR=razón de verosimilitudes, KS test= test de Kolmogorov-Smirnov. P-valores en paréntesis. Not rejected* indica un mejor ajuste en el KS test.

Tabla III.4 Estimaciones para la distribución del tamaño de la empresa bajo lognormal y log-SNP, Industria Finance, Insurance and Real Estate

Año	Panel A lognormal					Panel B log-SNP									Panel C LR
	μ	σ	logL	AIC	KS test	μ	σ	d_1	d_2	d_3	d_4	logL	AIC	KS test	
2004	4.9348	2.2025	-4414.58	8833.15	(0.0561)	4.1876	1.9570	0.3818	0.2061	0.1077	0.0984	-4381.61	8775.22	(0.4603)	65.94
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0011)	(<.0001)			Not rejected*	(<.0001)
2005	5.0986	2.1826	-4510.20	9024.40	(0.0179)	4.3265	1.9581	0.3943	0.1990	0.0990	0.0988	-4475.77	8963.54	(0.3788)	68.86
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(0.0016)	(0.0030)	(<.0001)			Not rejected*	(<.0001)
2006	5.2375	2.2216	-4606.95	9217.91	(0.0872)	4.4807	2.0436	0.3703	0.1594	0.0661	0.0923	-4569.33	9150.66	(0.3788)	75.24
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(0.0103)	(0.0457)	(<.0001)			Not rejected*	(0.0001)
2007	5.3224	2.2742	-4673.91	9351.82	(0.0017)	4.4390	2.1017	0.4203	0.1738	0.0480	0.0936	-4631.00	9274.00	(0.4184)	85.82
	(<.0001)	(<.0001)			Rejected	(<.0001)	(<.0001)	(<.0001)	(0.0070)	(0.1664)	(<.0001)			Not rejected*	(<.0001)
2008	5.3157	2.2614	-4666.27	9336.53	(0.0179)	4.4822	2.0080	0.4151	0.2203	0.0704	0.1056	-4623.26	9258.52	(0.3416)	86.01
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0384)	(<.0001)			Not rejected*	(<.0001)
2009	5.2705	2.3032	-4649.67	9303.34	(0.0481)	4.3158	2.1107	0.4523	0.1977	0.0478	0.1013	-4602.54	9217.07	(0.8284)	94.26
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(0.0019)	(0.1714)	(<.0001)			Not rejected*	(<.0001)
2010	5.2595	2.3541	-4656.39	9316.77	(0.0481)	4.3900	2.1587	0.4028	0.1757	0.0391	0.0993	-4608.45	9228.90	(0.4603)	95.88
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(0.0049)	(0.2504)	(<.0001)			Not rejected*	(<.0001)
2011	5.2829	2.3531	-4670.56	9345.12	(0.1317)	4.2784	2.1454	0.4682	0.2111	0.0443	0.1078	-4617.88	9247.76	(0.1501)	105.36
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(0.0011)	(0.2125)	(<.0001)			Not rejected*	(<.0001)
2012	5.2923	2.3979	-4688.05	9380.10	(0.0072)	4.3261	2.1399	0.4515	0.2297	0.0510	0.1062	-4640.82	9293.64	(0.3788)	94.46
	(<.0001)	(<.0001)			Rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.1470)	(<.0001)			Not rejected*	(<.0001)
2013	5.3019	2.4297	-4702.08	9408.16	(0.0017)	4.3472	2.1709	0.4398	0.2230	0.0523	0.0989	-4660.23	9332.46	(0.1004)	83.70
	(<.0001)	(<.0001)			Rejected	(<.0001)	(<.0001)	(<.0001)	(0.0011)	(0.1352)	(<.0001)			Not rejected*	(<.0001)
2014	5.3788	2.3615	-4732.01	9468.03	(0.1931)	4.4620	2.0670	0.4435	0.2510	0.0766	0.1003	-4699.90	9411.80	(0.5970)	64.23
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0252)	(<.0001)			Not rejected*	(<.0001)
2015	5.4117	2.4111	-4765.25	9534.49	(0.2180)	4.5211	2.2107	0.4028	0.1759	0.0376	0.0759	-4740.59	9493.18	(0.4603)	49.31
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(0.0238)	(0.2618)	(<.0001)			Not rejected*	(<.0001)

Esta tabla recoge la estimación ML para la industria Finance, Insurance and Real Estate. El Panel A muestra los parámetros estimados para la distribución lognormal. El Panel B muestra los parámetros estimados para la distribución log-SNP. El Panel C muestra la razón de verosimilitudes aplicada entre ambas distribuciones. μ y σ son los parámetros de localización y escala y d_s los parámetros de forma. logL=log-verosimilitud, AIC=Criterio de Información Akaike, LR=razón de verosimilitudes, KS test= test de Kolmogorov-Smirnov. P-valores en paréntesis. Not rejected* indica un mejor ajuste en el KS test.

Tabla III.5 Estimaciones para la distribución del tamaño de la empresa bajo lognormal y log-SNP, Economy-wide

Año	Panel A lognormal					Panel B log-SNP									Panel C LR
	μ	σ	logL	AIC	KS test	μ	σ	d_1	d_2	d_3	d_4	logL	AIC	KS test	
2004	5.3589	2.5722	-18140.36	36284.72	(0.5641)	4.0772	2.0039	0.6396	0.5283	0.1438	0.0797	-18113.55	36239.11	(0.9327)	53.62
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2005	5.5072	2.5439	-18462.70	36929.39	(0.4937)	4.1626	2.0102	0.6689	0.5245	0.1555	0.0738	-18439.89	36891.78	(0.1942)	45.61
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2006	5.6409	2.5371	-18770.48	37544.97	(0.4491)	4.2123	2.0282	0.7044	0.5305	0.1562	0.0705	-18745.70	37503.40	(0.8044)	49.57
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2007	5.7388	2.5396	-19002.83	38009.66	(0.0740)	4.1594	2.0343	0.7764	0.5807	0.1797	0.0837	-18967.08	37946.15	(0.5167)	71.51
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2008	5.7963	2.5334	-19132.03	38268.06	(0.0239)	4.1958	2.0424	0.7836	0.5764	0.1858	0.0830	-19096.15	38204.31	(0.5167)	71.75
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2009	5.7019	2.5199	-18897.95	37799.91	(0.0686)	3.9814	2.0941	0.8216	0.5616	0.1768	0.0751	-18862.74	37737.47	(0.2629)	70.44
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2010	5.7676	2.5449	-19075.26	38154.51	(0.2949)	3.8747	2.1719	0.8715	0.5662	0.1761	0.0714	-19034.35	38080.70	(0.4711)	81.81
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2011	5.8309	2.5992	-19273.51	38551.01	(0.5402)	3.9387	2.1652	0.8739	0.6023	0.1826	0.0797	-19222.03	38456.07	(0.9746)	102.95
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2012	5.8631	2.6176	-19365.85	38735.70	(0.1142)	3.9475	2.1432	0.8938	0.6453	0.2047	0.0904	-19308.43	38628.86	(0.3669)	114.84
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2013	5.8871	2.6442	-19445.94	38895.87	(0.0364)	3.8808	2.1976	0.9130	0.6407	0.2026	0.0815	-19388.47	38788.93	(0.5167)	114.94
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2014	5.9587	2.6022	-19576.70	39157.39	(0.1602)	4.1042	2.0662	0.8976	0.6959	0.2405	0.1001	-19522.27	39056.55	(0.5167)	108.85
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)
2015	5.9570	2.6027	-19572.92	39149.84	(0.1823)	4.0172	2.1084	0.9200	0.6852	0.2290	0.0886	-19519.64	39051.28	(0.5882)	106.55
	(<.0001)	(<.0001)			Not rejected	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			Not rejected*	(<.0001)

Esta tabla recoge la estimación ML para Economy-wide. El Panel A muestra los parámetros estimados para la distribución lognormal. El Panel B muestra los parámetros estimados para la distribución log-SNP. El Panel C muestra la razón de verosimilitudes aplicada entre ambas distribuciones. μ y σ son los parámetros de localización y escala y d_s los parámetros de forma. logL=log-verosimilitud, AIC=Criterio de Información Akaike, LR=razón de verosimilitudes, KS test= test de Kolmogorov-Smirnov. P-valores en paréntesis. Not rejected* indica un mejor ajuste en el KS test.

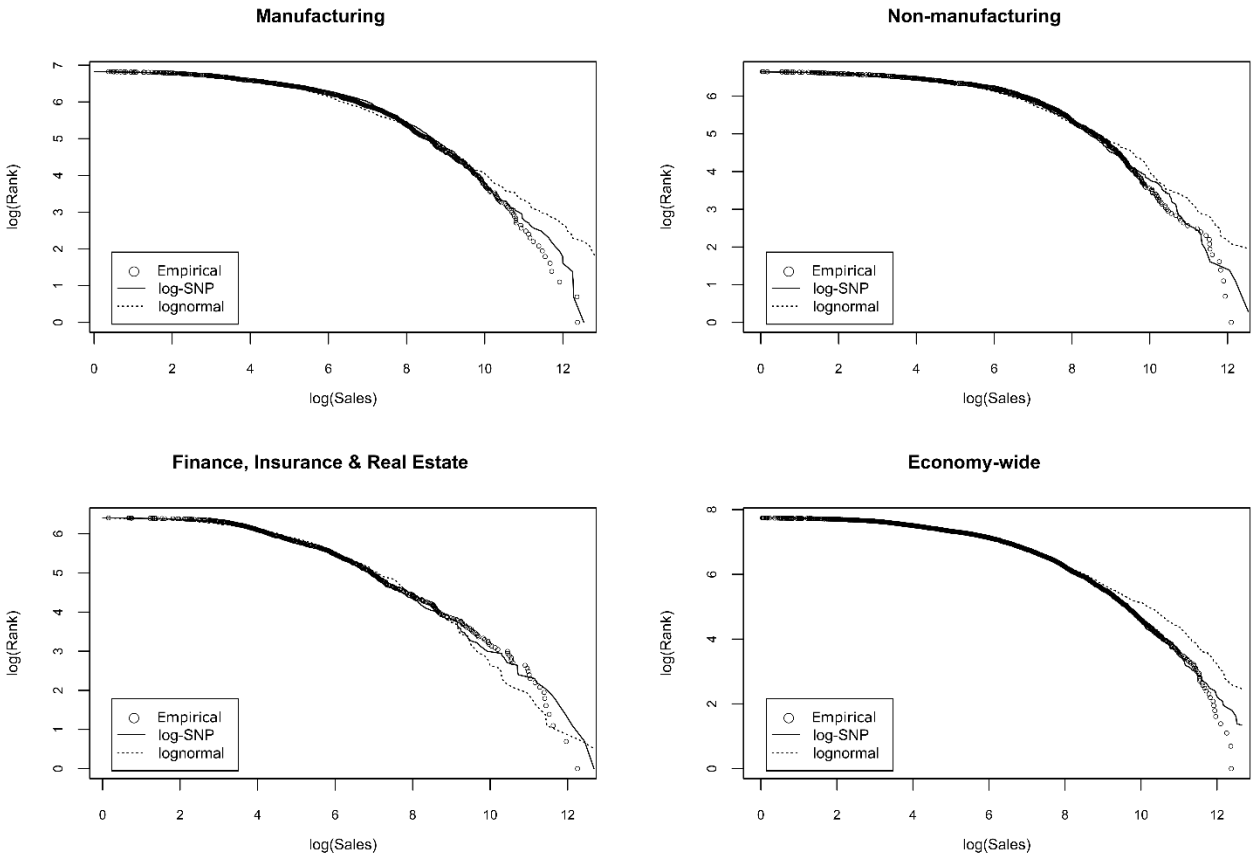
Además, los estadísticos LR incluidos en el Panel C, concluyen que, la mayoría de los años y de las industrias seleccionadas, la incorporación de los parámetros d_s mejoran la razón de verosimilitudes del modelo. Estos resultados son consistentes con el test de Kolmogorov-Smirnov (KS) aplicado a cada una de las distribuciones. Si se toma un nivel de significación de 0.01, en la mayoria de los años y a través de los cuatro grupos, el test no puede rechazar la hipótesis nula de que el proceso generador de los datos proviene de una distribución teórica lognormal o log-SNP. Sin embargo, en todos los años y en todas las industrias, la distribución log-SNP muestra un mejor ajuste. Obsérvese que a pesar de las diferencias en número de empresas y de la actividad económica que lleva a cabo cada uno de los cuatro grupos seleccionados, la forma de las distribuciones de tamaño de la empresa es similar.

Un ejemplo de la calidad del ajuste obtenido para cada una de las industrias en el año 2015 se recoge en la Figura III.2. La figura presenta, en escala logarítmica, la relación entre el rank y el tamaño de las ventas. Como se observa al comparar valores empíricos (puntos sin relleno) y los estimados teóricamente bajo una especificación lognormal (línea discontinua) y log-SNP (línea sólida), la distribución log-SNP recoge más adecuadamente no sólo los valores en torno a la media sino, especialmente, los valores extremos.[17]

Las Figura III.2 muestra como la distribución lognormal sobrestima los valores más extremos de la distribución. Estos resultados son consistentes con los obtenidos en investigaciones anteriores en las cuales la distribución lognormal subestima o sobreestima consistentemente los valores esperados teóricamente en la cola superior de la distribución del tamaño de la empresa (Stanley, et al., 1995; Hart & Oulton, 1997). La Tabla III.6 ilustra estos efectos para la industria Manufacturing calculando los cuantiles superiores empíricos y estimados bajo lognormal y log-SNP para niveles de confianza del 10, 5 y 1%.[18] Aquí se reportan los resultados para una sola industria, sin embargo se hizo el análisis para el resto de industrias y los resultados son cualitativamente similares.[19]

[17] El comportamiento para el resto de años es similar.
[18] Para obtener los cuantiles de la distribución log-SNP se usó la cdf presentada en la ecuación (III.3) y se aplicó el Método de la Transformada Inversa (ITM).
[19] Los resultados para las otras industrias se encuentran disponibles bajo requerimiento a los autores.

Figura III.2 Logaritmo del rank de la empresa vs logaritmo de las ventas de la empresa



La figura compara los valores empíricos (puntos sin relleno) y los estimados teóricamente bajo una especificación lognormal (línea discontinua) y log-SNP (línea sólida). Los ejes se encuentran en escala logarítmica y corresponden la relación entre el Rank y la variable ventas (log(Sales)). Para cada uno de los cuatro grupos de industrias se muestra el año 2015.

Analizando la tendencia de los valores de la cola superior de la distribución de las ventas en el periodo de estudio, se observa que la estructura paramétrica flexible de la distribución log-SNP permite ajustar mejor los valores esperados. La interpretación de los valores de la Tabla III.6 muestra claramente los errores inducidos en la estimación de la distribución del tamaño de la empresa mediante el uso de distribuciones paramétricas tradicionales como la lognormal.

Tabla III.6 Valor de las ventas observadas empíricamente versus valores esperados teóricamente bajo lognormal y log-SNP

Año	Valor de ventas observado (millones de dólares)			Valor de ventas esperado (millones de dólares)					
				Lognormal			Log-SNP		
	10%	5%	1.0%	10%	5%	1.0%	10%	5%	1.0%
2004	6,134.60	14,356.17	51,974.00	7,030.40	18,952.33	121,772.43	6,014.81	13,237.28	54,958.85
2005	6,603.66	15,043.38	55,868.84	7,684.46	20,512.97	129,392.69	6,598.52	14,322.48	58,345.11
2006	7,433.20	18,030.20	60,788.02	8,446.12	22,319.20	138,139.58	7,251.54	15,638.38	63,459.94
2007	7,750.20	19,133.65	60,310.24	9,224.81	24,328.55	150,016.27	7,887.45	16,788.62	66,459.65
2008	8,076.61	20,111.00	67,024.26	9,655.94	25,362.41	155,204.98	8,394.62	17,921.06	71,209.05
2009	7,218.28	17,030.90	65,292.00	8,271.88	21,539.06	129,678.04	7,250.99	15,625.23	63,327.55
2010	7,856.00	18,692.20	66,107.04	9,338.44	24,464.64	148,982.21	8,011.79	17,432.64	72,667.96
2011	8,369.81	19,970.51	75,190.16	11,240.25	30,330.24	195,230.32	8,991.23	19,557.65	81,846.26
2012	8,673.40	20,145.28	75,039.96	11,744.70	31,764.51	205,347.60	9,282.22	19,935.42	81,395.02
2013	8,772.35	20,860.57	78,253.70	12,218.77	33,222.39	216,919.80	9,454.11	20,367.08	84,142.50
2014	9,512.72	20,191.80	78,041.34	12,097.21	32,089.90	200,043.57	9,406.80	19,227.65	71,621.03
2015	9,330.04	20,170.31	68,975.52	11,859.39	31,285.45	193,014.95	9,175.00	18,691.47	69,169.41

Esta tabla compara valor de las ventas, en millones de dólares, observado empíricamente en la industria Manufacturing versus el esperado teóricamente bajo las distribuciones lognormal y log-SNP. Los valores 10%, 5% y 1%, 0.1% son percentiles de las distribuciones.

