Nereis. Revista Iberoamericana Interdisciplinar de Métodos, Modelización y Simulación	6	47-67	Universidad Católica de Valencia "San Vicente Mártir"	Valencia (España)	ISSN 1888-8550

Método gaussiano de suavización de datos experimentales

Fecha de recepción y aceptación: 21 de noviembre de 2013, 16 de diciembre de 2013

Juan Luis González-Santander Martínez y Germán Martín González

Facultad de Ciencias Experimentales y Matemáticas, Universidad Católica de Valencia "San Vicente Mártir".
Correspondencia: Calle Guillem de Castro, 94. 46001 Valencia. España. *E-mail*: martínez.gonzalez@ucv.es y german.martin@ucv.es

ABSTRACT

We provide a method for experimental data smoothing under a certain noise by using a statistical fitting considering gaussian weight functions. This method is quite useful when we have a large amount of experimental data, which are expected to approach an unknown theoretical curve. This allows us to find quite closely the derivative of the theoretical curve from the data and provides as well the error in the numerical integration of the data. The latter is not possible by using the typical discrete Fourier transform smoothing. On the other hand, the proposed method improves the typical smothening of the time series of financial data and allows the calculation of the volatility as a function of time.

KEYWORDS: *Curve smoothening, non-parametric regression, experimental data filtering.*

RESUMEN

Se presenta un método de suavización de datos experimentales sometidos a un cierto ruido, utilizando un ajuste estadístico con funciones peso de tipo gaussiano. Este método resulta bastante útil cuando disponemos de una gran cantidad de datos que presumiblemente se aproximan a una curva teórica desconocida. Dicho ajuste permite hallar con bastante aproximación la derivada de la curva teórica a partir de los datos y permite ofrecer el error cometido en su integración numérica. Esto último no es posible con la suavización usual que utiliza la transformada discreta de Fourier. Por otro lado, el método propuesto mejora las suavizaciones típicas de las series temporales de datos financieros y permite obtener la volatilidad en función del tiempo.

PALABRAS CLAVE: *Curva suavizada, regresión no paramétrica, filtrado de datos experimentales.*

INTRODUCCIÓN

En muchos experimentos se dispone de una gran cantidad de datos sujetos a un cierto ruido o error experimental. Dichos datos presumiblemente se ajustan a una cierta curva teórica. En estos casos, no es aconsejable utilizar como aproximación a la curva teórica la que se puede obtener por interpolación, ya sea lineal, polinómica o con "splines" [1], pues dicha curva tendría que pasar (artificialmente) por todos los puntos experimentales. Podemos optar entonces por suavizar los datos utilizando la transformada discreta de Fourier [2] y luego interpolar. La desventaja de este método es que no ofrece la banda de confianza en la que los datos se encuentran con una determinada probabilidad. Esto último sí es posible cuando realizamos un ajuste por mínimos cuadrados, siempre y cuando conozcamos la forma funcional de la curva teórica. El método presentado en este trabajo combina, por un lado, la eliminación del ruido experimental de los datos sin disponer de la forma funcional teórica a la que debieran ajustarse, y, por otro lado, ofrece la banda de confianza en la que dicha curva teórica puede hallarse con un cierto nivel de probabilidad. Las suavizaciones son muy útiles cuando necesitamos obtener la función derivada de los datos, mientras que la banda de confianza es útil cuando necesitamos dar una aproximación del error que se comete cuando realizamos la integral de esos mismos datos.

Para comprobar la bondad del método propuesto, en primer lugar, simularemos numéricamente los datos experimentales a partir de una curva teórica en la que hemos introducido un cierto nivel de ruido. A continuación, suavizaremos los datos simulados con el método propuesto y procederemos a comparar el resultado obtenido con la curva teórica de partida. Es más, el método propuesto permitirá inferir el perfil del ruido que tienen los datos a partir de los propios datos. Por último, analizaremos el comportamiento del método propuesto sobre datos reales en los que no se conoce ni la curva teórica a la que se ajustan ni el perfil de ruido que tienen.

