Andrés García Medina*

El uso de *Twitter* en el análisis financiero: aproximación desde la econofísica

Resumen | Se han utilizado técnicas matemáticas provenientes de la física estadística, especialmente de la teoría de matrices aleatorias (RMT, por sus siglas en inglés), para analizar datos textuales de *Twitter* en el contexto de los mercados financieros globales. Para esto, se ha analizado un periodo de tiempo de 7 meses a lo largo de 2014, considerando los retornos de 20 índices financieros globales para comparar los resultados. La información textual se logró extraer mediante el ensamblaje de distintos lenguajes de programación, construyendo series de tiempo de polaridad mediante el análisis de sentimiento. RMT reveló que existen correlaciones verdaderas en los índices financieros y las polaridades. Además, se encontró una buena concordancia entre el comportamiento temporal de los eigenvalores extremos de los retornos y polaridades, con resultados similares para la razón de participación inversa, lo cual nos da información acerca de la emergencia de factores comunes en la información financiera global, sin importar si estamos utilizando polaridades o retornos como fuente de datos. Nuestros resultados sugieren que al utilizar las polaridades de *Twitter* y *NYT* como un nuevo indicador financiero, proveen de información relevante acerca del comportamiento colectivo de los índices financieros globales. Esto genera una fuerte y novedosa evidencia en contra de la hipótesis de mercado eficiente, y apoya la tendencia de la economía conductual, la cual afirma que los precios de los mercados se ven afectados por las decisiones irracionales de los inversionistas, siendo estos influenciados por la tendencia de las noticias y redes sociales.

The use of Twitter in financial analysis: An approach from the econophysics

Abstract | Mathematical techniques from statistical physics, especially from random matrix theory (RMT), have been used to analyze textual data from Twitter in the context of global financial markets. For this, we have analyzed a period of time of 7 months along

Recibido: 23 de mayo de 2017. Aceptado: 21 de junio de 2017.

* Doctor en ciencias (física) por la Universidad de Sonora. Actualmente es profesor de asignatura en la Facultad de Ciencias de la Universidad Autónoma de Baja California, (UABC).

** Versión adaptada de [1].

Correos-e: andres.garcia.medina@uabc.edu.mx | andgarm@gmail.com

Apellidos, Nombres. «El uso de *Twitter* en el análisis financiero: aproximación desde la econofísica.» *Interdisciplina* 5, nº 12 (mayo–agosto 2017): 23–39.
DOI: http://dx.doi.org/10.22201/ceiich.24485705e.2017.12.61469

Volumen 5 | número 12 | mayo–agosto 2017
DOI: http://dx.doi.org/10.22201/ceiich.24485705e.2017.12.61469

2014, considering the returns of 20 global financial indices in order to compare the results. Textual information was extracted by assembling different programming languages, constructing time series of polarity through sentiment analysis. RMT revealed that there are true correlations in financial indices and polarities. In addition, a good concordance was found between the temporal behavior of the extreme eigenvalues of returns and polarities, with similar results for the inverse participation ratio, which gives us information about the emergence of common factors in global financial information, regardless if we are using polarities or returns as data source. Our results suggest that using polarity as a new financial indicator provide of useful information about collective and even individual behavior of global financial indices. This builds a strong and novel evidence against the efficient market hypothesis, and supporting the school of behavioral finance where the market prices are affected by the irrational decisions of investors, which are influenced by trending news and social networks.

Palabras clave | teoría de matrices aleatorias, *Twitter*, análisis de sentimiento, economía conductual

Key Words | random matrix theory, Twitter, sentiment analysis, behavioral finance

Introducción

Durante los últimos años ha surgido una enorme contribución de la física teórica hacia el entendimiento de los sistemas económicos, dando lugar a la emergencia de un nuevo campo de estudio denominado econofísica [2-4]. Dentro de esta área, entender la estructura de las correlaciones entre diferentes mercados financieros es una de las líneas de investigación que más rápidamente crece debido a la gran importancia en el contexto de la optimización de portafolios [5]. Un nuevo enfoque para entender este tipo de correlaciones viene de la teoría de matrices aleatorias (RMT, por sus siglas en inglés), la cual fue introducida en estadística matemática por Wishart en 1928 [6]. En la década de 1950, Wigner utiliza RMT para lidiar con la estadística de eigenvalores y eigenvectores de sistemas complejos de muchos cuerpos en el contexto de la física nuclear [7, 8, 9,10]. Muchos fenómenos de la física han sido resueltos exitosamente utilizando el formalismo de RMT, pero no fue hasta la aparición de los trabajos casi simultáneos de Stanley *et al.* [11] y Bouchoud *et al.* [12] que se incrementó considerablemente la cantidad de estudios dedicados a entender la estructura de los mercados financieros a través de la aplicación de métodos provenientes de RMT [13, 14, 15, 16, 17].

Por otro lado, la influencia de las noticias financieras y de las redes sociales no ha sido explorada exhaustivamente debido a la Hipótesis del Mercado Eficiente (EMH, por sus siglas en inglés). De acuerdo con la EMH, el precio de la acción

incorpora instantáneamente toda la información disponible del mercado, y su valor no depende del precio en el pasado [17]. No obstante, recientemente una serie de trabajos han comenzado a investigar la influencia de las fuentes textuales de Internet en los movimientos de los mercados [18-25], mostrando que la información extraída de *Twitter*, *StockTwits*, *Google Trends*, y la revista financiera *Financial Times*, dan indicaciones tempranas que pueden ayudar a predecir cambios en la bolsa de valores. Estos nuevos resultados están construyendo un fuerte sustento en contra del tan aceptado paradigma de mercado eficiente, apoyando la aproximación de la economía conductual [26]. En ese sentido, los resultados de este trabajo intentan mostrar una evidencia más en contra de este paradigma desde una nueva perspectiva.