III.3.3. La distribución log-SNP bivariante: Ventas vs. Activos

El tamaño de la empresa puede medirse por diferentes variables: ventas, activos, empleados o beneficios, entre otros (Delmar, Davidsson, & Gartner, 2003; Zhang , Chen, & Wang, 2009; Heinrich & Dai, 2016). Esta diversidad de medidas sugiere que no hay un indicador de tamaño universalmente aceptado y la elección depende principalmente de la disponibilidad de los datos (Barbosa & Eiriz, 2011). Con el fin de dar robustez a los resultados anteriormente obtenidos en este estudio, se tomó como medida adicional de tamaño el valor de los activos totales de la empresa en la industria Manufacturing. La Tabla III.7 recoge las estadísticas descriptivas de esa variable para todo el periodo estudiado.

Como se observa en la tabla, la distribución del tamaño de la empresa medido a partir de la variable activos muestra asimetría positiva, con la presencia de una cantidad muy alta de pequeñas empresas y un número bajo de grandes empresas. La forma de la distribución que sugieren estas estadísticas es consistente con la observada con la variable ventas. Sin embargo, el exceso de curtosis también muestra que la cola superior de la distribución es aún más pesada que la exhibida en la variable ventas.

Tabla III.7 Estadísticas descriptivas, activos totales

Industria	Estadístico	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Manufacturing (N=947)	Min $x10^6$	0.145	0.174	0.153	0.050	0.050	0.030	0.034	0.023	0.029	0.020	0.005	0.010
	Media $x10^9$	4.26	4.35	4.77	5.20	5.11	5.35	5.71	6.03	6.32	6.61	6.78	6.94
	Max $x10^9$	750.33	673.34	697.24	795.34	797.77	781.82	751.22	717.24	685.33	656.29	645.81	508.14
	Desv. $x10^{10}$	2.85	2.64	2.79	3.10	3.03	3.07	3.07	3.07	3.05	3.06	3.07	2.88
	Sesgo	20.73	19.10	18.20	19.26	20.12	18.63	17.08	15.62	14.44	13.35	12.92	10.18
	Curtosis	512.32	449.59	415.24	459.05	499.00	440.85	379.15	323.04	280.09	240.54	225.10	135.90

La tabla muestra las estadísticas descriptivas para la industria Manufacturing. N= número de empresas. Los valores presentados corresponden al valor de los activos totales en dólares (USD). Min=valor mínimo de los activos, Max=valor mínimo de los activos, Media=valor medio de los activos, Desv=desviación estándar del valor de los activos, Sesgo y Curtosis corresponden al coeficiente de asimetría y al exceso de curtosis respectivamente.

La Tabla III.8 incluye un ejemplo de los resultados de la estimación de la distribución conjunta del valor de los activos y de las ventas de la empresa para los años 2008 y 2009. Estos años son relevantes, ya que como se observa en las Tablas III.1 y III.7, en ellos se marca un cambio estructural en la tendencia de los momentos de la distribución debido a la crisis financiera mundial. Específicamente, se estimó por ML los parámetros del caso bivariante de las densidades, de la distribución lognormal y log-SNP, descritas en las secciones anteriores. Se implementó la estimación de manera secuencial comenzando con la densidad univariante más simple, la lognormal, y recursivamente se agregaron parámetros utilizando como valores iniciales para los modelos más complejos las estimaciones previas de los más sencillos.

En cuanto a los resultados, el Panel A de la Tabla III.8 muestra los parámetros estimados para la distribución lognormal y el Panel B muestra los parámetros estimados para la distribución log-SNP. Como se observa, solo se estimaron los parámetros d_{3i} y d_{4i} (i=1 activos, i=2 ventas) reforzando el hecho de que las densidades necesitan ser expandidas mediante polinomios de grado más alto para capturar la masa probabilística en el rango extremo de las colas. Nótese que para ambas distribuciones todos los parámetros resultan estadísticamente significativos. Respecto a la correlación estimada, ρ, también es significativa y aunque este parámetro no captura exactamente la correlación entre ambas variables, sí se observa una dependencia muy alta entre las mismas. Al comparar el Criterio de Información Bayesiano (BIC, que penaliza la inclusión de parámetros adicionales) para

las dos distribuciones, se encuentra que este criterio resulta consistentemente inferior para la distribución log-SNP bivariante. Esto sugiere que, tal como en el caso univariante, la modelización a partir de esta distribución resulta claramente superior.

Tabla III.8 Estimaciones caso bivariante activos-ventas

	Panel A lognormal		**Panel B log-SNP**	
	2008	**2009**	**2008**	**2009**
μ_1	5.8310	5.7997	5.2752	5.6858
	(<.0001)	(<.0001)	(<.0001)	(<.0001)
σ_1	2.5085	2.5385	3.3442	2.8759
	(<.0001)	(<.0001)	(<.0001)	(<.0001)
d_{31}			-2.1829	0.4319
			(<.0001)	(<.0001)
d_{41}			1.3327	0.4059
			(<.0001)	(<.0001)
μ_2	5.7685	5.6448	5.3791	5.6201
	(<.0001)	(<.0001)	(<.0001)	(<.0001)
σ_2	2.6589	2.6342	3.1727	2.8036
	(<.0001)	(<.0001)	(<.0001)	(<.0001)
d_{32}			2.3255	0.5729
			(<.0001)	(<.0001)
d_{42}			-0.1443	-0.1569
			(0.0210)	(<.0001)
ρ	0.9539	0.9547	0.9890	0.9871
	(<.0001)	(<.0001)	(<.0001)	(<.0001)
logL	-3343.27	-3339.48	-2832.68	-2983.98
BIC	3360.41	3356.61	2863.52	3014.82

Esta tabla muestra la estimación ML para la industria Manufacturing. El Panel A muestra los parámetros estimados para la distribución lognormal bivariante. El Panel B muestra los parámetros estimados para la distribución log-SNP bivariante. μ_i y σ_i son los parámetros de localización y escala, d_{3i} y d_{4i} los parámetros de forma y ρ=correlación estimada. i=1 activos, i=2 ventas. logL=log-verosimilitud, BIC=Criterio de Información Bayesiano. P-valores en paréntesis.

Adicionalmente, se realizó el Test de Wald con el fin de analizar la relación entre los parámetros estimados en la distribución log-SNP. La Tabla III.9 incorpora los resultados obtenidos. Obsérvese que en general, para los dos años seleccionados se rechaza la hipótesis nula de igualdad entre los valores de los coeficientes de la estimación, lo que indica que a pesar de que las series están altamente correlacionadas pueden encontrarse diferencias

significativas en cuanto al comportamiento de los valores extremos. Sin embargo, para el año 2009 la diferencia entre los parámetros d_{3i} no resulta estadísticamente distinta de cero.

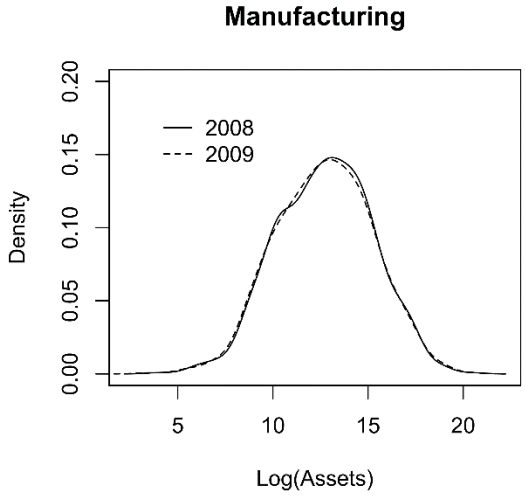
Tabla III.9 Test de Wald, distribución log-SNP

	2008	**2009**
μ_1-μ_2	-0.1039 (<.0001)	0.0657 (<.0001)
σ_1-σ_2	0.1714 (<.0001)	0.0723 (<.0001)
d_{31}-d_{32}	-4.5084 (<.0001)	-0.1411 (0.2110)
d_{41}-d_{42}	1.4770 (<.0001)	0.5628 (<.0001)
Chi Sq(4)	5049.63 (<.0001)	203.42 (<.0001)

La tabla muestra el test de Wald para la distribución log-SNP bivariante. μ_i y σ_i son los parámetros de localización y escala, d_{3i} y d_{4i} son los parámetros de forma. *i*=1 activos, *i*=2 ventas. P-valores en paréntesis.

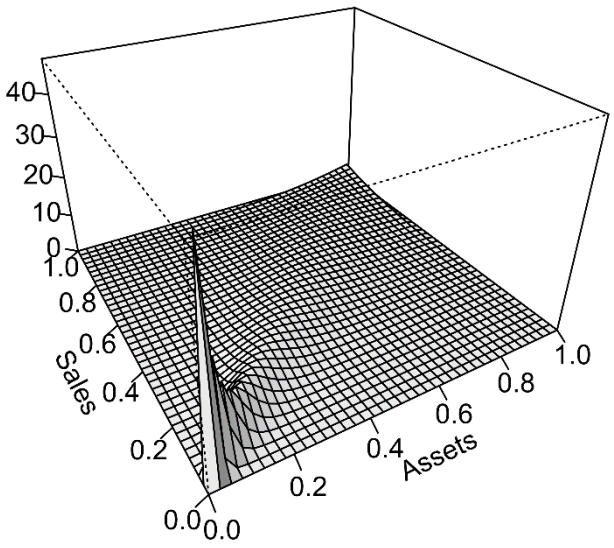
Como se muestra en la Figura III.3, en el año 2009 el valor de los activos, medido en logaritmos, se volvió más simétrico respecto al año anterior, lo que también resultó en una menor diferencia entre los parámetros de localización μ_i y de escala σ_i. De acuerdo con Pascoal, Augusto, & Monteiro (2016), una posible explicación para este comportamiento es que los activos son un mecanismo para la supervivencia en momentos de inestabilidad económica. Adicionalmente, una reducción de ventas, en presencia de grandes costos fijos, puede conducir a grandes pérdidas que resultan inmediatamente en la reducción de activos (que pueden actuar como un seguro en situaciones económicas adversas), lo que puede dar lugar a nuevas pérdidas en las ventas. En consecuencia, esto también puede implicar una reducción en la diferencia entre el tamaño de las empresas grandes y pequeñas. De esa manera, el valor de las ventas parece ser una mejor medida del tamaño de la empresa tal como lo indica la literatura al compararlas con otras medidas (Gaffeo, Gallegati, & Palestrini, 2003; Pascoal, Augusto, & Monteiro, 2016).

Figura III.3 Densidad empírica del logaritmo de los activos



La figura muestra la densidad del logaritmo de la variable valor de activos totales (log(Assets)) resultante de una suavización del histograma correspondiente para los años 2008 (línea sólida) y 2009 (línea discontinua) en la industria Manufacturing.

Figura III.4 Histograma distribución conjunta activos-ventas



La figura muestra el histograma de la distribución conjunta del valor de los activos y de las ventas de la empresa para el año 2009 en la industria Manufacturing.

Esos efectos también se pueden evidenciar en la Figura III.4. Esta figura muestra el histograma de la distribución conjunta del valor de los activos y de las ventas de la empresa para el año 2009. Normalmente las empresas que tienen valores de ventas bajos, tienen

valores de activos bajos (y viceversa). Pero pueden existir algunas empresas con valores de ventas altos y valores de activos con un nivel, relativo a esas ventas, inferior.

III.4. Conclusiones

Este capítulo aplica una novedosa metodología basada en la distribución log-SNP para la modelización de la distribución del tamaño de la empresa. Dicha distribución anida la lognormal incluyendo nuevos parámetros capaces de recoger mejor el comportamiento de la cola superior del tamaño de la empresa y permite contrastar las deficiencias de la distribución lognormal en esta dirección. El análisis empírico compara el desempeño de ambas distribuciones ajustando una muestra de empresas Estadounidenses en el periodo de 2004 a 2015.

Los resultados muestran que la distribución lognormal tiende a sobreestimar consistentemente los valores esperados en la cola superior de la distribución. Ese resultado pone de manifiesto la necesidad de proponer otras distribuciones que permitan obtener información más confiable sobre el grado de concentración industrial, los ciclos económicos y así poder implementar políticas de competencia. Tomando diferentes niveles de agregación por actividad económica, el presente estudio muestra que la log-SNP provee un mejor ajuste de la distribución del tamaño de la empresa. También, muestra ser más flexible que la lognormal cuando los datos son muy sesgados y hay posibles saltos en la cola superior debido a las observaciones extremas.

También, se desarrolla por primera vez una expresión para la densidad de la distribución log-SNP multivariante y se analiza la estimación de la distribución conjunta del valor de los activos y de las ventas de la empresa. Los resultados sugieren que las ventas son una mejor medida del tamaño de la empresa, tal como lo han obtenido otros estudios en la literatura. A pesar de la alta correlación entre del valor de los activos y de las ventas de la empresa, en periodos de crisis financiera, los activos pueden actuar como un seguro de supervivencia ante entornos económicos inestables. Ese hecho puede llevar a obtener conclusiones distorsionadas en el análisis del comportamiento la distribución del tamaño de la empresa a partir de esa variable.

Referencias

Aitchison, J., & Brown, J. (1957). *The Lognormal Distribution.* Cambridge: Cambridge University Press.

Axtell, R. (2001). Zipf distribution of U.S. Firm sizes. *Science, 293*, 1818–1820.

Barba Navaretti, G., Castellani, D., & Pieri, F. (2014). Age and firm growth: evidence from three European countries. *Small Business Economics, 43*(4), 823–837.

Barbosa, N., & Eiriz, V. (2011). Regional variation of firm size and growth: the Portuguese case. *Growth and Change, 42*(2), 125-158.

Blinnikov, S., & Moessner, R. (1998). Expansions for nearly Gaussian distributions. *Astronomy and astrophysics Supplement Series, 130*(1), 193–205.

Bottazzi, G., Cefis, E., Dosi, G., & Secchi, A. (2007). Invariances and diversities in the patterns of industrial evolution: some evidence from Italian manufacturing industries. *Small Business Economics, 29*(1), 137-159.

Bottazzi, G., Pirino, D., & Tamagni, F. (2015). Zipf law and the firm size distribution: a critical discussion of popular estimators. *Journal of Evolutionary Economics, 25*(3), 585–610.

Cabral, L. M., & Mata, J. (2003). On the evolution of the firm size distribution: facts and theory. *American Economic Review, 93*(4), 1075-1090.

Cefis, E., Marsili, O., & Schenk, H. (2009). The effects of mergers and acquisitions on the firm size distribution. *Journal of Evolutionary Economics, 19*(1), 1-20.

Cirillo , P., & Hüsler, J. (2009). On the upper tail of Italian firms' size distribution. *Physica A: Statistical Mechanics and its Applications, 388*(8), 1546-1554.

Coad, A. (2010). The exponential age distribution and the Pareto firm size distribution. *Journal of Industry, Competition and Trade, 10*(3), 389–395.

Cortés, L. M., Mora-Valencia, A., & Perote, J. (2016). The productivity of top researchers: a semi-nonparametric approach. *Scientometrics, 109*(2), 891-915.

Crosato, L., & Ganugi, P. (2007). Statistical regularity of firm size distribution: the Pareto IV and truncated Yule for Italian SCI manufacturing. *Statistical Methods and Applications, 16*(1), 85-115.

Delmar, F., Davidsson, P., & Gartner, W. B. (2003). Arriving at the high-growth firm. *Journal of Business Venturing, 18*(2), 189-216.

di Giovanni, J., Levchenko, A. A., & Rancière. , R. (2011). Power laws in firm size and openness to trade: measurement and implications. *Journal of International Economics, 85*(1), 42-52.

Dosi, G., Marsili, O., Orsenigo, L., & Salvatore, R. (1995). Learning, marketselection and the evolution of industrial structures. *Small BusinessEconomics, 7*(6), 411–436.

Gaffeo, E., Gallegati, M., & Palestrini, A. (2003). On the size distribution of firms: additional evidence from the G7 countries. *Physica A: Statistical Mechanics and its Applications, 324*(1-2), 117-123.

Gibrat, R. (1931). *Les Inégalités Economiques.* Paris: Recueil Sirey.

Goddard, J., Liu, H., Donal, M., & Wilson, J. O. (2014). The size distribution of US banks and credit unions. *International Journal of the Economics of Business, 21*(1), 139-156.