Este artículo se organiza de la siguiente manera: después de la introducción, en la siguiente sección propondremos una función de ajuste cuando tenemos una gran cantidad de datos que están afectados por un mismo error absoluto en la medida (datos homocedásticos). A continuación, dedicaremos una sección en la que generalizaremos el resultado obtenido en la sección anterior para datos que están afectados por un error no constante (datos heterocedásticos). En la siguiente sección analizaremos cuatro ejemplos numéricos (uno homocedástico y tres heterocedásticos) para evaluar la bondad de la función de ajuste propuesta: en el primer ejemplo, los datos estarán afectados por un error absoluto constante; en el segundo, los datos estarán afectados por un error relativo constante; en el tercer ejemplo, los datos se simularán con un error variable de perfil lorentziano, y en el último ejemplo, se aplicará el método propuesto de suavización a la serie temporal real de un índice del mercado bursátil. Las conclusiones se recogerán en la última sección.

AJUSTE PARA DATOS HOMOCEDÁSTICOS

Podemos entender los datos experimentales y_i como la predicción de la curva teórica $y(x)$ en unas determinadas abscisas x_i, $i = 1,\ldots, n$, más un cierto ruido aleatorio producido por la incertidumbre en la medida. Supongamos que dicho ruido aleatorio es el mismo en todas las abscisas (datos homocedásticos) y se distribuye normalmente[1] con una desviación típica σ, de tal modo que tenemos,

$$y_i \sim N\left(y(x_i),\sigma\right). \tag{1}$$

Sea $\langle y\rangle$ la media ponderada de los valores de puntos y_i con una determinada función peso p_i

$$\langle y\rangle = \frac{\sum_{i=1}^{n} y_i p_i}{\sum_{i=1}^{n} p_i}. \tag{2}$$

Para que (2) se corresponda con una buena función de ajuste f, observemos que, cuando evaluamos f en la abscisa x, los puntos x_i que han de tener un mayor peso son los que están cercanos a la abscisa x. Si suponemos que la influencia de un punto x_i tiene una distribución gaussiana, podemos proponer la siguiente función peso

$$p_i(x,s) := \frac{1}{s\sqrt{2\pi}} \exp\left[-\frac{\left(x-x_i\right)^2}{2s^2}\right]. \tag{3}$$

La magnitud del parámetro s da cuenta de lo localizada o dispersa que está la influencia de cada punto x_i en su entorno. Sustituyendo en (2) y simplificando, resulta la siguiente función de ajuste

$$f(x,s) := \frac{\sum_{i=1}^{n} y_i\, \rho_i(x,s)}{\sum_{i=1}^{n} \rho_i(x,s)}. \tag{4}$$

[1] Los errores experimentales se distribuyen normalmente por el Teorema del Límite Central [3].

donde hemos definido la función de influencia como

$$\rho_i(x,\mathrm{s}) := \exp\left[\frac{x_i(2x-x_i)}{2s^2}\right]. \tag{5}$$

Obsérvese que necesitamos una gran cantidad de datos para que el promedio local que supone la función de ajuste *f* se aproxime adecuadamente a la curva teórica.

Determinación de s

Observemos que cuanto más pequeño es *s*, más estrechamente se va a ajustar *f* a los puntos de la nube. En el límite $s \to 0$ la curva de ajuste proporciona una interpolación. Efectivamente, recordando que una forma de representar la delta de Dirac es por medio del siguiente límite de una gaussiana [4]

$$\lim_{s\to 0} \frac{1}{s\sqrt{2\pi}} \exp\left[-\frac{(x-x_i)^2}{2s^2}\right] = \delta(x-x_i),$$

resulta que

$$f_0(x) := \lim_{s\to 0} f(x,s) = \frac{\sum_{i=1}^{n} y_i \delta(x-x_i)}{\sum_{i=1}^{n} \delta(x-x_i)}.$$