La aproximación que se seguirá aquí para analizar los mercados financieros globales se sustenta en los resultados de RMT. Asimismo, en este estudio se incorpora la influencia de la red social *Twitter* mediante el análisis de sentimiento, el cual se basa en asignar un valor numérico a los datos textuales de acuerdo a su contenido. Aquí se utilizó la polaridad como medida del estado de ánimo, relacionando el puntaje obtenido con el precio al cierre de las bolsas de valores estudiadas.

Nuestro método de análisis se desarrolla desde el marco conceptual de la física estadística y de los sistemas complejos, ya que intentamos descubrir propiedades emergentes que escapan a los analistas financieros en econometría. Sin embargo, cuando se intentan crear paralelismos entre la física estadística y los mercados financieros, una cuestión importante que se debe tener en cuenta siempre es la complejidad del comportamiento humano, el cual es el origen de toda estrategia de comercio en la bolsa de valores [4].

Este trabajo está organizado como sigue: se describen los datos analizados, así como la metodología para extraer la colección pública de *tweets*; se explica cómo se llevó a cabo el análisis de sentimiento, y se describe el método seguido para construir las series de tiempo de polaridad. Asimismo, se muestran los resultados y análisis de los mismos, comenzando con la construcción de la matriz de correlación y terminando con el uso de la teoría de matrices aleatorias en el contexto del análisis multivariante. Finalmente, se da una conclusión general del trabajo.

Datos analizados

Nuestro análisis se llevó a cabo para dos conjuntos diferentes de datos. El primer conjunto de datos está compuesto por los precios diarios al cierre de 20 índices financieros alrededor del mundo. Los países donde cotizan estos índices, así como los símbolos correspondientes están listados en las dos primeras

Volumen 5 | número 12 | mayo–agosto 2017
DOI: http://dx.doi.org/10.22201/ceiich.24485705e.2017.12.61469

columnas de las tabla 1. El segundo conjunto de datos fue obtenido mediante la extracción de *tweets* asociados con cada uno de los índices financieros listados en el primer conjunto. Todas las consultas de *Twitter* fueron hechas en el horario universal (UTC), mientras que la consulta de los precios al cierre varía de acuerdo con la zona horaria donde cotizan los mercados involucrados de cada índice. Para ambos conjuntos de datos el periodo de tiempo bajo estudio comprendió del 22 de febrero al 13 de octubre de 2014, dando un total de L = 166 días de negociación, es decir, sin considerar los fines de semana. Los datos de los precios al cierre fueron obtenidos de la base de datos de *Bloomberg*, y siguieron el mismo preprocesamiento que en [27].

Por otro lado, todos los *tweets* fueron extraídos de la base de datos de *Twitter* mediante la interface *Twitter Search API*. Además, se utilizó un código *wrapper* en el lenguaje PYTHON para manejar de manera más eficiente los métodos del API de *Twitter*. Los *keyword* utilizados para obtener la colección de datos de *Twitter* se muestran en la tabla 1. Cabe resaltar que se obtuvieron aproximadamente 2,400 *tweets* por *keyword* por día, extrayendo en total cerca de 8 millones de *tweets* para este análisis.

Tabla 1: Lista de los índices financieros analizados en este trabajo.

País	Símbolo de *Bloomberg*	*Keyword Twiter*
México	MEXBOL	IPC_MEXICO
Estados Unidos	SPX	SP500
Argentina	MERVAL	MERVAL
Brasil	IBOV	IBOVESPA
Reino Unido	UKX	FTSE_UK_INDEX
Francia	CAC	CAC_40
Suiza	SMI	SWISS_MARKET_INDEX
Alemania	DAX	DAX_INDEX
Austria	ATX	ATX_INDEX
Egipto	CASE	EGX_EGYPT
Israel	TA-25	TEL_AVIV_STOCK
India	SENSEX	BSE_SENSEX
Indonesia	JCI	JAKARTA_STOCK
Malasia	FBMKLCI	BURSA_LALAYSIA
Singapur	FSSTI	STRAITS_TIMES_INDEX
Hong Kong	HSI	HANG_SENG_INDEX
Taiwan	TWSE	TAIWAN_STOCK
Korea del Sur	KOSPI	KOSPI
Japón	NKY	NIKKEI_INDEX
Australia	AS51	ALL_ORDINARIES

Notas: Primera columna: países donde cotizan los índices. Segunda columna: símbolo correspondiente en el sistema *Bloomberg*. Tercera columna: palabras clave elegidas para hacer la búsqueda en *Twitter*.
Fuente: Elaboración propia.

Análisis de sentimiento

El análisis de sentimiento es un campo de estudio del procesamiento de lenguaje natural, minería de opiniones, y lingüística computacional [28, 29]. El método de estados emocionales y el de polaridad son las dos aproximaciones principales para calificar numéricamente los datos textuales. Nosotros utilizamos la aproximación de polaridad debido a que se puede asociar de manera más directa con los movimientos positivos y negativos de los índices financieros. La polaridad, en esencia, mide la diferencia entre el número de palabras positivas y negativas encontradas en datos textuales, y está dada por la fórmula:

$$polarity = \frac{p - n}{p + n},$$

donde p y n se refieren al total de palabras positivas y negativas, respectivamente.

Por otro lado, la red social *Twitter* permite a sus usuarios mandar y recibir mensajes cortos de hasta 140 caracteres, a los cuales se les conoce como *tweets*. Un *tweet* es una combinación de caracteres intercalados con espacios en blancos, donde cada cadena de caracteres está compuesta por una secuencia alfanumérica, mezclada con caracteres especiales como son: @, $, #, %, &, etc. Por lo que cada *tweet* necesita de un preprocesamiento para eliminar los caracteres no deseados, los cuales introducen ruido para interpretar el significado de la información contenida en su texto. El primer paso es dividir cada uno de los *tweets* en las cadenas de caracteres que lo componen, evitando los símbolos no alfanuméricos. Como resultado se obtiene una colección de palabras individuales. A esta operación se le conoce como *tokenización*. Una vez que el *tweet* es *tokenizado*, cada elemento de la colección de palabras se simplifica o reduce mediante el método de *stemming*, el cual consiste en eliminar los afijos morfológicos dejando solamente la raíz de la palabra. Al finalizar esta etapa, el *tweet* ya se encuentra listo para ser categorizado por el análisis de sentimiento.