Gupta, H. M., Campanha, J. R., de Aguiar, D. R., Queiroz, G. A., & Raheja, C. G. (2007). Gradually truncated log-normal in USA publicly traded firm size distribution. *Physica A: Statistical Mechanics and its Applications, 375*(2), 643-650. doi:http://dx.doi.org/10.1016/j.physa.2006.09.025

Hart, P. E., & Oulton, N. (1997). Zipf and the size distribution of firms. *Applied Economics Letters, 4*(4), 205-206.

Heinrich, T., & Dai, S. (2016). Diversity of firm sizes, complexity, and industry structure in the Chinese economy. *Structural Change and Economic Dynamics, 37*, 90-106.

Hernández-Pérez, R. (2010). An analogy of the size distribution of business firms with Bose–Einstein statistics. *Physica A: Statistical Mechanics and its Applications, 389*(18), 3837-3843.

Jarrow, R., & Rudd, A. (1982). Approximate option valuation for arbitrary stochastic processes. *Journal of Financial Economics, 10*(3), 347-369.

Jondeau, E., & Rockinger, M. (2001). Gram-Charlier densities. *Journal of Economic Dynamics & Control, 25*(10), 1457-1483.

Kaizoji, T., Iyetomi, H., & Ikeda, Y. (2006). Re-examination of the size distribution of firms. *Evolutionary and Institutional Economics Review, 2*(2), 183–198.

Kuhs, W. F. (1988). The anharmonic temperature factor in crystallographic structure analysis. *Australian Journal of Physics, 41*(3), 369-382.

Leon, A., Mencía, J., & Sentana, E. (2009). Parametric properties of semi-nonparametric distributions, with applications to option valuation. *Journal of Business and Economic Statistics, 27*(2), 176-192.

Marsili, O. (2006). Stability and turbulence in the size distribution of firms: evidence from Dutch manufacturing. *International Review of Applied Economics, 20*(2), 255-272.

Martínez-Mekler, G., Martínez, R. A., del Río, M. B., Mansilla, R., Miramontes, P., & Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE, 4*(3), e4791.

Mauleón, I., & Perote, J. (2000). Testing densities with financial data: an empirical comparison of the Edgeworth-Sargan density to the Student's t. *European Journal of Finance, 6*(2), 225-239.

Newman, M. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323-351.

Ñíguez, T-M., Paya, I., Peel, D., & Perote, J. (2012). On the stability of the constant relative risk aversion (CRRA) utility under high degrees of uncertainty. *Economics Letters, 115*(2), 244-248.

Ñíguez, T-M., Paya, I., Peel, D., & Perote, J. (2013). Higher-order moments in the theory of diversification and portfolio composition. *Economics Working Paper Series 2013/003. Lancaster University*.

Pascoal, R., Augusto, M., & Monteiro, A. M. (2016). Size distribution of Portuguese firms between 2006 and 2012. *Physica A: Statistical Mechanics and its Applications, 458*, 342-355.

Perote, J. (2004). The multivariate Edgeworth-Sargan density. *Spanish Economic Review, 6*(1), 77-96.

Simon, H., & Bonini, C. (1958). The size distribution of business firms. *The American Economic Review, 48*(4), 607-617.

Stanley, M. H., Buldyrev, S. V., Havlin, S., Mantegna, R. N., Salinger, M. A., & Stanley, H. E. (1995). Zipf plots and the size distribution of firm. *Economics Letters, 49*(4), 453-457.

Voit, J. (2001). The growth dynamics of German business firms. *Advances in Complex Systems, 4*(1), 149-162.

Zhang , J., Chen, Q., & Wang, Y. (2009). Zipf distribution in top Chinese firms and an economic explanation. *Physica A: Statistical Mechanics and its Applications, 388*(10), 2020-2024.

CAPÍTULO IV. Distribución de probabilidad implícita en las opciones sobre el WTI: Black Scholes versus el enfoque semi-noparamétrico

IV.1. Introducción

La fluctuación sufrida por los precios del petróleo en los últimos años ha causado enorme preocupación entre los consumidores, las empresas y los gobiernos (Huang, Yu, Fabozzi, & Fukushima, 2009; Kallis & Sager, 2017). A nivel mundial, los pronósticos de variables macroeconómicas se ven impactados en gran medida por la evolución de las predicciones del precio del petróleo, ya que la actividad económica y la inflación dependen en gran medida de ella (He, Kwok, & Wan, 2010; Kallis & Sager, 2017). La dificultad radica en que el precio del petróleo está fuertemente influenciado por los niveles de inventario, el clima, los desequilibrios a corto plazo entre la oferta y la demanda y por asuntos políticos. Estos elementos dificultan el poder determinar un precio adecuado para el crudo (Huang, Yu, Fabozzi, & Fukushima, 2009; de Souza e Silva, Legey, & de Souza e Silva, 2010; Abhyankar, Xu, & Wang, 2013). Dado que la función de densidad neutral de riesgo (RND) incorpora las expectativas del mercado sobre el desarrollo futuro de los precios de los activos subyacentes (Jondeau & Rockinger, 2000; Liu, Shackleton, Taylor, & Xu, 2007; Monteiro, Tütüncü, & Vicente, 2008; Fabozzi, Tunaru, & Albota, 2009; Du, Wang, & Du, 2012; Lai, 2014; Taboga, 2016; Kiesel & Rahe, 2017), la convierte en una herramienta útil para la modelización del precio de este "commodity".

Una fuente valiosa de información para poder obtener la RND se encuentra en los precios de las opciones financieras (Liu, 2007; Rompolis, 2010; Völkert, 2015). Por ello, con la presente investigación se busca contribuir a la literatura sobre la estimación de la RND a partir de la modelización de opciones sobre el petróleo crudo West Texas Intermediate (WTI) cotizadas en la bolsa de materias primas New York Mercantile Exchange (NYMEX). Diferentes modelos se han desarrollado para extraer la RND. Sin embargo, su eficacia debe probarse en varios tipos de mercados y no solo en el mercado accionario en donde se han aplicado la mayoría de estudios (p.ej. Corrado & Su, 1996; Corrado & Su, 1997; Hartvig, Jensen, &

Pedersen, 2001; Lim, Martin, & Martin, 2005; Monteiro, Tütüncü, & Vicente, 2008; Birru & Figlewski, 2012; Christoffersen, Heston, & Jacobs, 2013; Kiesel & Rahe, 2017; Leippold & Schärer, 2017; entre otros). La importancia de estudiar el comportamiento de los precios de las opciones sobre el WTI radica en que el petróleo sigue siendo un componente energético principal de las economías modernas. En consecuencia, los cambios en los precios del petróleo pueden producir importantes efectos sobre la economía mundial, lo que hace necesario proponer métodos que permitan ajustar el proceso estocástico del precio futuro (Abhyankar, Xu, & Wang, 2013; Su, Li, Chang, & Lobonț, 2017).

La teoría bajo la cual se desarrolla la modelización de los precios de los activos financieros comenzó con la publicación del modelo de valoración de Black-Scholes (1973) y desde entonces ha servido como la base fundamental para numerosas generalizaciones y extensiones por parte de académicos y profesionales de las finanzas (Peña, Rubio, & Serna, 1999; Liu, 2007; León, Mencía, & Sentana, 2009; Rompolis, 2010; Du, Wang, & Du, 2012; Lai, 2014; Feng & Dang, 2016). Sin embargo, el modelo de Black-Scholes se ha vuelto cada vez menos fiable a lo largo del tiempo. Incluso en los mercados donde se esperaría que fuera más exacto, aparecen diferencias entre los precios teóricos y los precios de mercado (Jarrow & Rudd, 1982; Corrado & Su, 1996; Backus, Foresi, Li, & Wu, 1997; Birru & Figlewski, 2012; Christoffersen, Heston, & Jacobs, 2013).

Específicamente, se sabe que después de la crisis del mercado bursátil en octubre de 1987, el modelo de valoración de opciones de Black-Scholes tiende a infravalorar las opciones que están muy "en dinero" y muy "fuera de dinero" (véase Rubinstein (1994) para una discusión detallada de esta regularidad empírica). Este hecho es el resultado de la violación del supuesto bajo el cual todos los precios de opciones sobre el mismo activo subyacente con la misma fecha de expiración, pero con diferente precio de ejercicio deberían tener la misma volatilidad implícita (Corrado & Su, 1997; Lim, Martin, & Martin, 2005; Friesen, Zhang, & Zorn, 2012). La evidencia empírica muestra que la volatilidad implícita derivada del modelo de Black-Scholes parece ser diferente a través del precio de ejercicio dibujando la bien conocida sonrisa de volatilidad (Peña, Rubio, & Serna, 1999; Jondeau & Rockinger, 2000; Liu, 2007; Kiesel & Rahe, 2017).

Bajo Black-Scholes, la RND es lognormal, pero esta predicción se ha rechazado de manera convincente (MacBeth & Merville, 1979). Por ello, la literatura sobre valoración de opciones ha propuesto modelos que permiten incorporar ajustes tanto de sesgo como de exceso de curtosis en la RND con el fin de corregir los problemas anteriormente mencionados (Backus, Foresi, Li, & Wu, 1997; Nikkinen, 2003; Jondeau, Poon, & Rockinger, 2007, p. 365; Friesen, Zhang, & Zorn, 2012). La relevancia de este tipo de modelos radica en que asumir que el logaritmo del precio del activo es normal es poco realista. En particular porque las colas de la distribución resultan ser más pesadas que las de una distribución normal (Fama, 1965; Das & Sundaram, 1999; Dennis & Mayhew, 2002; Nikkinen, 2003; Huang, Yu, Fabozzi, & Fukushima, 2009; Feng & Dang, 2016).

Al respecto, la literatura académica más reciente ha procedido en dos direcciones con el fin de determinar la RND. Una de ellas consiste en especificar un proceso estocástico del precio alternativo al propuesto por Black-Scholes, que a su vez implican una RND alternativa. La otra busca desarrollar procedimientos para extraer RND implícitas de los precios observados de las opciones (Hartvig, Jensen, & Pedersen, 2001; Dennis & Mayhew, 2002; Lai, 2014). De acuerdo con este último enfoque, Breeden & Litzerberger (1978), Shimko (1993) y Jondeau, Poon, & Rockinger (2007, p. 398) sugieren hacer uso del hecho de que la RND es la segunda derivada del precio de la opción de compra (call) respecto al precio de ejercicio.

Sin embargo, otros autores sugieren otros modelos como los paramétricos, los cuales proponen una expresión directa para la RND sin referirse a ninguna dinámica del precio (Ritchey, 1990; Melick & Thomas, 1997; Anagnou-Basioudis, Bedendo, Hodges, & Tompkins, 2005; Fabozzi, Tunaru, & Albota, 2009; Völkert, 2015); los no-paramétricos que no tratan de dar una forma explícita a la RND (Jackwerth & Rubinstein, 1996; Aït-Sahalia & Lo, 1998); y los semi-noparamétricos (SNP) que proponen alguna aproximación de la RND (Jarrow & Rudd, 1982; Corrado & Su, 1996; Backus, Foresi, Li, & Wu, 1997; Rompolis & Tzavalis , 2007; León, Mencía, & Sentana, 2009; Taboga, 2016).

El enfoque del presente estudio busca extraer la RND implícita en los precios de opciones aplicando un modelo SNP. En particular, se verifica si el modelo SNP propuesto por Backus, Foresi, Li, & Wu (1997) permite mejorar la valoración sobre opciones sobre el WTI cotizadas en el periodo de enero de 2016 a enero de 2017. Además, con el fin de contrastar los

resultados obtenidos, las fechas de análisis se separaron entre eventos especiales en el mercado del petróleo o eventos políticos que pudieran afectar los mercados financieros y fechas de días de "relativa" calma. En una primera etapa, se calibraron los parámetros de sesgo y de exceso de curtosis haciendo uso de la distribución SNP. En una segunda etapa, se usaron los parámetros (sesgo y exceso de curtosis) estimados previamente con el fin de aproximar la distribución del precio del activo subyacente bajo una especificación que denotamos como log-SNP.[20] La ventaja de aplicar modelos SNP es que no son tan intensivos en datos como otros métodos que permiten extraer la RND (Aït-Sahalia & Lo, 1998; Taboga, 2016). Como se sabe, la dinámica de los mercados financieros ofrece un número limitado de precios durante un día de negociación, lo que hace necesario proponer métodos que se ajusten en la práctica (Liu, 2007; Feng & Dang, 2016).

Backus, Foresi, Li & Wu (1997) adoptan una expansión de series de Gram-Charlier tipo A (aquí expresada como SNP) alrededor de la función de densidad de la normal con el fin de obtener los términos de ajuste de asimetría y de exceso de curtosis para la fórmula de Black-Scholes. Estos autores tomaron como base del modelo propuesto por Jarrow & Rudd (1982) quienes fueron pioneros en proponer un modelo SNP de valoración de opciones usando una expansión de la serie de Edgewoth alrededor de la función de densidad de la lognormal. Posteriormente, Corrado & Su (1996) también derivaron un modelo de valoración de precios de la opción utilizando una expansión de series de Gram-Charlier alrededor de la función de densidad de la normal.

Aunque el modelo de Corrado & Su (1996) se deriva de Jarrow & Rudd (1982), operacionalmente los pioneros explican las desviaciones de sesgo y de exceso de curtosis de la lognormalidad del precio del activo, mientras que el modelo desarrollado por Corrado & Su (1996) explica las desviaciones de sesgo y exceso de curtosis de la normalidad de los rendimientos de las acciones. Es de notar que Brown & Robinson (2002) corrigen dos errores tipográficos de Corrado & Su (1996) y proveen ejemplos de cómo esos errores pueden tener significación económica. En este estudio se adopta el modelo de Backus, Foresi, Li, & Wu

[20] El enfoque de Black-Scholes considera que el precio del activo subyacente se distribuye bajo una especificación lognormal, de modo que las variaciones del mismo siguen una distribución normal. Como se muestra en la Sección IV.2, la distribución Gram-Charlier o SNP corresponde a una extensión de la distribución normal y la distribución log-SNP corresponde a una extensión de la distribución lognormal. En consecuencia, el modelo de valoración de opciones SNP es una generalización del modelo de Black-Scholes.

(1997) ya que ellos demuestran que algunos de los términos del modelo de Corrado & Su (1996) son numéricamente muy pequeños en los mercados reales y pueden ser eliminados del modelo de valoración. Así, Backus, Foresi, Li, & Wu (1997) proponen un modelo más parsimonioso que constituye una buena aproximación del precio de la opción.

Este capítulo se divide de la siguiente forma: la sección IV.2 presenta el modelo a estimar y la metodología aplicada. La sección IV.3 describe los datos a utilizar. La sección IV.4 recoge los resultados y se discute el método propuesto, y finalmente la sección IV.5 presenta las conclusiones del capítulo.

IV.2. Modelo y metodología

IV.2.1. Modelo

El primer intento de estimar la RND fue desarrollado por Breeden & Litzerberger (1978). Los autores mostraron que la RND se puede recuperar de la segunda derivada del precio call. Sea C (P) una opción Europea call (put) con precio de ejercicio K y tiempo al vencimiento τ,

$$C(K;\theta) = \int_K^\infty e^{-r\tau}\,(S_T - K)q(S_T;\theta)dS_T, \tag{IV.1}$$

$$P(K;\theta) = \int_0^K e^{-r\tau}\,(K - S_T)q(S_T;\theta)dS_T, \tag{IV.2}$$

donde r es la tasa libre de riesgo. Entonces,

$$\left.\frac{\partial^2 C}{\partial K^2}\right|_{K=S_T} = e^{-r\tau}q(S_T;\theta). \tag{IV.3}$$

El término $e^{-r\tau}q(S_T)$ se refiere generalmente como el *"state price density"* (SPD) y $q(.)$ es la RND no descontada (Jondeau, Poon, & Rockinger, 2007, p. 387). La estimación de la RND por (IV.3) requiere de una serie continua de precios de ejercicio para estimar los parámetros empleando un enfoque de diferencias finitas. Sin embargo, este procedimiento deriva resultados inestables y varios métodos como *(i)* volatilidad local o modelos de árboles implícitos, *(ii)* interpolación de la curva de volatilidad implícita, *(iii)* volatilidad estocástica y saltos, *(iv)* enfoque no-paramétrico y *(v)* la combinación de los enfoques paramétrico y no-paramétrico se han propuesto en la literatura (véase por ejemplo Fusai & Roncoroni, 2008 y

las referencias allí encontradas). El presente trabajo emplea el enfoque paramétrico clásico de la aproximación lognormal y su desempeño se compara con el enfoque propuesto SNP que se explica en la siguiente subsección.