Por tanto,

$$f_0(x_j) = y_j.$$

Es decir, la función de ajuste *f* pasa por todos los puntos de la dispersión. Por el contrario, cuando más grande es *s*, más se promediarán todos los datos entre sí, obteniendo una curva de ajuste con un valor constante. Es fácil ver, a partir de (5), que en el límite $s \to \infty$, la función de ajuste se corresponde con la media

$$\lim_{s\to\infty} f(x,s) = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Ahora bien, ¿cuál es el valor que debe tener *s* para que se promedien los errores debidos a la medición, pero se conserven las fluctuaciones debidas a la curva teórica? Para responder a esta pregunta expresemos (1) de la siguiente manera

$$y_i = y(x_i) + u_i, \qquad i = 1,\ldots,n, \tag{6}$$

en donde $u_i \sim N(0,\sigma)$. Como la función de ajuste *f* se ha de aproximar a la curva teórica

$$y(x_i) \approx f(x_i,s),$$

podemos reescribir (6) como

$$y_i - f(x_i, s) \approx u_i. \tag{7}$$

Tomando esperanzas en (7)

$$\mathrm{E}(y_i - f(x_i, s)) \approx \mathrm{E}(u_i) = 0. \tag{8}$$

Tomando varianzas en (7)

$$\mathrm{Var}(y_i - f(x_i, s)) \approx \mathrm{Var}(u_i) = \sigma^2. \tag{9}$$

De acuerdo con (8) y (9), resulta que

$$\frac{1}{n}\sum_{i=1}^{n}\left[f(x_i, s) - y_i\right]^2 \approx \sigma^2. \tag{10}$$

Como el error cuadrático medio de los residuos de la función de ajuste f con respecto a los puntos de la dispersión es

$$ECM(s) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[f(x_i, s) - y_i\right]^2}. \tag{11}$$

tendremos, teniendo en cuenta (10), que

$$\sigma \approx ECM(s). \tag{12}$$

Podemos interpretar (12) diciendo que si el error de medición en y_i se distribuye normalmente con una desviación típica igual a σ –véase (1)– es razonable pensar que los datos se desviarán en promedio con respecto a la curva de ajuste, tanto como se desvían con respecto a la curva teórica. Por tanto, si conocemos σ (el error en la medida), podemos aproximar $s \approx s_{\hat{fit}}$ en (4), resolviendo numéricamente la ecuación

$$ECM(s_{\hat{fit}}) - \sigma = 0. \tag{13}$$

Un buen punto de inicio s_0 para resolver numéricamente (13) consiste en tomar la distancia promedio entre dos abscisas consecutivas

$$s_0 = \frac{1}{n-1}\sum_{i=1}^{n-1}(x_{i+1} - x_i), \tag{14}$$

pues resulta ser un valor mínimo para que exista una cierta influencia entre puntos consecutivos. En conclusión, la función de ajuste vendrá dada por

$$f(x) := f(x, s_{\hat{fit}}). \tag{15}$$

Y la función peso por

$$\rho_i(x) := \rho_i(x, s_{fit}). \tag{16}$$

Función derivada de la función de ajuste

En muchas ocasiones, lo que interesa es determinar la derivada de la función a la que supuestamente se ajustan los datos. Para obtener una aproximación de dicha función derivada, derivemos la función *f* de ajuste (15)

$$f'(x) = \frac{\sum_{i=1}^{n} y_i \, \rho_i'(x)}{\sum_{i=1}^{n} \rho_i(x)} - \frac{\sum_{i=1}^{n} \rho_i'(x) \sum_{i=1}^{n} y_i \, \rho_i(x)}{\left[\sum_{i=1}^{n} \rho_i(x)\right]^2}. \tag{17}$$

Como, según (5)

$$\rho_i'(x) = \frac{x_i}{s_{fit}^2} \rho_i(x),$$

podemos simplificar (17), con lo que obtenemos

$$f'(x) = \frac{1}{s_{fit}^2} \left\{ \frac{\sum_{i=1}^{n} x_i \, y_i \, \rho_i(x)}{\sum_{i=1}^{n} \rho_i(x)} - \frac{\sum_{i=1}^{n} x_i \, \rho_i(x) \sum_{i=1}^{n} y_i \, \rho_i(x)}{\left[\sum_{i=1}^{n} \rho_i(x)\right]^2} \right\}. \tag{18}$$

Utilizando la notación utilizada en (2), podemos expresar (18) como

$$f'(x) = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{s_{fit}^2}. \tag{19}$$

Si definimos la covarianza promediada a la función peso ρ_i como

$$\mathrm{Cov}\langle x, y \rangle := \frac{\sum_{i=1}^{n} \left(x_i - \langle x \rangle\right)\left(y_i - \langle y \rangle\right) \rho_i(x)}{\sum_{i=1}^{n} \rho_i(x)}. \tag{20}$$