En nuestro estudio, el análisis de sentimiento se realizó con ayuda del código PYSENTIMENT$_3$. Este código implementa el diccionario *Harvard IV*, el cual ha tenido éxito en predecir el desempeño de los mercados de valores [27, 28]. Este diccionario está compuesto de más de 8000 palabras y 182 categorías. Al considerar solamente las categoría *positive* y *negative* fue posible crear un conjunto de series de tiempo de polaridad a partir de la colección publica de *tweets*. Cada una de estas series de tiempo asociada a un único *keyword* de la tabla 1, por lo que el conjunto está compuesto de 20 series de tiempo. Estas series de tiempo se construyeron como sigue. Primeramente, la colección pública de *tweets* se clasificó por *keyword* y fecha en una base de datos. En seguida, a cada *tweet* de la base de datos se le calculó la polaridad mediante de arriba. Después, se calculó

Volumen 5 | número 12 | mayo–agosto 2017
DOI: http://dx.doi.org/10.22201/ceiich.24485705e.2017.12.61469

el promedio de todas las polaridades para cada día y *keyword* dado. Finalmente, a este promedio se le consideró la polaridad $P_k(t)$ del *keyword* k al tiempo t, donde las unidades de tiempo se eligieron en días. En la tabla 2 se muestra esquemáticamente el proceso de cálculo de polaridad para un *tweet* individual.

Tabla 2. Ejemplo del proceso de cálculo de polaridad para un *tweet* individual.

Proceso	Resultado
Tweet	"#SP500 closed near the lows for the week: Expecting more downside Monday. $SPX"
Tokenización	{SP500, closed, near, the, lows, for, the, week, Expecting, more, downside, Monday, SPX}
Stemming	{sp500, closed, near, the, low, for, the, week, expect, more, downsid, monday, spx}
Categorización	{0, 0, 0, 0, negative, 0, 0, 0, 0, 0, negative, 0, 0}
Polaridad	−1

Fuente: Elaboración propia.

Análisis y resultados

En esta sección se presentan algunas técnicas matemáticas provenientes de la física estadística que nos ayudarán a entender la estructura de las matrices de correlación asociadas a nuestros datos empíricos, es decir, los datos provenientes de *Twitter* y de los índices financieros globales. Para los primeros, en lugar de los datos crudos se utilizaron las series de tiempo de polaridad, mientras que para los segundos datos se usaron los retornos diarios al cierre. Además, puesto que en general los mercados financieros no cotizan los fines de semana, se ajustaron las series de tiempo de polaridad a los días de cotización de los índices financieros, desfasando además sus valores por un día.

Matrices de correlación

Denotemos por $S_k(t)$ el precio al cierre del índice k al día t. Los retornos $R_k(t)$ para cada índice k = 1, … , 20 al tiempo t se obtienen mediante:

$$R_k(t) = \frac{S_k(t + \Delta t) - S_k(t)}{S_k(t)},$$

donde se eligió $t = 1$, tal que el intervalo de retorno sea de un día. Además, con el propósito de comparar nuestros datos empíricos con los resultados provenientes de la física estadística, las series de tiempo de polaridad y retornos son normalizadas. El retorno normalizado para el índice k al tiempo t está dado por:

$$r_k(t) = (R_k(t) - \langle R_k \rangle)/\sigma_k,$$

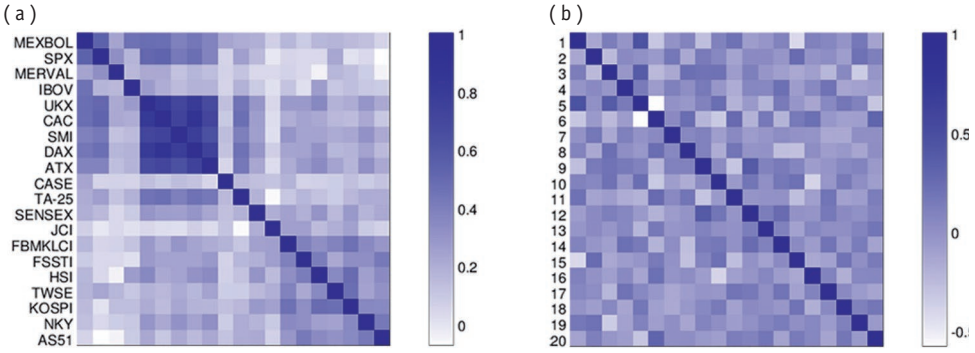
donde σ_k es la desviación estándar de R_k, y $< \cdots >$ denota el promedio temporal sobre el periodo estudiado. La polaridad se normalizó de la misma manera, y es denotada como $p_k(t)$ para el índice k al tiempo t. Por otro lado, la forma más simple de caracterizar los coeficientes de correlación entre series de tiempo normalizadas es mediante el cálculo de los elementos de matriz de Pearson:

$$c^{(x)}_{k,\,l} = \langle x_k(t) x_l(t) \rangle ,$$

donde el superíndice x es para denotar el tipo de serie de tiempo con la que se está trabajando, de tal manera que $c^{(p)}_{k,\,l}$ y $c^{(r)}_{k,\,l}$ son los elementos de la matriz de correlación k, l, construidos a partir de las series de tiempo de polaridad y retornos, respectivamente.