IV.2.2. Metodología

Para una fecha determinada y varios contratos de opciones call y put, con el mismo vencimiento y diferentes precios de ejercicio, se estima el conjunto de parámetros θ de los modelos Black-Scholes y SNP minimizando la suma de los errores cuadrados entre los precios de mercado observados y los precios teóricos, y el conjunto de parámetros se emplea para representar la RND para cada modelo. En un primer paso, se calibran los parámetros (μ,σ, δ_3, δ_4) empleando los modelos de Black-Scholes y SNP para opciones call y put. Esos parámetros se utilizan en un segundo paso para ajustar la función de densidad de probabilidad (pdf) asumiendo una distribución lognormal y log-SNP, respectivamente.[21]

El bien conocido precio teórico para la opción call de Black-Scholes viene dado por

$$C^{BS}(K;\theta) = S_T\Phi(d_1) - Ke^{-r\tau}\Phi(d_2), \qquad \text{(IV.4)}$$

donde $\Phi(.)$ denota la función de distribución acumulada (cdf) de la normal estándar, $d_1 = [ln(S_T/K) + (r + \sigma^2/2)\tau]/\sigma\sqrt{\tau}$ y $d_2 = d_1 - \sigma\sqrt{\tau}$. Por otra parte, el precio de la opción call del modelo SNP se puede formular como (Backus, Foresi, Li, & Wu, 1997; Christoffersen, 2012, p. 237):

$$C^{SNP}(K;\theta) = C^{BS}(K;\theta) + S_T\phi(d_1)\sigma\left[\delta_3\left(2\sqrt{\tau}\sigma - d_1\right) - \delta_4\sqrt{\tau}\left(1 - d_1^2 + 3d_1\sqrt{\tau}\sigma - 3\tau\sigma^2\right)\right]. \qquad \text{(IV.5)}$$

Los precios put se obtienen a través de la Paridad Put-Call. Para obtener (IV.5), los log-retornos se asumen distribuidos Gram-Charlier en lugar de normal como en el caso Black-Scholes. La pdf de la distribución Gram-Charlier es dada por

[21] Toda la metodología expuesta en el presente capítulo se desarrolló con base en códigos de la librería Risk Neutral Density Extraction Package (RND) del paquete R. Específicamente se realizaron modificaciones para programar las estimaciones del modelo SNP. El código está disponible bajo petición. Para información al respecto ver https://cran.r-project.org/web/packages/RND/index.html

$$f(x) = [1 + \sum_{s=1}^{n} \delta_s H_s(x)]\phi(x), \quad \text{(IV.6)}$$

donde δ_s son los parámetros y la distribución Gram-Charlier colapsa a la distribución normal cuando $\delta_s = 0$, y $H_s(x)$ es el polinomio de Hermite (HP) de orden s, el cual se puede definir en términos las derivadas de la densidad normal estándar $\phi(x)$, como $\frac{d^s\phi(x)}{dx^s} = (-1)^s H_s(x)\phi(x)$.

En particular, los cuatro primeros HP son:

$$H_1(x) = x, \quad \text{(IV.7)}$$

$$H_2(x) = x^2 - 1, \quad \text{(IV.8)}$$

$$H_3(x) = x^3 - 3x, \quad \text{(IV.9)}$$

$$H_4(x) = x^4 - 6x^2 + 3. \quad \text{(IV.10)}$$

Cabe mencionar que se pueden plantear otras expresiones para el precio de la opción SNP, véase por ejemplo Jarrow & Rudd (1982) y Corrado & Su (1996). La principal diferencia estriba en que la aproximación se hace sobre el logaritmo del precio en lugar del precio según el trabajo de Jarrow & Rudd (Backus, Foresi, Li, & Wu, 1997).

Cada modelo se calibra seleccionando un conjunto de parámetros θ, los cuales minimizan la suma de las diferencias cuadráticas entre los precios teóricos (Black-Scholes y SNP) y los precios de mercado observados para diferentes valores de N_c calls y N_p puts y el mismo tiempo al vencimiento. Los precios de mercado call y put se denotan por C_i^{mkt} y P_i^{mkt}, respectivamente.

$$\min_{\theta} \left\{ \sum_{i=1}^{N_c} \left(C_i^{mkt} - C(K_i;\theta) \right)^2 + \sum_{j=1}^{N_p} \left(P_i^{mkt} - P\left(K_j;\theta\right) \right)^2 \right\}. \quad \text{(IV.11)}$$

Para estimar la exactitud de cada método, se realiza una regresión lineal de los valores call (put) predichos por cada método como una variable dependiente y los respectivos valores de mercado como una variable independiente. Se considera que el método con la mínima media del valor absoluto de los residuales (MAE) es el mejor. Para obtener la gráfica de la RND no

descontada para el modelo de Black-Scholes, se emplea la densidad lognormal con los parámetros obtenidos del procedimiento de calibración,

$$q^{LogN}(S_T;\theta) = \frac{1}{S_T\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{\ln S_T-\mu}{\sigma}\right)^2}. \qquad \text{(IV.12)}$$

De forma similar, la gráfica de la RND no descontada para el modelo SNP se obtienen empleando la distribución log-SNP propuesta por Ñíguez, Paya, Peel, & Perote (2012) y los parámetros calibrados a partir del modelo SNP para los precios de opciones. Esta distribución ha mostrado resultados excepcionales en la literatura, donde la distribución lognormal se ha propuesto como un modelo de referencia (véase Cortés, Mora-Valencia, & Perote, 2016). La pdf de la log-SNP se especifica como

$$q^{logSNP}(S_T;\theta) = \left[1 + \sum_{s=1}^{n} \delta_s H_s\left(\frac{\ln S_T-\mu}{\sigma}\right)\right]\left(\frac{1}{S_T\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{\ln S_T-\mu}{\sigma}\right)^2}\right), \qquad \text{(IV.13)}$$

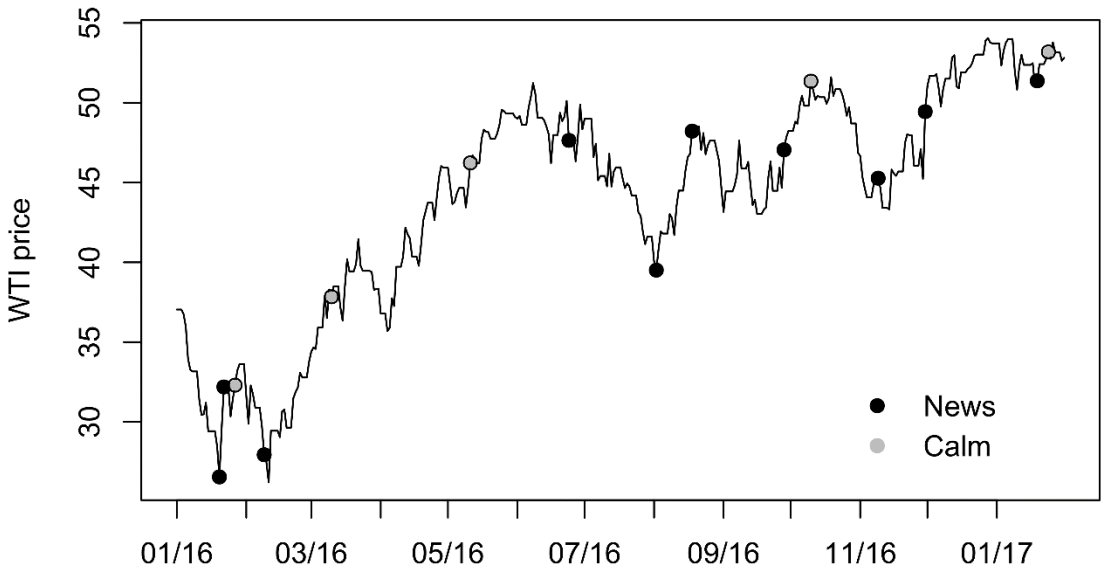
donde H_s denota el HP de orden s. Es de notar que la distribución lognormal se recupera de la log-SNP si $\delta_s = 0$. Además, si una variable aleatoria x se distribuye como log-SNP, entonces $\log(x)$ se distribuye como Gram-Charlier, lo cual se asemeja a la relación entre las variables aleatorias lognormal y normal, y teniendo en cuenta que la distribución Gram-Charlier puede entenderse como una expansión de la distribución normal en términos de los polinomios de Hermite.

IV.3. Descripción de los datos

La base de datos recopilada incluye los precios de cierre de contratos de opciones de compra y venta sobre petróleo crudo WTI cotizadas en NYMEX. En particular, se obtuvieron datos para 10 fechas espaciadas irregularmente teniendo en cuenta eventos especiales en el mercado del petróleo o eventos políticos que pudieran afectar los mercados financieros. Con el fin de contrastar los resultados obtenidos, también se seleccionaron al azar 5 fechas de días de "relativa" calma. La primera fecha fue el 20 de enero de 2016, mientras que la última fue el 24 de enero de 2017. Específicamente, para cada evento se seleccionaron opciones con

diferentes precios de ejercicio y con la misma fecha de expiración. Para obtener los precios de cotización sobre el WTI y de las opciones call y put se recurrió a la base de datos Bloomberg. Asimismo, las noticias se recopilaron de Financial Information Network de Bloomberg y del sitio web de la OPEC.[22] Los contratos seleccionados tienen como vencimiento aproximadamente entre 30 y 60 días, por lo que se toma como tasa libre de riesgo de referencia la London Interbank Offered Rate (LIBOR) a 1 y 2 meses dependiendo del contrato. La base de información para la LIBOR fue ICE Benchmark Administration (IBA).[23]

Figura IV.1 Evolución de los precios del petróleo crudo WTI



La figura corresponde a la serie de tiempo de la evolución del precio de contado del petróleo crudo WTI entre el 1 enero de 2016 y el 31 de enero de 2017. En la figura se pueden observar las marcas de las fechas seleccionadas en el presente estudio. La marca negra (News) representa una fecha con noticias destacadas que impactaron el precio del petróleo. La marca gris (Calm) corresponde a fechas sin noticias especiales sobre el petróleo.

El interés en analizar el mercado del petróleo radica en que cambios inesperados en el precio de este commodity han mostrado un impacto en la economía mundial, por lo tanto es un tema

[22] Ver: http://www.opec.org/opec_web/en/index.htm
[23] Ver: https://www.theice.com/iba/libor. Este estudio se limitará a la Libor sin tener en cuenta otras tasas de referencia que pueden ser estudiadas en otra futura investigación.

de interés para inversionistas y bancos centrales (Postali & Picchetti, 2006; He, Kwok, & Wan, 2010; Antonakakis, Chatziantoniou, & Filis, 2017). La Figura IV.1 corresponde a la serie de tiempo de la evolución del precio de contado del petróleo crudo WTI entre el 1 enero de 2016 y el 31 de enero de 2017.

Tabla IV.1 Noticias sobre el petróleo

Fecha del evento	Noticia	Símbolo Ticker opcion	Fecha de vencimiento	Tiempo al vencimiento en días
20/01/2016	El WTI se desploma a menos de US$28 alcanzando un nuevo mínimo en 13 años	CLH6	17/02/2016	28
22/01/2016	El WTI sube un 9,01 % y cerró en US$32.19, recuperando parte de las pérdidas que había acumulado desde el comienzo del año. La subida porcentual es la mayor desde el 27 de agosto de 2015	CLH6	17/02/2016	26
27/01/2016	Día de relativa calma	CLH6	17/02/2016	21
09/02/2016	El precio del petróleo cae por temores a exceso de suministro	CLJ6	16/03/2016	36
10/03/2016	Día de relativa calma	CLK6	15/04/2016	36
11/05/2016	Día de relativa calma	CLN6	16/06/2016	36
24/06/2016	El WTI cae un 5% siguiendo la debacle general de los mercados tras la votación de los británicos en favor de abandonar la Unión Europea	CLU6	17/08/2016	54
02/08/2016	El WTI cae por debajo de US$40 ante preocupación por oferta excesiva	CLV6	15/09/2016	44
18/08/2016	Los líderes mundiales del mercado petrolero se preparan para discutir la posibilidad de congelar los niveles de producción, lo que hace subir los precios del petróleo.	CLV6	15/09/2016	28
28/09/2016	El precio del WTI sube casi un 6% por acuerdo de la OPEC para limitar la producción en noviembre	CLZ6	16/11/2016	49
10/10/2016	Día de relativa calma	CLZ6	16/11/2016	37
09/11/2016	Jornada de precios volátiles tras la elección de Donald Trump	CLF7	15/12/2016	36
30/11/2016	Los precios del petróleo se disparan ante la perspectiva de que los países de la OPEP reunidos en Viena logren un acuerdo para limitar la producción y estimular los precios	CLG7	17/01/2017	48
19/01/2017	Los precios del petróleo cayeron significativamente durante la jornada de negociación ante las preocupaciones por el aumento de la producción de crudo de Estados Unidos, que pesaron más que las previsiones optimistas de la OPEP de aumento la demanda	CLH7	15/02/2017	27
24/01/2017	Día de relativa calma	CLH7	15/02/2017	22

La tabla muestra el detalle de las noticias sobre el petróleo en las fechas seleccionadas en el estudio y la información sobre las opciones call y put tomadas para cada uno de los eventos. Fuente: Financial Information Network de Bloomberg y sitio web de la OPEC.

En ese periodo, el precio del petróleo WTI alcanzó un máximo de US$54.06 y un mínimo de US$26.21, lo que se vio reflejado en una desviación estándar histórica del precio de US$6.89. Ese mismo comportamiento ha sido persistente en los últimos años. Desde el año 2014, los precios del petróleo han permanecido bajos, en un entorno económico en el que el crecimiento de varios países se redujo progresivamente. Adicionalmente, el descenso de los precios del petróleo ocasionó otros problemas pues también afectó los mercados bursátiles mundiales, la inflación de varias economías y llevó a los bancos centrales a elevar las tasas de interés (Kallis & Sager, 2017).

En la Tabla IV.1 se puede ver el detalle de las noticias sobre el petróleo en las fechas seleccionadas para el análisis y la información sobre las opciones call y put tomadas para cada uno de los eventos. Adicionalmente, en la Figura IV.1 también se pueden observar las marcas de las fechas seleccionadas para el análisis. La marca negra (News) se relaciona con noticias destacadas que impactaron el precio del petróleo. La marca gris (Calm) corresponde a momentos de relativa calma, sin noticias especiales sobre el petróleo.

IV.4. Resultados y discusión

La Tabla IV.2 resume los resultados de las estimaciones realizadas bajo el modelo de Black-Scholes (ecuación IV.4) y la distribución SNP (IV.5) propuesta por Backus, Foresi, Li, & Wu (1997). Haciendo uso de la ecuación IV.11 presentada en la metodología, subsección IV.2.2, se obtuvo cada uno de los parámetros para las distribuciones. Específicamente, para el modelo de Black-Scholes se muestra la desviación estándar implícita para cada una de las fechas seleccionadas (véase Panel A).

De manera similar, para el modelo SNP se presenta la desviación estándar implícita, la asimetría implícita y el exceso de curtosis implícito (véase Panel B). Puesto que las opciones call y put con el mismo precio de ejercicio y la misma fecha de vencimiento están relacionadas a través de la Paridad Put-Call, el estudio se enfoca solo en los resultados de las opciones call.

Tabla IV.2 Parámetros estimados, Black-Scholes versus SNP

Fecha del evento	Número de precios observados	Panel A Black-Scholes	Panel B SNP		
		Desviación estándar implícita	Desviación estándar implícita	Asimetría Implícita	Exceso de Curtosis Implícito
20/01/2016	82	0.65	0.70	-0.15	0.08
22/01/2016	82	0.59	0.61	-0.05	0.07
27/01/2016	82	0.65	0.68	-0.07	0.06
09/02/2016	63	0.70	0.74	-0.15	0.11
10/03/2016	60	0.49	0.51	-0.14	0.11
11/05/2016	46	0.40	0.41	-0.14	0.08
24/06/2016	57	0.40	0.42	-0.30	0.13
02/08/2016	44	0.44	0.46	-0.17	0.11
18/08/2016	46	0.34	0.35	-0.13	0.07
28/09/2016	94	0.42	0.43	-0.14	0.14
10/10/2016	95	0.35	0.36	-0.16	0.12
09/11/2016	46	0.40	0.41	-0.11	0.06
30/11/2016	44	0.40	0.41	-0.17	0.04
19/01/2017	59	0.29	0.30	-0.16	0.05
24/01/2017	61	0.29	0.30	-0.11	0.03

La tabla resume los resultados de las estimaciones realizadas bajo el modelo de Black-Scholes y de la distribución SNP. La primera columna muestra cada una de las fechas seleccionadas en el estudio y en la segunda columna se encuentra en el número de precios de mercado observados en cada fecha. El Panel A recoge la desviación estándar implícita del modelo Black-Scholes. El Panel B recoge la desviación estándar implícita, la asimetría implícita y el exceso de curtosis implícito del modelo SNP.