Podemos escribir (19) de la siguiente manera

$$f'(x) = \frac{\mathrm{Cov}\langle x, y \rangle}{s_{fit}^2}. \tag{21}$$

Suavización por transformada de Fourier

Podemos optar también por suavizar los datos utilizando la transformada discreta de Fourier. Para ello, transformamos la señal y_i con la transformada de Fourier en coseno F_C, [2]

$$\hat{y}_i = F_C[y_i], \quad i = 1, ..., n,$$

y tomamos de la señal transformada solo los $k \leq n$ primeros modos

$$\hat{y}_i(k) = \begin{cases} \hat{y}_i & i = 1, ..., k, \\ 0 & i = k+1, ..., n. \end{cases}$$

Anti-transformamos la señal $\hat{y}_i(k)$ para obtener la señal filtrada,

$$y_i(k) = F_C^{-1}(\hat{y}_i(k)), \quad i = 1, ..., n.$$

Definamos el error cuadrático medio entre la señal y_i y la señal filtrada $y_i(k)$ como

$$ECM_F(k) := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[y_i(k) - y_i\right]^2}.$$

A semejanza de (12), podemos determinar el número de modos k en la señal filtrada con la condición

$$ECM_F(k) \approx \sigma. \tag{22}$$

Debido a que cuantos más modos tomemos más se parecerá la señal filtrada a la original, la condición (22) se traduce en encontrar el número mínimo de modos $k_{\min}$ que satisfaga

$$\min_k \left\{ECM_F(k)\right\} \leq \sigma. \tag{23}$$

Según este método, la función de ajuste resulta de tomar la función interpoladora $f_{\text{int}}(x)$, que pasa por los puntos de la señal filtrada, es decir, por los puntos $(x_i, y_i(k_{\min}))$.

Banda de confianza

Análogamente a (4), podemos hacer una media cuadrática ponderada de lo que se desvían los datos con respecto a la función de ajuste

$$\Delta f(x) := \sqrt{\frac{\sum_{i=1}^{n}\left[f(x_i) - y_i\right]^2 \rho_i(x)}{\sum_{i=1}^{n}\rho_i(x)}}. \tag{24}$$

De este modo, podemos dar la siguiente banda de confianza en la curva de ajuste

$$f(x) \pm \Delta f(x).$$

La banda de confianza permite dar un error en la integración numérica de los datos en el intervalo en el que estos se encuentran, $x \in (x_1, x_n)$

$$I = \int_{x_1}^{x_n} f(x)\,dx \pm \int_{x_1}^{x_n} \Delta f(x)\,dx,$$

lo cual no es posible si obtenemos una función interpoladora de ajuste $f_{int}(x)$

$$I_{\text{int}} = \int_{x_1}^{x_n} f_{\text{int}}(x)\,dx.$$

AJUSTE PARA DATOS HETEROCEDÁSTICOS

Cuando el error al que están sometidos los datos experimentales se distribuye normalmente, pero no es el mismo para todos los datos, análogamente a (1), podemos escribir

$$y_i \sim N\left(y(x_i), \sigma(x_i)\right), \quad i = 1, \ldots, n. \tag{25}$$

Consideremos ahora (24), pero para un parámetro *s* cualquiera

$$\Delta f(x,s) := \sqrt{\frac{\sum_{i=1}^{n} \left[f(x_i, \mathrm{s}) - y_i\right]^2 \rho_i(x,s)}{\sum_{i=1}^{n} \rho_i(x,s)}}. \tag{26}$$

La integral de los residuos cuadráticos entre el error teórico $\sigma(x)$ y el error del ajuste (26), viene dada por

$$G(s) := \int_{x_1}^{x_n} \left[\Delta f(x,s) - \sigma(x)\right]^2 dx. \tag{27}$$