En la figura 1 se muestran las matrices de correlación de las polaridades y retornos como mapas de calor. En estas figuras las etiquetas del *keyword* son remplazadas por el nombre del país que representa al índice financiero o a la serie de tiempo de polaridad asociada. En esta representación los cuadros de color más oscuro denotan las correlaciones fuertes, mientras que los cuadros más claros representan las correlaciones débiles o anticorrelaciones. Los casos extremos son $c_{k,\,l} = 1(-1)$, lo que corresponde a una perfecta correlación (anticorrelación), mientras que $c_{k,\,l} = 0$ significa que la correlación es nula entre los elementos k y l. Se puede ver en la figura 1a, que emergen algunos patrones en los datos de retorno, como es la fuerte correlación entre el sector europeo, Estados Unidos de Norteamérica y México, así como Norteamérica con Europa, y el sector asiático, lo cual refleja la codependencia de las economías debido a su

Figura 1. *Twitter* e índices financieros.



Notas: (a) Elementos de la matriz de correlación para datos de retorno. (b) Elementos de la matriz de correlación para datos de polaridad. En esta escala de representación de colores, el valor mínimo corresponde al blanco, mientras que los valores más grandes corresponden a colores azul intenso.
Fuente: Elaboración propia.

Volumen 5 | número 12 | mayo–agosto 2017
DOI: http://dx.doi.org/10.22201/ceiich.24485705e.2017.12.61469

posición geográfica. No obstante, si seguimos la posición de estos índices en la figura 1b no encontramos la misma estructura. De aquí surge la necesidad de realizar un análisis más profundo para averiguar si existe una estructura de correlación oculta en la series de tiempo de polaridad.

Ensemble de Wishart

Deseamos ahora introducir una herramienta fundamental para el análisis multivariante proveniente de RMT. Sea W una matriz de dimensión $N \times T$, cuyos elementos son variables gaussianas estadísticamente independientes con media cero y varianza fija. La matriz $H = WW^{+}$ es conocida en RMT como matriz de Wishart y al *ensemble* (conjunto) generado por estas matrices como *ensemble* de Wishart (WE). Por construcción, estas matrices están formadas por N series de tiempo no correlacionadas de longitud finita T.

La densidad de probabilidad del espectro de eigenvalores se puede resolver analíticamente en el límite $N, T \to \infty$, y $Q = T/N \geq 1$, para el caso en que las entradas de la matriz H son números reales. A lo que se conoce como ley de Marcenko–Pastur [30]:

$$p(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda},$$

la cual presenta las cotas $\lambda_- \leq \lambda \leq \lambda_+$ donde:

$$\lambda_\pm = \sigma^2 (1 + 1/Q \pm 2\sqrt{1/Q}).$$

La cuestión de interés para nosotros es que *si no existen correlaciones entre las series de tiempo, entonces la distribución de los eigenvalores de la matriz de correlación debe estar acotada dentro de la ley de Marcenko–Pastur*. Estas predicciones son conocidas como resultados universales de las matrices de Wishart, y constituyen la hipótesis nula de la ausencia de correlaciones entre las variables de estudio, en nuestro caso entre los índices financieros globales y las polaridades de *Twitter*.

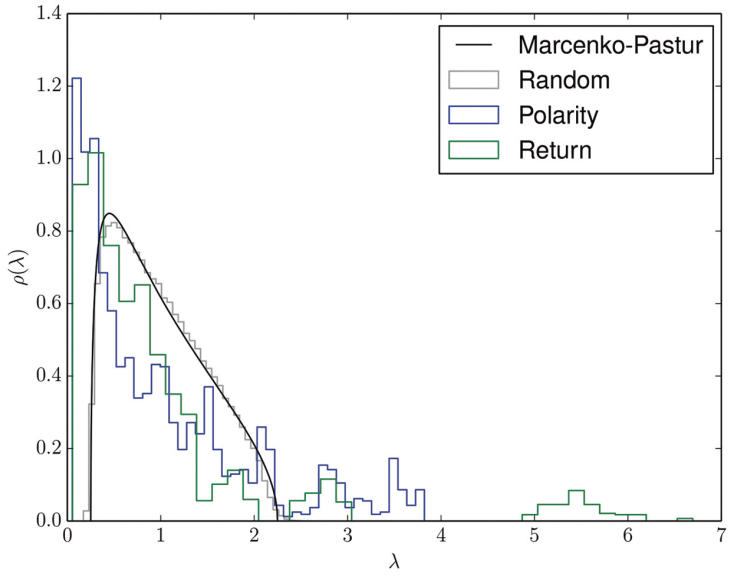
Aunque los resultados universales de las matrices de Wishart son válidos únicamente para dimensiones asintóticas ($N, T \to \infty$,), compararlos con nuestros datos empíricos sigue siendo útil, ya que nos puede proporcionar indicios acerca de la presencia de correlaciones ocultas. Para este propósito, hemos construido un conjunto de matrices de correlación muestra a partir de ventanas de tiempo de $T = 80$ días de cotización, deslizándolas por un día. De esta manera, hemos obtenido dos muestras de $M = 86$ matrices de correlación, uno para los valores

de polaridad, y el otro conjunto para los de retorno. Dentro de estos conjuntos, cada matriz de correlación tiene dimensiones N, $T = 20 \times 80$, con $Q = T/N = 4$, por lo que el espectro de eigenvalores está acotado entre los límites $\lambda_- = 0.25$ y $\lambda_+ = 2.25$, y uno esperaría como hipótesis nula, que la gran mayoría de los eigenvalores no presenten correlaciones y se encuentren dentro de estos límites.