Los resultados muestran desviaciones estándar implícitas muy próximas entre sí para cada uno de los modelos, Black-Scholes y SNP. Sin embargo, como se observa en el Panel B, los resultados sugieren que las distribuciones implícitas son leptocúrticas y sesgadas negativamente. Estos hallazgos son consistentes con los obtenidos por Corrado & Su (1996), Backus, Foresi, Li, & Wu (1997), Corrado & Su (1997) y Nikkinen (2003). Adicionalmente, refuerzan la evidencia que existe sobre el comportamiento de los rendimientos y precios de activos subyacentes, los cuales suelen no presentar comportamiento normal y lognormal respectivamente (MacBeth & Merville, 1979).

A partir de los parámetros de entrada presentados en la Tabla IV.2, se obtuvo el precio teórico para las opciones call en cada una de las fechas seleccionadas en el estudio. En la Tabla IV.3 se puede comparar el precio promedio observado en el mercado respecto al precio promedio teórico bajo una RND lognormal (véase Panel A) y una RND log-SNP (véase Panel B).

Tabla IV.3 Comparación precio promedio de mercado para opciones de compra versus precio teórico

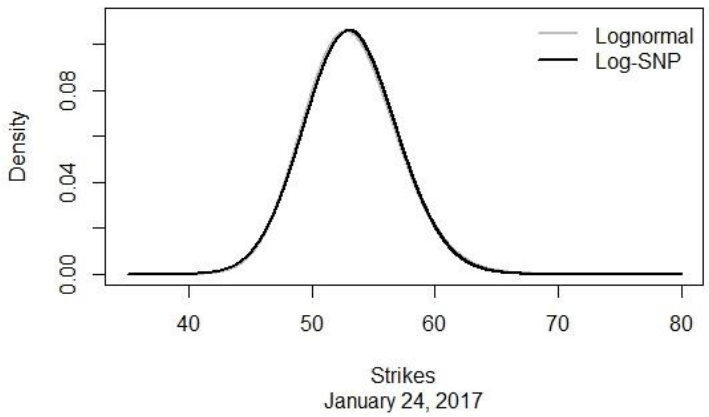
Fecha del evento	Número de precios observados	Precio promedio de mercado-opción call ($US)	Panel A Lognormal		Panel B Log-SNP	
			Precio promedio teórico-opción call ($US)	Diferencia precio promedio de mercado y teórico-opción call ($US)	Precio promedio teórico-opción call ($US)	Diferencia precio promedio de mercado y teórico-opción call ($US)
20/01/2016	82	0.507	0.501	0.005	0.504	0.003
22/01/2016	82	1.203	1.188	0.015	1.200	0.003
27/01/2016	82	1.221	1.206	0.014	1.218	0.002
09/02/2016	63	1.156	1.145	0.010	1.155	0.000
10/03/2016	60	3.984	3.965	0.019	3.982	0.002
11/05/2016	46	6.940	6.912	0.028	6.938	0.002
24/06/2016	57	6.411	6.371	0.040	6.408	0.003
02/08/2016	44	1.829	1.832	-0.003	1.827	0.002
18/08/2016	46	5.551	5.532	0.019	5.548	0.003
28/09/2016	94	3.834	3.820	0.013	3.832	0.002
10/10/2016	95	5.275	5.259	0.015	5.271	0.003
09/11/2016	46	2.515	2.514	0.001	2.514	0.001
30/11/2016	44	4.227	4.219	0.008	4.226	0.000
19/01/2017	59	4.862	4.855	0.007	4.859	0.003
24/01/2017	61	5.238	5.231	0.007	5.235	0.003

La tabla compara el precio promedio observado en el mercado respecto al precio promedio teórico. La primera columna muestra cada una de las fechas seleccionadas en el estudio, la segunda columna muestra el número de precios de mercado observados y la tercera columna muestra el precio promedio de mercado en cada fecha seleccionada en el estudio. El Panel A recoge el precio promedio teórico que sigue una RND lognormal. El Panel B recoge el precio promedio teórico que sigue una RND log-SNP.

Al comparar los precios promedio de mercado y los precios promedio teóricos para cada una de las distribuciones se encuentra que si los precios siguen una RND lognormal, estos tienden a infravalorar sistemáticamente en una mayor cantidad monetaria los precios de las opciones de compra. En especial, para las fechas 11 de mayo y 24 de junio de 2016 en las que el promedio de las opciones estuvieron más en dinero la diferencia es más notoria. Este

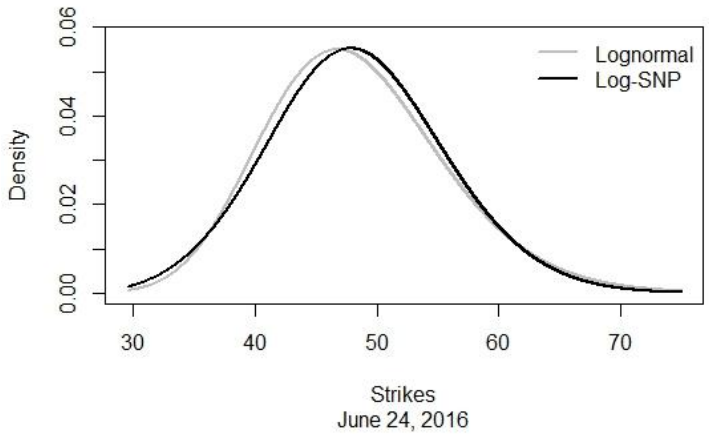
resultado no es sorprendente ya que esta regularidad empírica la obtuvo Rubinstein (1994) para opciones sobre el índice S&P500.

Figura IV.2 Densidad neutral al riesgo



La figura muestra la función de densidad neutral de riesgo (RND) para el 24 de enero de 2017, día de relativa calma en los mercados financieros (Calm). La línea gris corresponde a especificación lognormal y la línea negra corresponde a una especificación Log-SNP.

Figura IV.3 Densidad neutral al riesgo



La figura muestra la función de densidad neutral de riesgo (RND) para el 24 de junio de 2016, día de noticias que afectaron los mercados financieros (News). La línea gris corresponde a especificación lognormal y la línea negra corresponde a una especificación Log-SNP.

Un ejemplo de las RNDs lognormal (ecuación IV.12) y log-SNP (ecuación IV.13) se recoge en las Figuras IV.2 y IV.3. Al azar se seleccionó una de las fechas de relativa calma y una de las fecha con un evento que afectó el comportamiento del WTI. La primera de ellas corresponde al 24 de enero de 2017 (Figura IV.2). La segunda fecha corresponde al 24 de junio de 2016, día de reacción adversa en los mercados financieros tras la votación de los británicos en favor de abandonar la Unión Europea (Figura IV.3). Obsérvese que para la fecha de relativa calma (Calm) las RNDs bajo ambas especificaciones no parecen ser muy diferentes. Sin embargo, para la segunda fecha (News), la RND que sigue la distribución log-SNP es más sesgada que la lognormal, lo que al parecer, permite recoger mejor la evolución de los precios de las opciones.

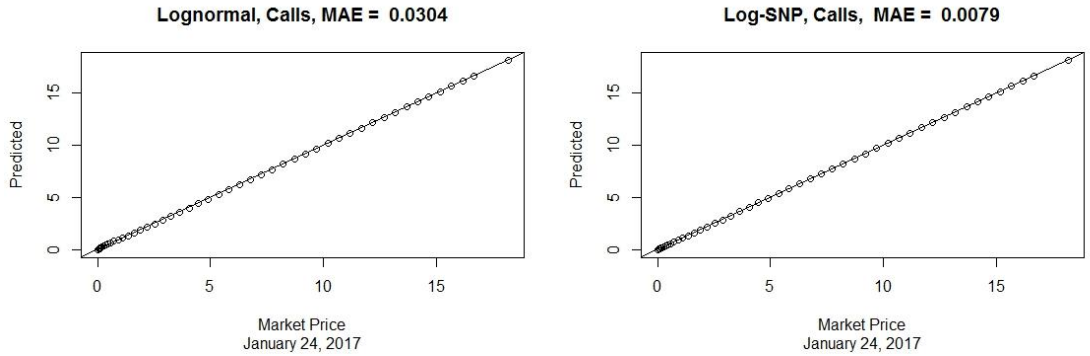
Adicionalmente a la diferencia monetaria entre los precios promedio de mercado y los teóricos, mostradas en la Tabla IV.3, como una medida de bondad de ajuste se presenta la media del valor absoluto de los residuales (MAE) calculada como se explicó en la subsección IV.2.2. Como se observa en la Tabla IV.4, para todas las fechas de estudio, los precios que siguen una RND log-SNP presentan consistentemente un MAE menor.

Tabla IV.4 Media del valor absoluto de los residuales, lognormal versus log-SNP

Fecha del evento	Lognormal	Log-SNP
	Media error absoluto	Media error absoluto
20/01/2016	0.0204	0.0051
22/01/2016	0.0165	0.0039
27/01/2016	0.0186	0.0039
09/02/2016	0.0342	0.0033
10/03/2016	0.0449	0.0068
11/05/2016	0.0516	0.0061
24/06/2016	0.1076	0.0186
02/08/2016	0.0397	0.0103
18/08/2016	0.0445	0.0067
28/09/2016	0.0478	0.0090
10/10/2016	0.0451	0.0076
09/11/2016	0.0326	0.0076
30/11/2016	0.0464	0.0091
19/01/2017	0.0458	0.0080
24/01/2017	0.0304	0.0079

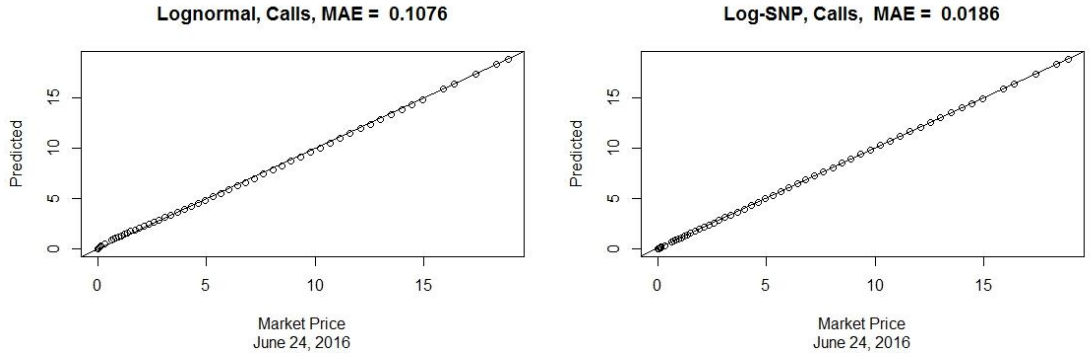
La tabla recoge la media del valor absoluto de los residuales (MAE) en la estimación de los precios de las opciones de compra. La primera columna muestra cada una de las fechas seleccionadas en el estudio. La segunda columna muestra el MAE bajo una especificación lognormal y la tercera columna muestra el MAE bajo una especificación log-SNP.

Figura IV.4 Media del valor absoluto de los residuales



La figura muestra la media del valor absoluto de los residuales (MAE) en la estimación de los precios de las opciones de compra (call) en la fecha 24 de enero de 2017 día de relativa calma en los mercados financieros (Calm). La figura de la izquierda corresponde al MAE bajo una especificación lognormal y la figura de la derecha corresponde al MAE bajo una especificación log-SNP.

Figura IV.5 Media del valor absoluto de los residuales



La figura muestra la media del error del valor absoluto de los residuales (MAE) en la estimación de los precios de las opciones de compra (call) 24 de junio de 2016 día de noticias que afectaron los mercados financieros (News). La figura de la izquierda corresponde al MAE bajo una especificación lognormal y la figura de la derecha corresponde al MAE bajo una especificación log-SNP.

Las Figuras IV.4 y IV.5 soportan gráficamente los resultados presentados en la Tabla IV.4. En concreto, se ofrece un ejemplo para las mismas fechas de calma (Figura IV.4) y de noticias en el mercado (Figura IV.5) seleccionadas anteriormente. La principal diferencia entre las RNDs modelizadas es la capacidad de capturar los momentos de alto orden, tales como sesgo y exceso de curtosis. Especialmente para las fechas de más incertidumbre en el mercado, los resultados sugieren un mejor ajuste para los precios a partir de la distribución log-SNP.

Estudiar modelos que permitan obtener un mejor ajuste entre los precios de mercado y los precios teóricos es fundamental, no solo desde el punto de vista de la valoración de opciones, sino también como propósito en la gestión de riesgos. Por ejemplo, dentro de la gestión de riesgos uno de los efectos más importantes que se trata de cuantificar es el cambio del precio de la opción frente a un cambio en el precio del activo subyacente (Backus, Foresi, Li, & Wu, 1997).

En ese caso, como estrategia de cobertura ante el riesgo se requiere el cálculo de medidas como el Delta de la opción. Esta medida precisamente, cuantifica la sensibilidad del precio de la opción al cambio en el precio del activo subyacente. En el caso del modelo de Black-Scholes se puede demostrar que el Delta (Δ^{BS}) está dado por $\Delta_{call}^{BS} = \Phi(d_1)$ en el caso de la opción call y por $\Delta_{put}^{BS} = \Phi(d_1) - 1$ para la opción put (véase la prueba en el Apéndice IV.A).

Sin embargo, como se mostró anteriormente, los resultados obtenidos en la Tabla IV.2 sugieren que las distribuciones implícitas en los precios de las opciones son leptocúrticas y negativamente sesgadas. Por ello, se requiere que el Delta también recoja los efectos del sesgo y del exceso de curtosis. En este caso el Delta del modelo SNP viene dado por

$$\Delta_{call}^{SNP} = \Phi(d_1) - \frac{\delta_3}{\sqrt{\tau}}\emptyset(d_1)\left(1 - d_1^2 + 3d_1\sigma\sqrt{\tau} - 2\sigma^2\tau\right) +$$

$$\delta_4\emptyset(d_1)\left[3d_1(1 - 2\sigma^2\tau) - d_1^3 + 4d_1^2\sigma\sqrt{\tau} - 4\sigma\sqrt{\tau} + 3\sigma^3\tau^{3/2}\right], \qquad \text{(IV.14)}$$

para la opción call y por

$$\Delta_{put}^{SNP} = \frac{\partial P^{SNP}}{\partial S_T} - 1 \qquad \text{(IV.15)}$$

para la opción put (véase la prueba en el Apéndice IV.B).

Como se observa en las ecuaciones anteriores la aproximación tradicional de Black-Scholes, que se usa frecuentemente en la cobertura y en la gestión del riesgo de las opciones, puede diferir sustancialmente cuando el precio de la opción muestra asimetría y exceso de curtosis. Como consecuencia, se puede llegar a decisiones de cobertura erradas que pueden conllevar a enfrentar pérdidas severas.

IV.5. Conclusiones

En el presente estudio se adopta el modelo SNP propuesto por Backus, Foresi, Li, & Wu (1997) quienes siguen una expansión de series de Gram-Charlier tipo A alrededor de la función de densidad de la normal. Como modelo de comparación se toma a Black-Scholes, el cual es un estándar universal empleado en la valoración de opciones. A partir de precios de opciones sobre petróleo crudo WTI transadas en NYMEX en el periodo de enero de 2016 a enero de 2017, se calibran los parámetros de sesgo y de exceso curtosis haciendo uso de la distribución SNP. En relación con una distribución normal, se encuentra una asimetría negativa y un exceso de curtosis positivo. Esos resultados fueron consistentes para 10 fechas seleccionadas teniendo en cuenta eventos especiales en el mercado del petróleo o eventos políticos que pudieran afectar los mercados financieros y para 5 fechas de días de relativa calma.

Adicionalmente, al comparar el precio promedio observado en el mercado respecto al precio promedio teórico, se halla que los precios de las opciones bajo una RND lognormal tienden a infravalorarse sistemáticamente. Ese resultado se hace más evidente en las fechas de mayor incertidumbre en los mercados financieros. En resumen, puede concluirse que los términos de ajuste de asimetría y de exceso de curtosis producen una precisión significativamente mejorada para obtener los precios de opciones sobre el WTI y ajustarlos a una RND log-SNP. Adicionalmente considerar estos términos es fundamental a la hora de realizar estrategias adecuadas de cobertura y gestión del riesgo.

Referencias

Abhyankar, A., Xu, B., & Wang, J. (2013). Oil price shocks and the stock market: evidence from Japan. *The Energy Journal, 34(2)*, 199-222.

Aït-Sahalia, Y., & Lo, A. W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *The Journal of Finance, 53(2)*, 499–547.

Anagnou-Basioudis, I., Bedendo, M., Hodges, S. D., & Tompkins, R. (2005). Forecasting accuracy of implied and GARCH-based probability density functions. *Review of Futures Markets 11(1)*, 41-66.