Si conocemos la función $\sigma(x)$, el parámetro *s* óptimo, s_{fit}, se puede hallar minimizando numéricamente la función *G(s)*

$$\min[G(s)] = G(s_{fit}).$$

Datos con error relativo constante

Si tenemos unos datos experimentales con un error relativo σ_r constante

$$\sigma_r = \frac{\sigma(x)}{y(x)}.$$

De acuerdo con (25), resulta que

$$y_i \sim \mathrm{N}(y(x_i), \sigma_r\ y(x_i)), \qquad i = 1, \ldots, n.$$

Teniendo en cuenta que la curva de ajuste se debe aproximar a la curva teórica

$$y(x) \approx f(x), \tag{28}$$

de acuerdo con (27), la función a minimizar se aproxima a

$$G(s) \approx \sum_{j=1}^{n} \left[\Delta f(x_j, s) - \sigma_r\ f(x_j, \mathrm{s}) \right]^2. \tag{29}$$

Igual que anteriormente, un buen punto de inicio en la minimización numérica de (29) puede ser (14). Una vez evaluado s_{fit}, la banda de confianza se determina de acuerdo con (24).

Datos con error desconocido

Si la función $\sigma(x)$ es una función desconocida, su valor se puede aproximar a partir de las oscilaciones que presentan los datos en abscisas consecutivas. Efectivamente, a partir de (1) podemos escribir para $j = 1 \ldots n - 1$

$$y_j = y\left(x_j\right) + u_j, \tag{30}$$

$$y_{j+1} = y\left(x_{j+1}\right) + u_{j+1}, \tag{31}$$

donde u_j, u_{j+1} son variables aleatorias normales:

$$u_i \sim \mathrm{N}(0, \sigma(x_j)),$$
$$u_{i+1} \sim \mathrm{N}(0, \sigma(x_{j+1})).$$

Restando (31) de (30), se obtiene

$$y_{j+1} - y_j = y(x_{j+1}) - y(x_j) + u_{j+1} - u_j. \tag{32}$$

Ahora bien, el desarrollo de Taylor de primer orden (aproximación lineal) de una función derivable cualquiera $z(x)$ es

$$z(x + h) \approx z(x) + h\, z'(x), \tag{33}$$

o bien

$$z(x - h) \approx z(x) - h\, z'(x). \tag{34}$$

Aplicando estos resultados a la función teórica $y(x)$ y restando (33) y (34), se obtiene, escogiendo como valor x

$$x = m_j = \frac{x_{j+1} + x_j}{2}, \tag{35}$$

y como valor h

$$h = \frac{\Delta x_j}{2} = \frac{x_{j+1} + x_j}{2}, \tag{36}$$

de tal modo que

$$y(x_{j+1}) - y(x_j) \approx y'(m_j)\,\Delta x_j. \tag{37}$$

Como el valor de la función de ajuste se debe aproximar al valor de la curva teórica (28), resulta que (37) se puede escribir como

$$y(x_{j+1}) - y(x_j) \approx f'(m_j, \mathrm{s})\,\Delta x_j, \tag{38}$$

donde la derivada de la función de ajuste se puede calcular de acuerdo con (21). Sustituyendo (38) en (21), resulta entonces

$$y(x_{j+1}) - y(x_j) \approx f'(m_j, \mathrm{s})\,\Delta x_j \approx u_{j+1} - u_j. \tag{39}$$

Tomando esperanzas en (39), y teniendo en cuenta que u_{j+1} y u_j son variables independientes (las medidas son independientes entre sí)

$$\mathrm{E}[y(x_{j+1}) - y(x_j) - f'(m_j, \mathrm{s})\,\Delta x_j] \approx \mathrm{E}(u_{j+1} - u_j) = \mathrm{E}(u_{j+1}) - \mathrm{E}(u_j) = 0. \tag{40}$$

Tomando esperanzas en (39), y teniendo en cuenta que u_{j+1} y u_j son variables independientes (las medidas son independientes entre sí)

$$\mathrm{Var}[y(x_{j+1}) - y(x_j) - f'(m_j, \mathrm{s})\,\Delta x_j] \approx Var(u_{j+1} - u_j) = \sigma^2(u_{j+1}) + \sigma^2(u_j). \tag{41}$$