Sin embargo, se encontró que los eigenvalores extremos para el conjunto de matrices de correlación de polaridad están entre $\lambda_- = 0.0526$ y $\lambda_- = 3.8215$, mientras los de retornos se encuentran entre entre $\lambda_- = 0.0556$ y $\lambda_- = 6.6942$. Además, se encontró que solamente el 68:02% de los eigenvalores de polaridad y 72:27% de los de retorno caen dentro de los resultados universales de la RMT, es decir, dentro de la zona asociada a ruido, donde no se presentan correlaciones. La distribución de eigenvalores para los datos empíricos, así como de las matrices de correlación se han graficado en la figura 2, superponiendo la ley Marcenko–Pastur en la misma figura.

Es importante recordar que la distribución de Marcenko–Pastur es válida solamente para el límite asintótico, por lo que las distribuciones finitas siempre presentan desviaciones a este resultado. Además, entre más grande es el valor T/N, más confiables son los resultados, y las fluctuaciones son descritas de for-

Figura 2. Distribución de eigenvalores de las matrices de correlación.



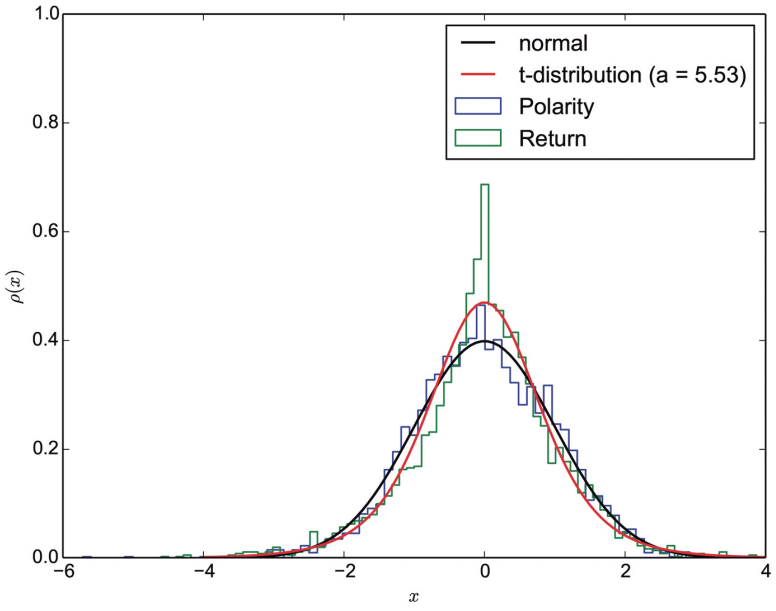
Notas: La línea negra muestra la ley de Macenko–Pastur. La línea gris representa los resultados numéricos para 10,000 miembros de *WE*, la línea azul los resultados para las polaridades, y la línea verde para los retornos.
Fuente: Elaboración propia.

Volumen 5 | número 12 | mayo–agosto 2017
DOI: http://dx.doi.org/10.22201/ceiich.24485705e.2017.12.61469

ma más realista por la varianza de los datos. Pero si T/N es un número pequeño, los resultados se verán afectados fuertemente por la finitud de la matriz de datos. En estos casos es una práctica común utilizar técnicas de *noise dressing* para omitir el ruido intrínseco de la matriz de correlación [31-33]. Aun así, estas técnicas funcionan bien hasta dimensiones cercanas a $N = 50$, para dimensiones más pequeñas (que es nuestro caso) se debe proceder de manera diferente.

Otro hecho que puede generar desviaciones de los resultados universales de las matrices de Wishart se debe a que la distribución de los retornos usualmente tiene colas más largas que la distribución normal [1], la cual se asume de manera idealizada en la derivación de la ley de Marcenko-Pastur. Para observar este fenómeno, en la figura 3 se ha graficado la distribución de los datos empíricos junto con la distribución normal y la distribución *t–Student* para caracterizar la distribución de los retornos. Para nuestro periodo de estudio, el parámetro a = 5.53 fue el que mejor se ajustó al caracterizar la distribución de los retornos, mientras la distribución de las polaridades parece ajustarse mucho mejor con la distribución normal en este caso: cabe resaltar que los datos de polaridad parecen romper la regla de colas largas encontrada por muchos autores para los retornos.

Figura 3. Distribución de datos empíricos, donde se ha superpuesto la distribución normal y la distribución *t–Student* con el parámetro que mejor ajusta los valores de retornos.



Fuente: Elaboración propia.

Eigenvalores extremos

En el área de riesgo financiero y optimización de portafolios, los eigenvalores más grandes y más pequeños representan cantidades muy importantes, pues están asociados con los casos extremos de riesgo en una cartera de inversión [5]. Los eigenvalores más grandes corresponden a una mezcla arriesgada de acciones o mercados financieros, mientras que los eigenvalores más pequeños están relacionados con un portafolio de bajo riesgo [34]. El eigenvalor más grande es el factor que representa la información colectiva de los índices, y el eigenvector correspondiente es conocido como el modo del mercado. Este eigenvector nos dice si los índices como conjunto van a la alza o a la baja, siendo su tendencia condicionada al estado actual del mercado [34].

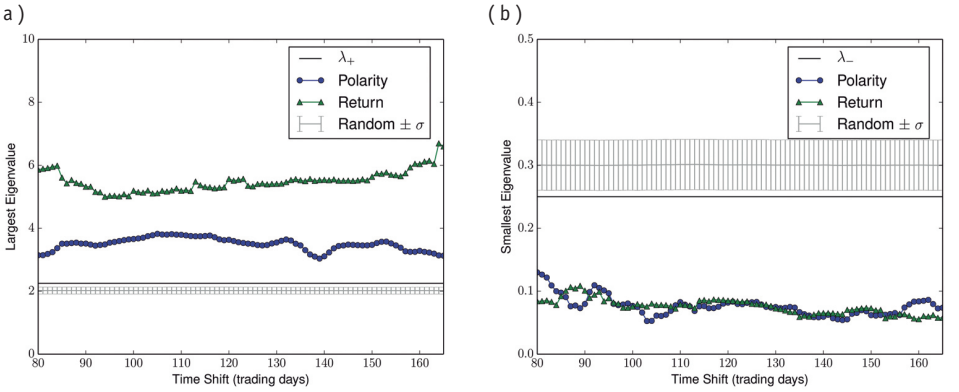
Estas características han sido exploradas exhaustivamente con datos provenientes de los índices financieros. Sin embargo, hasta el momento no se ha hecho un estudio con datos textuales. Es por ello que aquí se explora si este fenómeno también emerge al trabajar con la información proveniente de *Twitter*. Para este fin se analizó el comportamiento temporal del eigenvalor más grande y más pequeño de las matrices de correlación empíricas para cada periodo de estudio, utilizando la misma muestra de matrices ($M = 86$).