Antonakakis, N., Chatziantoniou, I., & Filis, G. (2017). Oil shocks and stock markets: dynamic connectedness under the prism of recent geopolitical and economic unrest. *International Review Of Financial Analysis, 50*, 1-26.

Backus, D., Foresi, S., Li, K., & Wu, L. (1997). Accounting for Biases in Black-Scholes. *Manuscript, The Stern School at New York University*, 1-46.

Birru, J., & Figlewski, S. (2012). Anatomy of a meltdown: the risk neutral density for the S&P 500 in the fall of 2008. *Journal of Financial Markets, 15(2)*, 151-180.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy, 81(3)*, 637–659.

Breeden, D., & Litzerberger, R. (1978). Prices of state-contingent claims implicit in options prices. *Journal of Business, 51(4)*, 621–651.

Brown, C. A., & Robinson, D. M. (2002). Skewness and kurtosis implied by option prices: a correction. *The Journal of Financial Research, 25(2)*, 279–282.

Christoffersen, P. (2012). *Elements of financial risk management (2nd ed.).* Waltham: Academic Press.

Christoffersen, P., Heston, S., & Jacobs, K. (2013). Capturing option anomalies with a variance-dependent pricing kernel. *Review of Financial Studies, 26(8)*, 1963-2006.

Corrado, C. J., & Su, T. (1996). Skewness and kurtosis in S&P 500 index returns implied by option prices. *The Journal of Financial Research, 19(2)*, 175-192.

Corrado, C. J., & Su, T. (1997). Implied volatility skews and stock return skewness and kurtosis implied by stock option prices. *The European Journal of Finance, 3(1)*, 73–85.

Cortés, L. M., Mora-Valencia, A., & Perote, J. (2016). The productivity of top researchers: a semi-nonparametric approach. *Scientometrics*, *109*(2), 891-915.

Das, S. R., & Sundaram, R. K. (1999). Of smiles and smirks: a term structure perspective. *The Journal of Financial and Quantitative Analysis, 34(2)*, 211-239.

de Souza e Silva, E. A., Legey, L. F., & de Souza e Silva, E. A. (2010). Forecasting oil price trends using wavelets and hidden Markov models. *Energy Economics, 32(6)*, 1507-1519.

Dennis, P., & Mayhew, S. (2002). Risk-neutral skewness: evidence from stock options. *Journal of Financial and Quantitative Analysis, 37(3)*, 471–493.

Du, Y., Wang, C., & Du, Y. (2012). Inversion of option prices for implied risk-neutral probability density functions: general theory and its applications to the natural gas market. *Quantitative Finance, 12(12)*, 1877-1891.

Fabozzi, F. J., Tunaru, R., & Albota, G. (2009). Estimating risk-neutral density with parametric models in interest rate markets. *Quantitative Finance, 9(1)*, 55-70.

Fama, E. (1965). The behaviour of stock market prices. *Journal of Business, 38(1)*, 34–105.

Feng, P., & Dang, C. (2016). Shape constrained risk-neutral density estimation by support vector regression. *Information Sciences, 333(C)*, 1-9.

Friesen, G., Zhang, Y., & Zorn, T. (2012). Heterogeneous beliefs and risk-neutral skewness. *Journal of Financial and Quantitative Analysis, 47(4)*, 851-872.

Fusai, G., & Roncoroni, A. (2008). *Implementing models in quantitative finance (1st ed.).* Berlin: Springer.

Hartvig, N. V., Jensen, J. L., & Pedersen, J. (2001). A class of risk neutral densities with heavy tails. *Finance and Stochastics, 5(1)*, 115-128.

He, A., Kwok, J., & Wan, A. (2010). An empirical model of daily highs and lows of West Texas Intermediate crude oil prices. *Energy Economics, 32(6)*, 1499-1506.

Huang, D., Yu, B., Fabozzi, F., & Fukushima, M. (2009). CAViaR-based forecast for oil price risk. *Energy Economics, 31(4)*, 511-518.

Jackwerth, J., & Rubinstein, M. (1996). Recovering probability distributions from options prices. *Journal of Finance, 51(5)*, 1611–1631.

Jarrow, R., & Rudd, A. (1982). Approximate option valuation for arbitrary stochastic processes. *Journal of Financial Economics,10(3)*, 347-369.

Jondeau, E., & Rockinger, M. (2000). Reading the smile: the message conveyed by methods which infer risk neutral densities. *Journal of International Money and Finance, 19(6)*, 885–915.

Jondeau, E., Poon, S.-H., & Rockinger, M. (2007). *Financial Modeling Under Non-Gaussian Distributions.* London: Springer.

Kallis, G., & Sager, J. (2017). Oil and the economy: a systematic review of the literature for ecological economists. *Ecological Economics, 131*, 561-571.

Kiesel, R., & Rahe, F. (2017). Option pricing under time-varying risk-aversion with applications to risk forecasting. *Journal Of Banking & Finance, 76(C)*, 120-138.

Lai, W.-N. (2014). Comparison of methods to estimate option implied riskneutral densities. *Quantitative Finance, 14(10)*, 1839-1855.

Leippold, M., & Schärer, S. (2017). Discrete-time option pricing with stochastic liquidity. *Journal Of Banking & Finance, 75*, 1-16.

León, A., Mencía, J., & Sentana, E. (2009). Parametric properties of semi-nonparametric distributions, with applications to optionvaluation. *Journal of Business & Economic Statistics, 27(2)*, 176-192.

Lim, G. C., Martin, G. M., & Martin, V. L. (2005). Parametric pricing of higher order moments in S&P500 options. *Journal of Applied Econometrics,20(3)*, 377–404.

Liu, X. (2007). Bid–ask spread, strike prices and risk-neutral densities. *Applied Financial Economics, 17(11)*, 887–900.

Liu, X., Shackleton, M. B., Taylor, S. J., & Xu, X. (2007). Closed-form transformations from risk-neutral to real-world distributions. *Journal of Banking & Finance, 31(5)*, 1501-1520.

MacBeth, J., & Merville, L. (1979). An empirical examination of the Black-Scholes call option pricing model. *Journal of Finance, 34(5)*, 1173-1186.

Melick, W. R., & Thomas, C. P. (1997). Recovering an asset's implied PDF from option prices: an application to crude oil during the Gulf crisis. *Journal of Financial and Quantitative Analysis, 32(1)*, 91–115.

Monteiro, A. M., Tütüncü, R. H., & Vicente, L. N. (2008). Recovering risk-neutral probability density functions from options prices using cubic splines and ensuring nonnegativity. *European Journal of Operational Research, 187(2)*, 525-542.

Nikkinen, J. (2003). Normality tests of option-implied risk-neutral densities: evidence from the small Finnish market. *International Review of Financial Analysis, 12(2)*, 99–116.

Ñíguez, T-M., Paya, I., Peel, D., & Perote, J. (2012). On the stability of the constant relative risk aversion (CRRA) utility under high degrees of uncertainty. *Economics Letters, 115(2)*, 244-248.

Peña, I., Rubio, G., & Serna, G. (1999). Why do we smile? On the determinants of the implied volatility function. *Journal of Banking & Finance, 23(8)*, 1151 – 1179.

Postali, F., & Picchetti, P. (2006). Geometric Brownian Motion and structural breaks in oil prices: a quantitative analysis. *Energy Economics, 28(4)*, 506-522.

Ritchey, R. (1990). Call option valuation for discrete normal mixtures. *Journal of Financial Research, 13(4)*, 285–296.

Rompolis, L. S. (2010). Retrieving risk neutral densities from European option prices based on the principle of maximum entropy. *Journal of Empirical Finance, 17(5)*, 918-937.

Rompolis, L. S., & Tzavalis , E. (2007). Retrieving risk neutral densities based on risk neutral moments through a Gram–Charlier series expansion. *Mathematical and Computer Modelling, 46(1–2)*, 225-234.

Rubinstein, M. (1994). Implied binomial trees. *The Journal of Finance, 49(3)*, 771-818.

Shimko, D. (1993). Bounds of Probability. *RISK Magazine, 6*, 33-37.

Su, C., Li, Z., Chang, H., & Lobonț, O. (2017). When will occur the crude oil bubbles? *Energy Policy, 102*, 1-6.

Taboga, M. (2016). Option-implied probability distributions: How reliable? How jagged? *International Review Of Economics & Finance, 45*, 453-469.

Völkert, C. (2015). The distribution of uncertainty: evidence from the VIX options market. *Journal of Futures Markets, 35(7)*, 597-624.

Apéndice IV.A

Este apéndice deriva la expresión para el Delta del modelo de Black-Scholes (Δ^{BS}):

De la ecuación (IV.4) se sabe que el precio teórico para la opción call de Black-Scholes se obtiene como

$$C^{BS}(K;\theta) = S_T\Phi(d_1) - Ke^{-r\tau}\Phi(d_2),$$

donde $\Phi(d_1) = \int_{-\infty}^{d_1} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du$ y $\Phi(d_2) = \int_{-\infty}^{d_2} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du$ es la cdf de la distribución normal estándar, $d_1 = \frac{[ln(S_T/K)+(r+\sigma^2/2)\tau]}{\sigma\sqrt{\tau}}$ y $d_2 = d_1 - \sigma\sqrt{\tau}$.

Primero, se calcula

$$\frac{\partial\Phi(d_1)}{\partial d_1} = \frac{1}{\sqrt{2\pi}} e^{\frac{-d_1^2}{2}}, \tag{IV.A.1}$$

$$\frac{\partial\Phi(d_2)}{\partial d_2} = \frac{1}{\sqrt{2\pi}} e^{\frac{-d_2^2}{2}},$$

$$= \frac{1}{\sqrt{2\pi}} e^{\frac{-(d_1-\sigma\sqrt{\tau})^2}{2}}.$$

Desarrollando el binomio al cuadrado y reemplazando d_1 se

obtiene $\frac{1}{\sqrt{2\pi}} e^{\frac{-d_1^2}{2}} e^{\frac{[ln(S_T/K)+(r+\sigma^2/2)\tau]\sigma\sqrt{\tau}}{\sigma\sqrt{\tau}}} e^{\frac{\sigma^2\tau}{2}}$, de donde

$$\frac{\partial\Phi(d_2)}{\partial d_2} = \frac{1}{\sqrt{2\pi}} e^{\frac{-d_2^2}{2}} \left(\frac{S_T}{K}\right) e^{r\tau}. \tag{IV.A.2}$$

Además, se cumple que

$$\frac{\partial d_1}{\partial S_T} = \frac{\partial d_2}{\partial S_T} = \frac{1}{S_T\sigma\sqrt{\tau}}. \tag{IV.A.3}$$

El Delta de una opción se define como la derivada parcial del precio de la opción con respecto al precio del activo subyacente, para la call se tiene

$$\Delta_{call}^{BS} = \frac{\partial C^{BS}}{\partial S_T}.$$

$$\frac{\partial C^{BS}}{\partial S_T} = \Phi(d_1) + S_T \frac{\partial \Phi(d_1)}{\partial S_T} - Ke^{r\tau} \frac{\partial \Phi(d_2)}{\partial S_T},$$

$$= \Phi(d_1) + S_T \frac{\partial \Phi(d_1)}{\partial d_1} \frac{\partial d_1}{\partial S_T} - Ke^{r\tau} \frac{\partial \Phi(d_2)}{\partial d_2} \frac{\partial d_2}{\partial S_T},$$

reemplazando las ecuaciones (IV.A.1), (IV.A.2) y (IV.A.3) se obtiene

$$\Delta_{call}^{BS} = \Phi(d_1). \square \quad \text{(IV.A.4)}$$

Para la opción put se puede demostrar que

$$\Delta_{put}^{BS} = \Phi(d_1) - 1. \quad \text{(IV.A.5)}$$

Apéndice IV.B

Este apéndice deriva la expresión para el Delta del modelo SNP (Δ^{SNP}):

De la ecuación (IV.5) se sabe que el precio teórico para la opción call SNP se obtiene como

$$C^{SNP}(K;\theta) = C^{BS}(K;\theta) + S_T \phi(d_1) f(d_1),$$

$$\text{con } f(d_1) = \sigma\left[\delta_3\left(2\sqrt{\tau}\sigma - d_1\right) - \delta_4\sqrt{\tau}\left(1 - d_1^2 + 3d_1\sqrt{\tau}\sigma - 3\tau\sigma^2\right)\right], \quad \text{(IV.B.1)}$$

donde $\phi(d_1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-d_1^2}{2}}$ es la pdf de la distribución normal estándar y $d_1 = \frac{\left[ln(S_T/K) + \left(r + \sigma^2/2\right)\tau\right]}{\sigma\sqrt{\tau}}$.

Primero, se calcula

$$\frac{\partial \emptyset(d_1)}{\partial S_T} = \frac{\partial \emptyset(d_1)}{\partial d_1} \frac{\partial d_1}{\partial S_T},$$

donde $\frac{\partial \emptyset(d_1)}{\partial d_1} = -d_1 \emptyset(d_1)$ y $\frac{\partial d_1}{\partial S_T}$ es la ecuación hallada en (IV.A.3). Además se obtiene que

$$\frac{\partial \emptyset(d_1)}{\partial S_T} = \frac{-d_1 \emptyset(d_1)}{S_T \sigma\sqrt{\tau}}. \quad \text{(IV.B.2)}$$

Además, la derivada de (IV.B.1) está dada por $\frac{\partial \mathrm{f}(d_1)}{\partial S_T} = \frac{\partial \mathrm{f}(d_1)}{\partial d_1} \frac{\partial d_1}{\partial S_T}$,

donde $\frac{\partial \mathrm{f}(d_1)}{\partial d_1} = \sigma\left[-\delta_3 - \delta_4\sqrt{\tau}\left(-2d_1 + 3\sqrt{\tau}\sigma\right)\right]$ y $\frac{\partial d_1}{\partial S_T}$ es la ecuación hallada en (IV.A.3). De forma que se obtiene

$$\frac{\partial \mathrm{f}(d_1)}{\partial S_T} = \frac{\left[-\delta_3 - \delta_4\sqrt{\tau}\left(-2d_1 + 3\sqrt{\tau}\sigma\right)\right]}{S_T\sqrt{\tau}}. \qquad \text{(IV.B.3)}$$

Así el Delta del modelo SNP, para la call se define por la derivada parcial

$$\Delta_{call}^{SNP} = \frac{\partial C^{SNP}}{\partial S_T}.$$

$$\frac{\partial C^{SNP}}{\partial S_T} = \frac{\partial C^{BS}}{\partial S_T} + S_T\left(\emptyset(d_1)\frac{\partial \mathrm{f}(d_1)}{\partial S_T} + \mathrm{f}(d_1)\frac{\partial \emptyset(d_1)}{\partial S_T}\right) + \emptyset(d_1)\mathrm{f}(d_1).$$

Reemplazando las ecuaciones (IV.A.4), (IV.B.1), (IV.B.2) y (IV.B.3) se obtiene

$$\Delta_{call}^{SNP} = \Phi(d_1) - \frac{\delta_3}{\sqrt{\tau}}\emptyset(d_1)\left(1 - d_1^2 + 3d_1\sigma\sqrt{\tau} - 2\sigma^2\tau\right) + \delta_4\emptyset(d_1)\left[3d_1(1 - 2\sigma^2\tau) - d_1^3 + 4d_1^2\sigma\sqrt{\tau} - 4\sigma\sqrt{\tau} + 3\sigma^3\tau^{3/2}\right]. \square \qquad \text{(IV.B.4)}$$

Para la opción put se puede demostrar que

$$\Delta_{put}^{SNP} = \frac{\partial P^{SNP}}{\partial S_T} - 1. \qquad \text{(IV.B.5)}$$

CONCLUSIONES

En la presente tesis se proponen transformaciones logarítmicas de una distribución semi-noparamétrica (log-SNP) que son extensiones de una distribución lognormal que permiten aproximar cualquier distribución empírica mediante la introducción de parámetros adicionales. Con esta transformación se busca mantener la flexibilidad de los parámetros de las distribuciones de Gram-Charlier, pero restringiendo el dominio a valores positivos. Con el fin de evaluar la metodología propuesta, se estudiaron diferentes fenómenos económicos y financieros, cuyas series exhiben colas superiores pesadas.