Si ahora $z(x) = \sigma^2(x)$, podemos sumar las ecuaciones (33) y (34) escogiendo de nuevo como valores x y h, (35) y (36) respectivamente, de tal modo que

$$\sigma^2(x_j) + \sigma^2(x_{j+1}) \approx 2\sigma^2(m_j). \tag{42}$$

Por tanto, a partir de (40)-(42) podemos dar la siguiente aproximación del error que tienen los datos en m_j

$$\sigma(m_j, s) \approx \frac{\left| y_{j+1} - y_j - f'(m_j, s)\,\Delta x_j \right|}{\sqrt{2}}. \tag{43}$$

Como ahora tenemos solo aproximaciones del error en los puntos medios m_j, análogamente a (29) podemos considerar la siguiente función para minimizar

$$G(s) \approx \sum_{j=1}^{n-1} [\Delta f(m_j, s) - \sigma(m_j, s)]^2. \tag{44}$$

RESULTADOS NUMÉRICOS

Datos con error absoluto constante

Para evaluar la bondad del método de ajuste descrito en la sección anterior, hemos utilizado la curva teórica $y(x) = \mathrm{sinc}(x)$, que aparece frecuentemente en la teoría de las transformadas de Fourier [5]

$$y(x) = \mathrm{sinc}(x) = \begin{cases} \dfrac{\sin x}{x}, & x \neq 0, \\ 1, & x = 0. \end{cases} \tag{45}$$

Hemos tomado $n = 100$ puntos equiespaciados en las abscisas en el intervalo con un $x \in [-2\pi, 2\pi]$, error absoluto normal en y_i de $\sigma = 0{,}03$. En la figura 1 se presenta la curva teórica (45) y los puntos de la dispersión. El error cuadrático medio de los puntos de la dispersión con respecto a la curva teórica

$$ECM_{th} := \sqrt{\frac{1}{n} \sum_{i=1}^{n} [y(x_i) - y_i]^2},$$

en el caso representado en la figura 1 tiene un valor cercano al error utilizado en la simulación, $\sigma = 0{,}03$

$$ECM_{th} \approx 0{,}0308197.$$

En la figura 2 se presentan los datos junto a las funciones interpoladoras de grado 3 de los datos y de los datos filtrados, aplicando la transformada discreta de Fourier. Resolviendo algorítmicamente (23) se ha obtenido para la señal filtrada,

$$k_{min} = 13.$$

Como se puede observar, la función interpoladora de los datos sin filtrar da una idea pobre de la curva teórica a la cual deben ajustarse los datos. El valor obtenido para s_{fit} a partir de la evaluación numérica de (13) es

$$s_{fit} \approx 0{,}29989. \tag{46}$$

También se puede obtener s_{fit} minimizando la función dada en (27)

$$s'_{fit} \approx 0{,}348215. \tag{47}$$

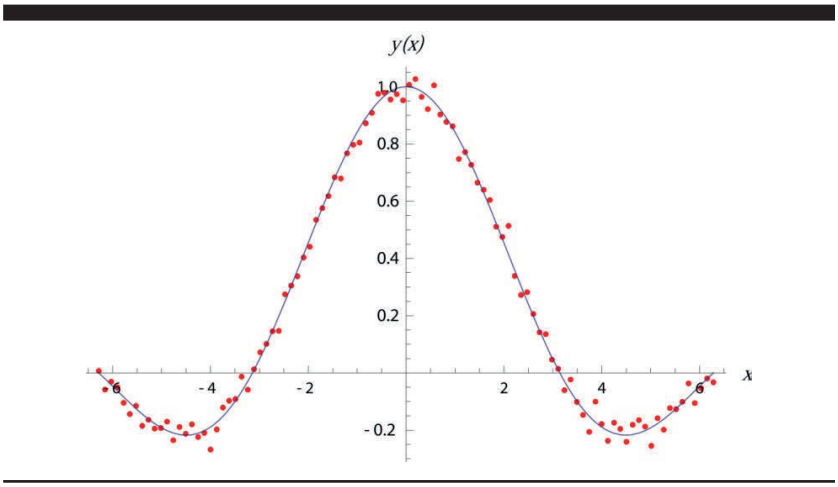

Figura 1. Dispersión de puntos obtenida a partir de $y(x) = \mathrm{sinc}(x)$.