En la figura 4 se muestran los resultados empíricos para nuestro periodo de estudio, junto con la media y desviación estándar de una simulación numérica de 10,000 miembros de *WE*, donde cada punto se calcula teniendo en cuenta los 80 días de transacción anteriores. Se puede observar que los resultados empíricos están lejos de los bordes teóricos, así como a más de tres desviaciones estándar de los resultados numéricos. Además, se encontró una anticorrelación fuerte entre polaridades y retornos al comparar el comportamiento temporal de sus eigenvalores más grandes. El coeficiente de Pearson encontrado fue $Pc = 0.70$, el de Spearman $Sc = 0.69$, ambos con valores de confidencia menores a 1×10^{-12}. Por el contrario, el comportamiento temporal de los eigenvalores más pequeños muestra correlaciones positivas más moderadas, con valores de $Pc = 0.45$ y $Sc = 0.49$.

El hecho de que el comportamiento temporal de los eigenvalores más grandes de los retornos y polaridades estén anticorrelacionados mutuamente, puede deberse a un retraso en la transmisión de información de *Twitter* hacia los precios de los mercados financieros globales. Pudiendo esto constituir una evidencia más en contra de la hipótesis de mercado eficiente. Además, el coeficiente de correlación que se encontró para el comportamiento temporal de los eigenvalores empíricos más pequeños revela que el portafolio de menor riesgo se preserva aproximadamente sin importar si usamos polaridades o retornos para su cálculo, por lo que *Twitter* resulta ser una fuente de información muy interesante para el análisis de portafolios.

Volumen 5 | número 12 | mayo–agosto 2017
DOI: http://dx.doi.org/10.22201/ceiich.24485705e.2017.12.61469

Figura 4. Eigenvalores extremos para *Twitter* e índices financieros.



Notas: (a) Comportamiento temporal de los eigenvalores más grandes. (b) Comportamiento temporal de los eigenvalores más pequeños. La línea azul representa los resultados para las polaridades, la verde para retornos, y la línea negra los límites predichos por RMT para las matrices de Wishart, mientras que la línea gris representa la media y desviación estándar para una simulación numérica con 10,000 miembros de *WE*.
Fuente: Elaboración propia.

En general, estos resultados proveen evidencias acerca del surgimiento de factores comunes en la información financiera global, sin la necesidad de discriminar si los datos provienen de los retornos o de las polaridades, en otras palabras, con la información proveniente de *Twitter* parecer ser posible caracterizar el comportamiento colectivo de los mercados financieros globales.

Razón de participación inversa

La razón de participación inversa (IPR, por sus siglas en inglés) es una manera simple de cuantificar en cuántos estados está distribuida una partícula cuando existe cierta incertidumbre acerca de dónde se encuentra. Históricamente [35], la razón de participación (PR, por sus siglas en inglés) fue introducida para ayudar a clasificar las vibraciones atómicas en las redes cristalinas desordenadas [36]. Esta cantidad describe la fracción del número total de sitios que participan en un modo vibracional correspondiente al eigenvector $x = (x_1, \ldots, x_n) \in R^n$, y toma el valor

$$PR = \frac{(\mu^1)^2}{\mu^0 \mu^2}$$

donde $\mu^r = \sum_{i=1} |x_i|^{2r}$ puede ser visto como el momento r de la energía cinética del modo. Si un modo dado envuelve el movimiento de un solo átomo, se caracteriza como localizado y tiene el valor $PR = 1/N$. Por el contrario, un modo vi-

bracional que contenga a todos los átomos participando por igual es llamado extendido y tiene el valor $PR = 1$. Asimismo, una medida equivalente ha sido utilizada para estudiar el grado de localización de los eigenestados electrónicos en el modelo de Anderson [37] con implicaciones en la existencia de transiciones localizadas (no–localizadas) de un metal aislante en presencia de desorden.

En econofísica, una manera simple de extraer información a partir de los eigenvectores es estimando esta *ipr*, la cual nos permite conocer el número de índices que participan significativamente en cada eigenvector (o portafolio). Esta medida exhibe la distinción entre los eigenvectores asociados a los extremos y aquellos que pertenecen al resto del conjunto, dentro de la zona de ruido. Si se considera el *k–ésimo* eigenvector normalizado de la matriz de correlación $|V_k| = 1$, la *ipr* del eigenvector V_k se puede escribir como:

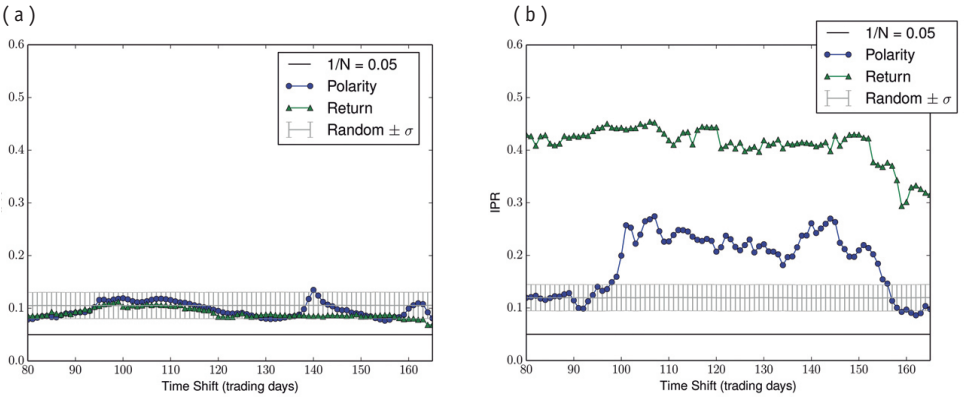
$$IPR_k = \sum_{j=1}^{N} |V_j^k|^4$$

cuyo valor siempre cae entre los límites $1/N$ y uno. Si el eigenvector V_k se encuentra localizado solamente en un componente, entonces $IPR_k = 1$. Por el contrario, si se encuentra distribuido uniformemente sobre los N componentes, entonces $IPR_k = 1/N$. Es de esperarse que los valores para IPR_N fluctúen cerca del límite inferior $1/N$, ya que corresponde al portafolio más diversificado, mientras que para IPR_1 se esperan valores más altos, ya que está asociado al eigenvalor más pequeño, y, por lo tanto, al portafolio menos diversificado [4]. Asimismo, para valores de $1 < k < N$, dentro de la región considerada como ruido, es de esperarse que surjan combinaciones aleatorias de los componentes, y, en consecuencia, valores de IPR_k comprendidos entre los de IPR_N y los de IPR_1.

En la figura 5 se muestra el comportamiento temporal de IPR_N (figura 5(a)) y de IPR_1 (figura 5(b)) para los datos empíricos de *Twitter* e índices financieros. En estas mismas figuras se muestra la media y desviación estándar de una simulación numérica de 10,000 miembros de *WE*, donde de nuevo cada punto se calcula teniendo en cuenta los 80 días de transacción anteriores. Se puede ver en la figura 5(a) que ambos resultados empíricos presentan un comportamiento suave y fluctúan alrededor del límite inferior como es de esperarse, cayendo en la región de los resultados numéricos, lo cual confirma que cada uno de los indicadores financieros involucrados participa significativamente en V_k, y como consecuencia todos los índices se mueven como uno solo en este eigenmodo. Es interesante observar que esta misma característica emerge cuando trabajamos con las polaridades. Para este caso se ha encontrado un $Pc = 0.6$ entre ambos comportamientos empíricos.

Volumen 5 | número 12 | mayo–agosto 2017
DOI: http://dx.doi.org/10.22201/ceiich.24485705e.2017.12.61469

Figura 5. *IPR* para *Twitter* e índices financieros.



Notas: (a) Comportamiento temporal de *IPR* correspondiente a los eigenvalores más grandes. (b) Comportamiento temporal de *IPR* para el eigenvalor más pequeño. La línea azul representa los resultados para los retornos, la línea verde para polaridades, y la línea negra el límite inferior $1/N$. Además, la línea gris representa la media y la desviación estándar de los resultados de la simulación numérica de 10,000 miembros de *WE*.
Fuente: Elaboración propia.

Si ahora fijamos nuestra atención en la figura 5(b), podemos observar que el IPR_1 se comporta de manera bastante diferente en ambos datos empíricos. Los resultados para retornos se mantienen fluctuando más de tres desviaciones estándar sobre lo esperado para los resultados numéricos, mientras que los resultados de polaridad se encuentran lejos de los valores numéricos la mayor parte del tiempo, aunque hay periodos en que caen dentro de los valores de la simulación. Esto último podría implicar la presencia de ruido en la adquisición de los datos de polaridad, principalmente al comienzo del periodo de estudio. Sin embargo, incluso así se presenta una correlación positiva entre el comportamiento temporal de los datos empíricos, con un $Pc = 0.49$.

Conclusión

Se logró extraer información de *Twitter* mediante el ensamblaje de distintos lenguajes de programación, cuantificando su contenido mediante técnicas de análisis de sentimiento. Donde las técnicas matemáticas provenientes de la física estadística mostraron que los datos extraídos de estas fuentes contienen información relevante para el análisis de los índices financieros globales estudiados aquí.

Mediante el análisis de RMT, se ha encontrado que los datos de *Twitter* comparten la misma estructura de correlaciones encontrada para los datos de retorno asociados con cada periodo de estudio. Asimismo, se han encontrado desvia-

ciones largas de los eigenvalores, más allá de los límites predichos por la ley de Marcenko–Pastur, lo cual nos dice que existen correlaciones verdaderas entre los índices financieros y las polaridades. Por lo cual el hecho de que en este momento no seamos capaces de generar estrategias de compra en las bolsas de valores o de prevenir posibles crisis financieras a partir de la información textual, no imposibilita que en un futuro podamos hacerlo.

Asimismo, el estudio de RMT ha permitido observar una correlación moderada entre los comportamientos temporales de los eigenvalores extremos de las polaridades y retornos en ambos periodos de estudio. Esto implica que la información colectiva de los índices financieros globales emerge también al analizar las polaridades, y, por lo tanto, el portafolio de inversión óptimo o más diversificado es preservado al utilizar este tipo de información. Además, se encontró que los valores de IPR_N fluctúan cerca del límite inferior $1/N$, mientras que para IPR_1 se obtuvieron los valores más altos como es de esperarse para el caso del portafolio más y menos diversificado, respectivamente. Es notable observar que esta característica surge cuando trabajamos con las polaridades. Ello implica que la estructura de las correlaciones globales puede ser preservada independientemente de si estamos trabajando con las fuentes de *Twitter* o la información financiera (retornos). Este conjunto de resultados obtenidos por medio de RMT sugiere que los retornos y polaridades comparten una estructura de correlaciones común para los países y periodos de tiempo estudiados aquí.