El principal objetivo del capítulo II fue medir la productividad investigadora y determinar los cuantiles que delimitan a los investigadores más productivos en cada área del conocimiento mediante la distribución univariante log-SNP. Los resultados permiten concluir que los métodos tradicionales basados en distribuciones de uno o dos parámetros pueden conducir a estimaciones sesgadas, particularmente en la cola de distribución y, por lo tanto, los cuantiles calculados a partir de la distribución de productividad teórica pueden resultar en medidas de productividad engañosas. Así, en comparación con la distribución lognormal, la distribución log-SNP provee un mejor ajuste de la distribución de desempeño de los investigadores. Independientemente de los campos particulares que analizamos, se muestra que la metodología propuesta proporciona resultados más precisos y, en consecuencia, representa una herramienta útil para medir la productividad científica, particularmente interesante en contextos donde el desempeño de autores, instituciones o campos tiene que ser comparado o agregado para implementar políticas basadas en el desempeño académico.

El objetivo del capítulo III fue modelizar la distribución del tamaño de la empresa y ajustar los cuantiles superiores de la distribución mediante la distribución univariante log-SNP. Al contrastar los resultados obtenidos a partir de la distribución lognormal univariante respecto a la distribución log-SNP univariante, se concluye que ambos modelos capturan de manera adecuada los parámetros de localización, escala y, en el caso de la distribución log-SNP, los parámetros de forma son altamente significativos. Sin embargo, al analizar criterios de

bondad de ajuste de las distribuciones, los resultados sugieren que una modelización a partir de la distribución log-SNP es claramente superior. Adicionalmente, al comparar los cuantiles estimados teóricamente bajo las especificaciones lognormal y log-SNP, esta última recoge más adecuadamente, no sólo los valores en torno a la media sino, especialmente, los valores extremos. Estos resultados son de gran interés para los encargados o responsables de formular e implementar políticas de competencia en los distintos sectores. A partir de estos resultados, se puede obtener información del tamaño crítico de las empresas por cada industria, de modo que no se alcancen tamaños que puedan atentar contra la competencia de mercado.

En el capítulo III también se desarrolló, por primera vez, una expresión para la densidad de la distribución log-SNP multivariante. Especificamente se estudió el caso bivariante de las distribuciones lognormal y log-SNP. Tal como en el caso univariante, se puede concluir que aunque para ambas distribuciones todos los parámetros resultan estadísticamente significativos, la modelización a partir de distribución log-SNP bivariante resulta claramente superior. La estimación conjunta permitió concluir que al tomar las variables de ventas y total de activos como proxies de tamaño de la empresa, la primera de ellas permite modelizar de manera más fiable el comportamiento de los valores extremos en la distribución del tamaño de la empresa. Aunque son dos variables altamente correlacionadas, los activos pueden actuar como un seguro de supervivencia ante entornos económicos inestables, lo que la hace una variable más sensible a factores que pueden alterar la correcta modelación de la distribución del tamaño de la empresa.

Finalmente, el capítulo IV buscó modelizar la función de densidad neutral de riesgo (RND) log-SNP usando los precios de opciones sobre petróleo crudo West Texas Intermediate (WTI). Con el objetivo de poder obtener conclusiones sobre las mejoras obtenidas al ajustar una RND log-SNP, se modelizó también la tradicional RND lognormal derivada del modelo de Black-Scholes. A partir de los resultados se concluye que bajo una RND lognormal el precio promedio, para las opciones sobre el WTI, tiende a ser infravalorado sistemáticamente en una mayor cantidad monetaria que el precio promedio modelizado bajo una RND log-SNP. Los resultados muestran que las distribuciones implícitas en los precios de las opciones son leptocúrticas y negativamente sesgadas. Por ello, asumir una distribución lognormal como proceso estocástico del precio conlleva a valoraciones de las opciones incorrectas y

puede conducir a una gestión inadecuada del riesgo. El estudio de modelos que permitan caracterizar mejor los precios del WTI es relevante, dado que es un commodity cuya valoración ha mostrado un gran impacto en la economía mundial, especialmente en épocas de alta en volatilidad los precios del petróleo. Por lo tanto, este se convierte en un tema de interés para inversionistas y bancos centrales.

Como conclusión general se puede decir que la distribución log-SNP es un proceso de generación de datos que se adapta a toda la distribución empírica de variables económicas y financieras. Adicionalmente, es una distribución flexible que logra capturar saltos en la cola superior con precisión. De esta manera, este proceso es más fiable que el de la distribución lognormal (que está anidada en el modelo propuesto) ya que la distribución log-SNP permite modelizar el sesgo y el exceso de curtosis presente en las variables aquí estudiadas.

Como futuras investigaciones se propone lo siguiente: en el capítulo II se puede hacer el estudio de productividad teniendo en cuenta las edades de los autores. Esto puede permitir comparar el desempeño investigador por grupos de acuerdo con la trayectoria académica. Por otra parte, se pueden verificar los resultados obtenidos usando una base de datos diferente a Web of Science, por ejemplo, se podrían utilizar datos recopilados por Scopus o Google Scholar. Aunque estas bases de datos contienen un número similar de amplios campos del conocimiento, pueden permitir contrastar los resultados obtenidos en la presente tesis. Otra posibilidad es investigar la solidez de los resultados ajustando o fraccionando el número de artículos por investigador coescritos por dos o más personas.

Respecto al capítulo III, se puede modelizar la distribución del tamaño de la empresa teniendo en cuenta la dinámica de cada sector industrial, es decir, considerando los procesos de creación y desaparición de empresas. Para ello, sería necesario estudiar un periodo más extenso de investigación. También, es relevante estudiar los determinantes del crecimiento empresarial, por ejemplo, restricciones financieras, estructura de capital, coyuntura industrial o política, entre otros, con el fin de analizar el impacto de esas variables en la forma de la distribución. Estos temas se pueden enmarcar en otras economías diferentes a la de Estados Unidos, la cual se estudió en la presente tesis.

En el capítulo IV, se puede aprovechar el hecho de que se conocen fechas especiales, con noticias del mercado del petróleo y políticos, con el fin de realizar un estudio de eventos que verifique, si en esas fechas, realmente hubo un impacto significativo en los rendimientos del WTI. Adicionalmente, es interesante hacer un estudio de la evolución de los parámetros estimados con el modelo SNP, modelizando los precios de las opciones para varias fechas anteriores y posteriores a las noticias, y así ver si ellos dan señales de comportamientos adversos en el mercado.

Finalmente, determinar la viabilidad de obtener funciones de densidad de probabilidad en términos de funciones ortogonales diferentes a los polinomios de Hermite. A la luz de esos resultados, evaluar la calidad de las estimaciones de las funciones de distribución desarrolladas mediante su aplicación a la modelización de variables de los campos económico y financiero.

BIBLIOGRAFÍA

Abhyankar, A., Xu, B., & Wang, J. (2013). Oil price shocks and the stock market: evidence from Japan. *The Energy Journal, 34(2)*, 199-222.

Abramo, G., & D'Angelo, C. A. (2014). Assessing national strengths and weaknesses in research fields. *Journal of Informetrics, 8(3)*, 766–775.

Abramo, G., D'Angelo, A. C., & Pugini, F. (2008). The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology. *Scientometrics, 76(2)*, 225–244.

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover Publications.

Aguinis, H., O'Boyle, E., Gonzalez-Mulé, E., & Joo, H. (2015). Cumulative advantage: Conductors and insulators of heavy-tailed productivity distributions and productivity tars. *Personnel Psychology*, http://dx.doi.org/10.1111/peps.12095 (in press).

Ahn, H., & Powell, J. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal Of Econometrics, 58(1-2)*, 3-29.

Aitchison, J., & Brown, J. (1957). *The Lognormal Distribution.* Cambridge: Cambridge University Press.

Aït-Sahalia, Y., & Lo, A. W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *The Journal of Finance, 53(2)*, 499–547.

Albarrán, P., Juan, A. C., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics, 88(2)*, 385-397.

Allen, D., & Lunneborg, C. (1996). Modeling experimental and observational data. *Technometrics, 38(3)*, 291-291.

Anagnou-Basioudis, I., Bedendo, M., Hodges, S. D., & Tompkins, R. (2005). Forecasting accuracy of implied and GARCH-based probability density functions. *Review of Futures Markets 11(1)*, 41-66.

Antonakakis, N., Chatziantoniou, I., & Filis, G. (2017). Oil shocks and stock markets: dynamic connectedness under the prism of recent geopolitical and economic unrest. *International Review Of Financial Analysis, 50*, 1-26.

Axtell, R. (2001). Zipf distribution of U.S. Firm sizes. *Science, 293*, 1818–1820.

Backus, D., Foresi, S., Li, K., & Wu, L. (1997). Accounting for Biases in Black-Scholes. *Manuscript, The Stern School at New York University*, 1-46.

Barba Navaretti, G., Castellani, D., & Pieri, F. (2014). Age and firm growth: evidence from three European countries. *Small Business Economics, 43*(4), 823–837.

Barbosa, N., & Eiriz, V. (2011). Regional variation of firm size and growth: the Portuguese case. *Growth and Change, 42*(2), 125-158.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy, 44(2),* 451-466.

Birkmaier, D., & Wohlrabe, K. (2014). The Matthew effect in economics reconsidered. *Journal of Informetrics, 8(4),* 880–889.

Birru, J., & Figlewski, S. (2012). Anatomy of a meltdown: the risk neutral density for the S&P 500 in the fall of 2008. *Journal of Financial Markets, 15(2),* 151-180.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy, 81(3),* 637–659.

Blinnikov, S., & Moessner, R. (1998). Expansions for nearly Gaussian distributions. *Astronomy and astrophysics Supplement Series, 130(1),* 193–205.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45(1),* 199–245.

Borokhovich, K. A., Bricker, R. J., Brunarski, K. R., & Simkins, B. J. (1995). Finance research productivity and influence. *The Journal of Finance, 50(5),* 1691-1717.

Bottazzi, G., Cefis, E., Dosi, G., & Secchi, A. (2007). Invariances and diversities in the patterns of industrial evolution: some evidence from Italian manufacturing industries. *Small Business Economics, 29*(1), 137-159.

Bottazzi, G., Pirino, D., & Tamagni, F. (2015). Zipf law and the firm size distribution: a critical discussion of popular estimators. *Journal of Evolutionary Economics, 25*(3), 585–610.

Boudt, K., Peterson, B., & Croux, C. (2008). Estimation and decomposition of downside risk for portfolios with non-normal returns. *The Journal Of Risk, 11(2),* 79-103.

Box, G. (1976). Science and statistics. *Journal Of The American Statistical Association, 71(356),* 791-799.

Breeden, D., & Litzerberger, R. (1978). Prices of state-contingent claims implicit in options prices. *Journal of Business, 51(4),* 621–651.

Broadus, R. N. (1987). Toward a definition of 'bibliometrics'. *Scientometrics, 12(5-6),* 373-379.

Brown, C. A., & Robinson, D. M. (2002). Skewness and kurtosis implied by option prices: a correction. *The Journal of Financial Research, 25(2)*, 279–282.

Cabral, L. M., & Mata, J. (2003). On the evolution of the firm size distribution: facts and theory. *American Economic Review, 93*(4), 1075-1090.

Campanario, J. M. (2015). Providing impact: The distribution of JCR journals according to references they contribute to the 2-year and 5-year journal impact factors. *Journal of Informetrics, 9(2)*, 398–407.

Cefis, E., Marsili, O., & Schenk, H. (2009). The effects of mergers and acquisitions on the firm size distribution. *Journal of Evolutionary Economics, 19*(1), 1-20.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. Heckman, & E. Leamer, *Handbook of Econometrics, Vol. 6. Part B* (pp. 5549-5632). Amsterdam: Elsevier.

Christoffersen, P. (2012). *Elements of financial risk management (2nd ed.).* Waltham: Academic Press.

Christoffersen, P., & Gonçalves, S. (2005). Estimation risk in financial risk management. *The Journal Of Risk, 7(3)*, 1-28.

Christoffersen, P., Heston, S., & Jacobs, K. (2013). Capturing option anomalies with a variance-dependent pricing kernel. *Review of Financial Studies, 26(8)*, 1963-2006.

Chung, K. H., & Cox, R. A. (1990). Patterns of productivity in the finance literature: a study of the bibliometric distributions. *The Journal of Finance, 45(1) 1, 301-309*, 301-309.

Cirillo , P., & Hüsler, J. (2009). On the upper tail of Italian firms' size distribution. *Physica A: Statistical Mechanics and its Applications, 388*(8), 1546-1554.

Coad, A. (2010). The exponential age distribution and the Pareto firm size distribution. *Journal of Industry, Competition and Trade, 10*(3), 389–395.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance, 1(2)*, 223-236.

Corrado, C. J., & Su, T. (1996). Skewness and kurtosis in S&P 500 index returns implied by option prices. *The Journal of Financial Research, 19(2)*, 175-192.

Corrado, C. J., & Su, T. (1997). Implied volatility skews and stock return skewness and kurtosis implied by stock option prices. *The European Journal of Finance, 3(1)*, 73–85.

Cortés, L. M., Mora-Valencia, A., & Perote, J. (2016). The productivity of top researchers: a semi-nonparametric approach. *Scientometrics, 109*(2), 891-915.

Coupé, T. (2003). Revealed performances. Worldwide rankings of economists and economics departments. *Journal of the European Economic Association, 1(6)*, 1309–1345.

Cox, J., & Ross, S. (1976). The valuation of options for alternative stochastic processes.,. *Journal Of Financial Economics, 3(1-2)*, 145-166.

Cramér, H. (1925). On some classes of series used in mathematical statistics. *Sixth Scandinavian Congress of Mathematicians*, (págs. 399-425). Copenhagen.

Crespo, J. A., Ortuño-Ortín, I., & Ruiz-Castillo, J. (2012). The citation merit of scientific publications. *PLoS ONE 7(11)*, e49156.

Crosato, L., & Ganugi, P. (2007). Statistical regularity of firm size distribution: the Pareto IV and truncated Yule for Italian SCI manufacturing. *Statistical Methods and Applications, 16*(1), 85-115.

Da Silva, R., Kalil, F., De Oliveira, J. M., & Martinez, A. S. (2012). Universality in bibliometrics. *Physica A: Statistical Mechanics and its Applications, 391(5)*, 2119-2128.

Das, S. R., & Sundaram, R. K. (1999). Of smiles and smirks: a term structure perspective. *The Journal of Financial and Quantitative Analysis, 34(2)*, 211-239.

Day, T. E. (2015). The big consequences of small biases: A simulation of peer review. *Research Policy, 44(6)*, 1266–1270.

de Souza e Silva, E. A., Legey, L. F., & de Souza e Silva, E. A. (2010). Forecasting oil price trends using wavelets and hidden Markov models. *Energy Economics, 32(6)*, 1507-1519.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2007). *A modern introduction to probability and statistics: understanding why and how*. London: Springer.

Del Brio, E. B., & Perote, J. (2012). Gram–Charlier densities: Maximum likelihood versus the method of moments. *Insurance: Mathematics and Economics, 51(3)*, 531-537.

Del Brio, E., Mora-Valencia, A., & Perote, J. (2014a). Semi-nonparametric VaR forecasts for hedge funds during the recent crisis. *Physica A: Statistical Mechanics And Its Applications, 401*, 330-343.

Del Brio, E., Mora-Valencia, A., & Perote, J. (2014b). VaR performance during the subprime and sovereign debt crises: An application to emerging markets. *Emerging Markets Review, 20*, 23-41.

Del Brio, E., Mora-Valencia, A., & Perote, J. (2017). The kidnapping of Europe: high-order moments' transmission between developed and emerging markets. *Emerging Markets Review*, In press.

Delmar, F., Davidsson, P., & Gartner, W. B. (2003). Arriving at the high-growth firm. *Journal of Business Venturing, 18*(2), 189-216.

Dennis, P., & Mayhew, S. (2002). Risk-neutral skewness: evidence from stock options. *Journal of Financial and Quantitative Analysis, 37(3)*, 471–493.

di Giovanni, J., Levchenko, A. A., & Rancière. , R. (2011). Power laws in firm size and openness to trade: measurement and implications. *Journal of International Economics, 85*(1), 42-52.

Dosi, G., Marsili, O., Orsenigo, L., & Salvatore, R. (1995). Learning, marketselection and the evolution of industrial structures. *Small BusinessEconomics, 7*(6), 411–436.

Du, Y., Wang, C., & Du, Y. (2012). Inversion of option prices for implied risk-neutral probability density functions: general theory and its applications to the natural gas market. *Quantitative Finance, 12(12)*, 1877-1891.

Duch, J., Zeng, X. T., Sales-Pardo, M., Radicchi, F., Otis , S., Woodruff, T. K., & Nunes Amaral, L. A. (2012). The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS ONE 7(12)*, e51332.

Dundar, H., & Lewis, D. (1998). Determinants of research productivity in higher education. *Research in Higher Education, 39(6)*, 607-631.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics.* Kidlington, UK: Elsevier Academic Press.

Ellison, G. (2013). How does the market use citation data? the hirsch index in economics. *American Economic Journal: Applied Economics, 5(3)*, 63-90.