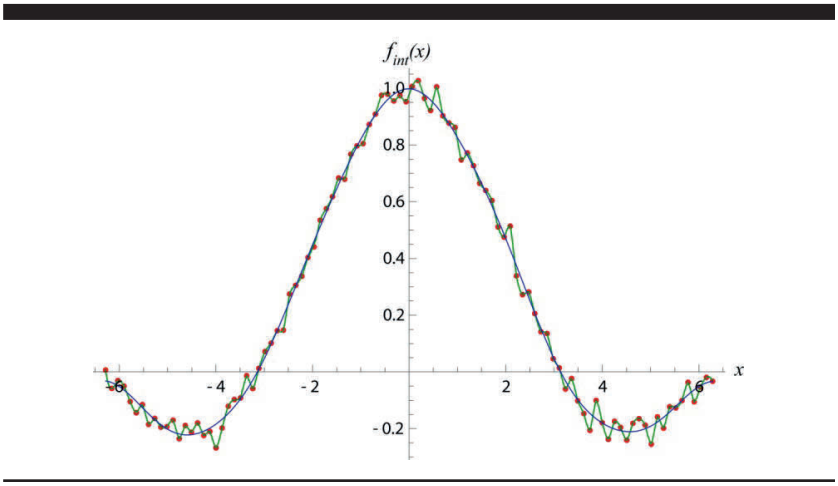

Figura 2. Representación de la dispersión de puntos (rojo) y de las funciones interpoladoras de los puntos (verde) y de los puntos suavizados por Fourier (azul).

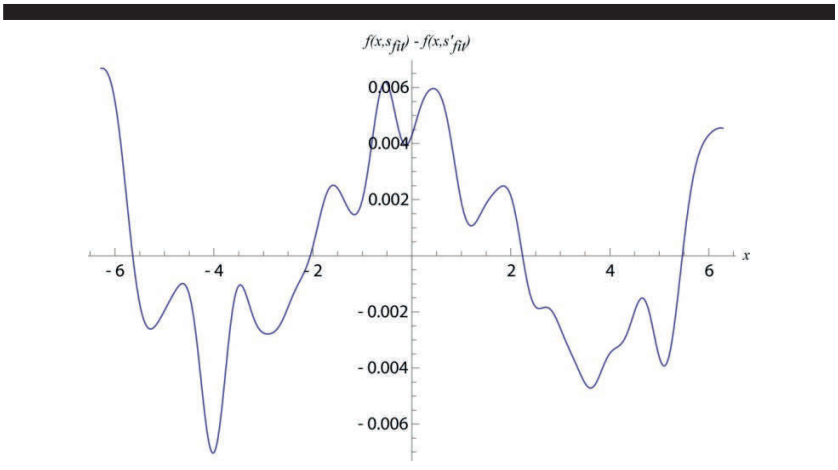

Figura 3. Diferencia entre las funciones de ajuste $f(x, s_{fit})$ y $f(x, s'_{fit})$ en el intervalo en el que se presentan los datos $x \in (-2\pi, 2\pi)$.

En la figura 3 se puede apreciar que la diferencia entre las funciones de ajuste $f(x,s_{fit})$ y $f(x,s'_{fit})$ queda un orden de magnitud por debajo del error $\sigma = 0{,}03$, por lo que resulta indistinto escoger cualquiera de los métodos desde el punto de vista de la bondad del ajuste. De todas maneras, en las gráficas que siguen en esta sección se ha escogido (46), pues numéricamente es bastante más rápido de calcular que (47).

En la figura 4 se ha representado la curva de ajuste $f(x)$ para la misma dispersión de puntos que aparece en la figura 1, así como la banda de confianza, que contiene un 66% de los puntos. Esto se corresponde bastante bien con el error normal que tienen los datos, pues teóricamente debería haber un $\mathrm{erf}(1/\sqrt{2}) \approx 68{,}2689\%$ los datos dentro de la banda de confianza. En la figura 5 se ha representado gráficamente las discrepancias entre la curva de ajuste $f(x)$ y la función interpoladora de los datos filtrados $f_{int}(x)$ con respecto a la curva teórica $y(x)$. Se puede observar que ambas discrepancias están contenidas en el error experimental de medición $\pm\sigma$, excepto en los extremos del intervalo, debido a que fuera del intervalo no tenemos puntos a los que ajustar.