En suma, estos nuevos resultados apoyan el paradigma de la economía conductual, es decir, de que las decisiones de los inversionistas se ven influenciados por la información de los medios de comunicación y redes sociales (Twitter) lo cual influye para generar juicios o estrategias precipitadas de comercio en el mercado de valores, influyendo finalmente estas decisiones en el precio final de las acciones.

Referencias

[1]. García, A. «Global financial indices and twitter sentiment: A random matrix theory approach.» *Physica A*, 461, 509, 2016.
[2]. Mantenga, R. N. y H. E. Stanley. *An introduction to econophysics: Correlations and complexity in finance*. Cambridge: Cambridge University Press, 2000.
[3]. Bouchaud, J. P. y M. Potters. *Theory of financial risks: From statistical physics to risk management*. Cambridge University Press, Cambridge, 2000.
[4]. Voit, J. *The statistical mechanics of financial markets*. Berlín: Springer–Verlag, 2005.
[5]. Markowitz, H. *Portfolio selection: Eficient diversification of investment*. Nueva York: John Wiley & Sons, 1959.

Volumen 5 | número 12 | mayo–agosto 2017
DOI: http://dx.doi.org/10.22201/ceiich.24485705e.2017.12.61469

[6]. Wishart, J. En *Biometrika*, 20A, 1928, 32.
[7]. Wigner, E. En *Ann. Math.*, 1955, 548.
[8]. Mehta, M. L. *Random matrices and the statistical theory of energy levels*. Nueva York: Academic Press, 1967.
[9]. Brody,T. A., J. Flores, J. B. French, P. A. Mello, A. Pandey S. S. M. y Wong. En *Rev. Mod. Phys.*, 53, 1981, 385.
[10]. Guhr, T., A. Muller–Groeling y H. A. Weidenmuller. En *Phys. Rep.*, 299, 1998, 189.
[11]. Plerou, V., P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral y H. E. Stanley. En *Phys. Rev. Lett.*, 83 1999, 1471.
[12]. Laloux, L., P. Cizeau, J. P. Bouchaud y M. Potters. En *Phys. Rev. Lett.*, 83 1999, 1467.
[13]. Plerou, V., P. Gopikrishman, B. Rosenow, L. A. N. Amaral y H. E. Stanley. En *Physica A*, 287, 2000, 374.
[14]. Plerou, V., P. Gopikrishman, B. Rosenow, L. A. N. Amaral, T. Guhr y H. E. Stanley. En *Phys. Rev. E*, 2002, 066126.
[15]. Potters, M., J.–P. Bouchaud y L. Laloux. En Acta. *Phys. Pol. B*, 36, 2005, 2767.
[16]. Medina, L. y R. Mansilla. En *J. Manag., Finance and Econ.*, 2, 2008, 125.
[17]. Munnix, M. C., R. Schafer y T. Guhr. En *Physica A*, 389, 2010, 76.
[18]. Zhang, X., H. Fuehres y P. A. Gloor. En *Procedia Soc. Behav. Sci.*, 26, 2010, 55.
[19]. Bollen, J., H. Mao y X. Zeng. En *J. Comput. Phys.*, 2, 2011, 1.
[20]. Smailovic, Jasmina , Miha Grcar, Nada Lavrac y Martin Znidarsic. *Human–computer interaction and knowledge discovery in complex, unstructured, big data*, vol. 7947 of the series *Lecture notes in computer science*. Berlín Heidelberg: Springer-Verlag, 2013, 77.
[21]. Oliveira, Nuno, Paulo Cortez y Nelson Areal. *Progress in artificial intelligence*, vol. 8154 of the series *Lecture notes in computer science*. Berlín Heidelberg: Springer-Verlag, 2013, 355.
[22]. Preis, T., H. S. Moat y H. E. Stanley. En *Sci. Rep.*, 3, 2013, 1684.
[23]. Alanyali, M., H. S. Moat y T. Preis. En *Sci. Rep.*, 3, 2013, 3578.
[24]. Zheludev, I., R. Smith y T. Aste. En *Sci. Rep.*, 4. 2014, 4213.
[25]. Plakandaras, Vasilios, Theophilos Papadimitriou, Periklis Gogas y Konstantinos Diamantaras. En *Algorithmic Finance*, 4, 2015, 69.
[26]. Barberis, Nicholas. *Handbook of the economics of finance*. North Holland: Elsevier Science B.V., 2003, 1051.
[27]. Sandoval, L. Jr., I. D. P. Franca. En *Physica A*, 391, 2012, 187.
[28]. Pang, B. y L. L. Found. En *Trends. Network*, 2, 1986, 1.
[29]. Jurafsky, D. y J. H. Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition*. Englewood Cliffs NJ Prentice-Hall, 2000.

[30]. Marcenko, V. A. y L. A. Pastur. En *Sb. Math.*, 72, 1967, 507.
[31]. Laloux, L., P. Cizeau, J. P. Bouchaud y M. Potters. En *Int. J. Theor. Appl. Finance*, 3, 2000, 391.
[32]. Rosenow, B., V. Plerou, P. Gopikrishnan y H.E. Stanley. En *Europhys. Lett.*, 59, 2002, 500.
[33]. Sandoval, L., A. B. Bortoluzzo y M. K. Venezuela. En *Physica A*, 410, 2014, 94.
[34]. J. P. Bouchaud, J. P y M. Potters. «Financial applications of random matrix theory: A short review.» En Por Gemot Akemann, Jinho Balk y Philippe Di Francesco (eds.), *The Oxford handbook of random matrix theory*. Oxford: Oxford University Press, 2011, cap. 40, 824.
[35]. Clark, T. B. P. y A. Del Maestro. En *arXiv*, 2015, arXiv: 1506.02048.
[36]. Bell, R. J. y P. Dean. En *Discuss. Faraday Soc.*, 50 (1970), 55.
[37]. Visscher, W. M. En *J. Non–Cryst. Solids*, 477, 1972, 8.