Eom, Y. H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE, 6(9)*, e24926.

Fabozzi, F. J., Tunaru, R., & Albota, G. (2009). Estimating risk-neutral density with parametric models in interest rate markets. *Quantitative Finance, 9(1)*, 55-70.

Fama, E. (1965). The behaviour of stock market prices. *Journal of Business, 38(1)*, 34–105.

Feng, P., & Dang, C. (2016). Shape constrained risk-neutral density estimation by support vector regression. *Information Sciences, 333(C)*, 1-9.

Finardi, U. (2013). Correlation between Journal Impact Factor and Citation Performance: An experimental study. *Journal of Informetrics, 7(2)*, 357–370.

Frandsen, T. F. (2005). Geographical concentration. The case of economics journals. *Scientometrics, 63(1)*, 69-85.

Friesen, G., Zhang, Y., & Zorn, T. (2012). Heterogeneous beliefs and risk-neutral skewness. *Journal of Financial and Quantitative Analysis, 47(4)*, 851-872.

Fusai, G., & Roncoroni, A. (2008). *Implementing models in quantitative finance (1st ed.).* Berlin: Springer.

Gabaix, X., Lasry, J. M., Lions, P. L., & Moll, B. (2016). The dynamics of inequality. *Econometrica, 84(6)*, 2071–2111.

Gaffeo, E., Gallegati, M., & Palestrini, A. (2003). On the size distribution of firms: additional evidence from the G7 countries. *Physica A: Statistical Mechanics and its Applications, 324*(1-2), 117-123.

Gallant, A. R., & Nychka, D. W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica, 55(2)*, 363–390.

Garfield, E. (1980). Bradford's Law and related statistical pattern. *Essays of an Information Scientist, 4(19)*, 476-483.

Genest, C. (1997). Statistics on statistics: measuring research productivity by journal publications between 1985 and 1995. *The Canadian Journal of Statistics, 25(4)*, 427-443.

Gibrat, R. (1931). *Les Inégalités Economiques.* Paris: Recueil Sirey.

Goddard, J., Liu, H., Donal, M., & Wilson, J. O. (2014). The size distribution of US banks and credit unions. *International Journal of the Economics of Business, 21*(1), 139-156.

Guerrero-Bote, V. P., Zapico-Alonso, F., Espinosa-Calvo, M. E., Gomez-Crisostomo, R., & Moya-Anegon, F. (2007). Import–export of knowledge between scientific subject categories: The iceberg hypothesis. *Scientometrics, 71(3)*, 423–441.

Gupta, H. M., Campanha, J. R., de Aguiar, D. R., Queiroz, G. A., & Raheja, C. G. (2007). Gradually truncated log-normal in USA publicly traded firm size distribution. *Physica A: Statistical Mechanics and its Applications, 375*(2), 643-650. doi:http://dx.doi.org/10.1016/j.physa.2006.09.025

Hart, P. E., & Oulton, N. (1997). Zipf and the size distribution of firms. *Applied Economics Letters, 4*(4), 205-206.

Hartvig, N. V., Jensen, J. L., & Pedersen, J. (2001). A class of risk neutral densities with heavy tails. *Finance and Stochastics, 5(1)*, 115-128.

Harzing, A. (2008). Publish or Perish: A citation analysis software program. *Available from http://www.harzing.com/resources.htm.*

He, A., Kwok, J., & Wan, A. (2010). An empirical model of daily highs and lows of West Texas Intermediate crude oil prices. *Energy Economics, 32(6)*, 1499-1506.

Heberger, A. E., Christie, C. A., & Alkin, M. C. (2010). A bibliometric analysis of the academic influences of and on evaluation theorists' published works. *American Journal of Evaluation, 31(1)*, 24-44.

Heinrich, T., & Dai, S. (2016). Diversity of firm sizes, complexity, and industry structure in the Chinese economy. *Structural Change and Economic Dynamics, 37*, 90-106.

Hernández-Pérez, R. (2010). An analogy of the size distribution of business firms with Bose–Einstein statistics. *Physica A: Statistical Mechanics and its Applications, 389*(18), 3837-3843.

Hodgson, G. M., & Rothman, H. (1999). The editors and authors of economics journals: A case of institutional oligopoly? *The Economic Journal, 109(453)* , 165–186.

Hsieh, D. (1988). The statistical properties of daily foreign exchange rates: 1974–1983. *Journal Of International Economics, 24(1-2)*, 129-145.

Huang, D., Yu, B., Fabozzi, F., & Fukushima, M. (2009). CAViaR-based forecast for oil price risk. *Energy Economics, 31(4)*, 511-518.

Jackwerth, J., & Rubinstein, M. (1996). Recovering probability distributions from options prices. *Journal of Finance, 51(5)*, 1611–1631.

Jarrow, R., & Rudd, A. (1982). Approximate option valuation for arbitrary stochastic processes. *Journal of Financial Economics,10(3)*, 347-369.

Jondeau, E., & Rockinger, M. (2000). Reading the smile: the message conveyed by methods which infer risk neutral densities. *Journal of International Money and Finance, 19(6)*, 885–915.

Jondeau, E., & Rockinger, M. (2001). Gram-Charlier densities. *Journal of Economic Dynamics & Control, 25*(10), 1457-1483.

Jondeau, E., & Rockinger, M. (2006). Optimal portfolio allocation under higher moments. *European Financial Management, 12(1)*, 29-55.

Jondeau, E., Poon, S.-H., & Rockinger, M. (2007). *Financial Modeling Under Non-Gaussian Distributions.* London: Springer.

Jurczenko, E., Maillet, B., & Negrea, B. (2004). A note on skewness and kurtosis adjusted option pricing models under the Martingale restriction. *Quantitative Finance, 4(5)*, 479-488.

Kaizoji, T., Iyetomi, H., & Ikeda, Y. (2006). Re-examination of the size distribution of firms. *Evolutionary and Institutional Economics Review, 2*(2), 183–198.

Kallis, G., & Sager, J. (2017). Oil and the economy: a systematic review of the literature for ecological economists. *Ecological Economics, 131*, 561-571.

Kaur, J., Ferrara, E., Menczer, F., Flammini, A., & Radicchi, F. (2015). Quality versus quantity in scientific impact. *Journal of Informetrics, 9(4)*, 800-808.

Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics, 7(4)*, 924–932.

Kendall, M., & Stuart, A. (1977). *The Advanced Theory of Statistics, Vol. I, 4th ed.* London: C. Griffin.

Kiesel, R., & Rahe, F. (2017). Option pricing under time-varying risk-aversion with applications to risk forecasting. *Journal Of Banking & Finance, 76(C)*, 120-138.

Kocher, M. G., Luptacik, M., & Sutter, M. (2006). Measuring productivity of research in economics: A cross-country study using DEA. *Socio-Economic Planning Sciences, 40(4)*, 314-332.

Kretschmer, H., & Kretschmer, T. (2007). Lotka's distribution and distribution of co-author pairs' frequencies. *Journal of Informetrics, 1(4)*, 308–337.

Kuhs, W. F. (1988). The anharmonic temperature factor in crystallographic structure analysis. *Australian Journal of Physics, 41*(3), 369-382.

Kumar, S., Sharma, P., & Garg, K. C. (1998). Lotka's law and institutional productivity. *Information Processing & Management, 34(6)*, 775–783.

Lai, W.-N. (2014). Comparison of methods to estimate option implied riskneutral densities. *Quantitative Finance, 14(10)*, 1839-1855.

Lancho-Barrantes, B. S., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). The iceberg hypothesis revisited. *Scientometrics, 85(2)*, 443–461.

Leippold, M., & Schärer, S. (2017). Discrete-time option pricing with stochastic liquidity. *Journal Of Banking & Finance, 75*, 1-16.

León, A., Mencía, J., & Sentana, E. (2009). Parametric properties of semi-nonparametric distributions, with applications to optionvaluation. *Journal of Business & Economic Statistics, 27(2)*, 176-192.

Lim, G. C., Martin, G. M., & Martin, V. L. (2005). Parametric pricing of higher order moments in S&P500 options. *Journal of Applied Econometrics,20(3),* 377–404.

Liu, X. (2007). Bid–ask spread, strike prices and risk-neutral densities. *Applied Financial Economics, 17(11),* 887–900.

Liu, X., Shackleton, M. B., Taylor, S. J., & Xu, X. (2007). Closed-form transformations from risk-neutral to real-world distributions. *Journal of Banking & Finance, 31(5),* 1501-1520.

Loretan, M., & Phillips, P. (1994). Testing the covariance stationarity of heavy-tailed time series: an overview of the theory with applications to several financial datasets. *Journal Of Empirical Finance, 1(2),* 211-248.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science, 16(12),* 317-323.

MacBeth, J., & Merville, L. (1979). An empirical examination of the Black-Scholes call option pricing model. *Journal of Finance, 34(5),* 1173-1186.

Mandelbrot, B. (2003). Heavy tails in finance for independent or multifractal price increments. In S. Rachev, *Handbook of heavy tailed distributions in finance (1st ed.)* (pp. 4-32). Amsterdam: Elsevier.

Marsili, O. (2006). Stability and turbulence in the size distribution of firms: evidence from Dutch manufacturing. *International Review of Applied Economics, 20*(2), 255-272.

Martínez-Mekler, G., Martínez, R. A., del Río, M. B., Mansilla, R., Miramontes, P., & Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE, 4*(3), e4791.

Mauleón, I. (1997). Instability and long memory in conditional variances. *Journal de la Societe de Statistique de Paris, 134(4),* 67-88.

Mauleón, I., & Perote, J. (2000). Testing densities with financial data: an empirical comparison of the Edgeworth-Sargan density to the Student's t. *European Journal of Finance, 6*(2), 225-239.

Melick, W. R., & Thomas, C. P. (1997). Recovering an asset's implied PDF from option prices: an application to crude oil during the Gulf crisis. *Journal of Financial and Quantitative Analysis, 32(1),* 91–115.

Merton, R. (1971). Optimum consumption and portfolio rules in a continuous-time model. *Journal Of Economic Theory, 3(4),* 373-413.

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research, 246(1),* 1-19.

Monteiro, A. M., Tütüncü, R. H., & Vicente, L. N. (2008). Recovering risk-neutral probability density functions from options prices using cubic splines and ensuring nonnegativity. *European Journal of Operational Research, 187(2)*, 525-542.

Newman, M. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323-351.

Nicholls, P. T. (1989). Bibliometric modelling processes and the empirical validity of Lotka's law. *Journal of the American Society for Information Science, 40(6)*, 379–385.

Nicholls, T. P. (1986). Empirical validation of Lotka's law. *Information Processing & Management, 22(5)*, 417–419.

Nicolaisen, J., & Hjørland, B. (2007). Practical potentials of Bradford's law: a critical examination of the received view. *Journal of Documentation, 63(3)*, 359 - 377.

Nikkinen, J. (2003). Normality tests of option-implied risk-neutral densities: evidence from the small Finnish market. *International Review of Financial Analysis, 12(2)*, 99–116.

Ñíguez, T-M., & Perote, J. (2016). Multivariate moments expansion density: Application of the dynamic equicorrelation model. *Journal Of Banking & Finance, 72(S)*, S216-S232.

Ñíguez, T-M., Paya, I., Peel, D., & Perote, J. (2012). On the stability of the constant relative risk aversion (CRRA) utility under high degrees of uncertainty. *Economics Letters, 115*(2), 244-248.

Ñíguez, T-M., Paya, I., Peel, D., & Perote, J. (2013). Higher-order moments in the theory of diversification and portfolio composition. *Economics Working Paper Series 2013/003. Lancaster University*.

O'Boyle, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology, 65(1)*, 79–119.

Pascoal, R., Augusto, M., & Monteiro, A. M. (2016). Size distribution of Portuguese firms between 2006 and 2012. *Physica A: Statistical Mechanics and its Applications, 458*, 342-355.

Peña, I., Rubio, G., & Serna, G. (1999). Why do we smile? On the determinants of the implied volatility function. *Journal of Banking & Finance, 23(8)*, 1151 – 1179.

Perc, M. (2010). Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. *Journal of Informetrics, 4(2)*, 358–364.

Perote, J. (2004). The multivariate Edgeworth-Sargan density. *Spanish Economic Review, 6*(1), 77-96.

Perote, J., & Del Brio, E. (2003). Measuring Value-at-Risk under the conditional Edgeworth-Sargan distribution. *Finance Letters, 1(3)*, 23-40.

Phillips, P. B. (1977). A general theorem in the theory of asymptotic expansions as approximations to the finite sample distributions of econometric estimators. *Econometrica, 45(6)*, 1517–1534.

Plerou, V., Gopikrishnan, P., Nunes Amaral, L., Gabaix, X., & Eugene Stanley, H. (2000). Economic fluctuations and anomalous diffusion. *Physical Review E, 62(3)*, R3023-R3026.

Postali, F., & Picchetti, P. (2006). Geometric Brownian Motion and structural breaks in oil prices: a quantitative analysis. *Energy Economics, 28(4)*, 506-522.

Price, D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27(5)*, 292–306.

Radicchi, F., Fortunado, S., & Castellano, C. (2008). Universality of citation distribution: Towards an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America, 105(45)*, 17268–17272.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems, 4(2)*, 131-134.

Ritchey, R. (1990). Call option valuation for discrete normal mixtures. *Journal of Financial Research, 13(4)*, 285–296.

Rompolis, L. S. (2010). Retrieving risk neutral densities from European option prices based on the principle of maximum entropy. *Journal of Empirical Finance, 17(5)*, 918-937.

Rompolis, L. S., & Tzavalis , E. (2007). Retrieving risk neutral densities based on risk neutral moments through a Gram–Charlier series expansion. *Mathematical and Computer Modelling, 46(1–2)*, 225-234.

Rousseau, R. (1994). Bradford curves. *Information Processing and Management, 30(2)*, 267–277.

Rubinstein, M. (1994). Implied binomial trees. *The Journal of Finance, 49(3)*, 771-818.

Rubinstein, M. (1998). Edgeworth binomial trees. *Journal of Derivatives, 5(3)*, 20-27.

Ruiz-Castillo, J., & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics, 8(4)*, 917–934.

Ruppert, D. (2010). *Statistics and data analysis for financial engineering.* New York: Springer-Verlag Inc.

Sabharwal, M. (2013). Comparing research productivity across disciplines and career stages. *Journal of Comparative Policy Analysis: Research and Practice, 15(2)*, 141-163.

Sargan, D. (1975). Gram-Charlier approximation applied t ratios or k-class estimatiors. *Econometrica, 43(2)*, 327-346.

Seggie, S. H., & Griffith, D. A. (2009). What does it take to get promoted in marketing academia? Understanding exceptional publication productivity in the leading marketing journals. *Journal of Marketing, 73(1)*, 122-132.

Shimko, D. (1993). Bounds of Probability. *RISK Magazine, 6*, 33-37.

Simon, H., & Bonini, C. (1958). The size distribution of business firms. *The American Economic Review, 48*(4), 607-617.

Stanley, M. H., Buldyrev, S. V., Havlin, S., Mantegna, R. N., Salinger, M. A., & Stanley, H. E. (1995). Zipf plots and the size distribution of firm. *Economics Letters, 49*(4), 453-457.

Stein, E., & Stein, J. (1991). Stock price distributions with stochastic volatility: an analytic approach. *Review Of Financial Studies, 4(4)*, 727-752.

Su, C., Li, Z., Chang, H., & Lobonț, O. (2017). When will occur the crude oil bubbles? *Energy Policy, 102*, 1-6.

Taboga, M. (2016). Option-implied probability distributions: How reliable? How jagged? *International Review Of Economics & Finance, 45*, 453-469.

Verhoeven, P., & McAleer, M. (2004). Fat tails and asymmetry in financial volatility models. *Mathematics And Computers In Simulation, 64(3-4)*, 351-361.

Voit, J. (2001). The growth dynamics of German business firms. *Advances in Complex Systems, 4*(1), 149-162.

Völkert, C. (2015). The distribution of uncertainty: evidence from the VIX options market. *Journal of Futures Markets, 35(7)*, 597-624.

Wallace, D. L. (1958). Asymptotic approximations to distributions. *Annals of Mathematical Statistics, 29(3)*, 635-654.

Williamson, I. O., & Cable, D. M. (2003). Predicting early career research productivity: The case of management faculty. *Journal of Organizational Behavior, 24(1)*, 25-44.

Zhang , J., Chen, Q., & Wang, Y. (2009). Zipf distribution in top Chinese firms and an economic explanation. *Physica A: Statistical Mechanics and its Applications, 388*(10), 2020-2024.

Zhao, Z. (2008). Parametric and nonparametric models and methods in financial econometrics. *Statistics Surveys, 2(0)*, 1-42.