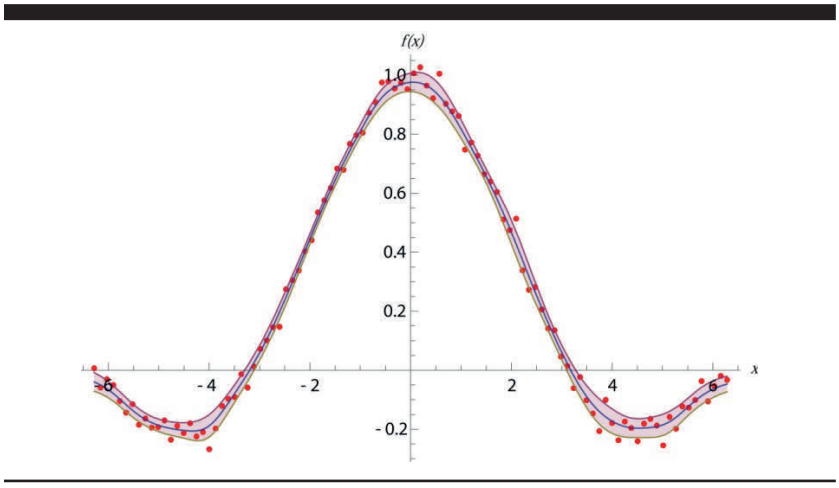

Figura 4. Función de ajuste y banda de confianza de los puntos obtenidos en la figura 1.

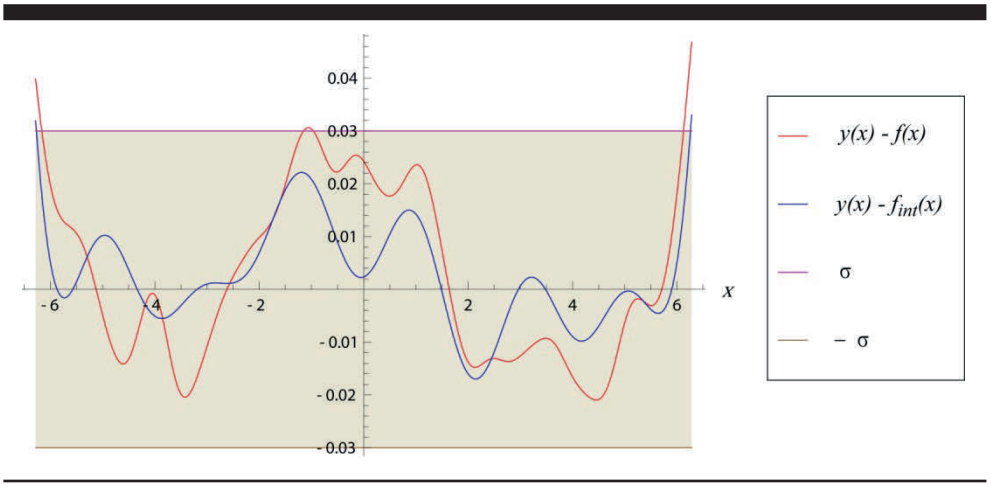

Figura 5. Representación gráfica de la diferencia entre la curva teórica $y(x)$ y, por un lado, la función de ajuste $f(x)$, y, por otro lado, la función interpoladora de los datos filtrados $f_{int}(x)$.

La integración de la curva teórica en el intervalo en que se presentan los datos da como resultado

$$\int_{-2\pi}^{2\pi} \operatorname{sinc}(x)\,dx = 2\operatorname{Si}(2\pi) \approx 2,8363.$$

Asimismo, la integración numérica de la función interpoladora de grado 3 de los datos filtrados $f_{int}(x)$ es

$$\int_{-2\pi}^{2\pi} f_{\text{int}}(x)\,dx \approx 2,8047.$$

Mientras que la función de ajuste es

$$\int_{-2\pi}^{2\pi} f(x)\,dx \pm \int_{-2\pi}^{2\pi} \Delta f(x)\,dx = 2,80 \pm 0,37.$$

En la figura 6 se presentan las derivadas de la función teórica, de ajuste e interpoladora: $y'(x)$, $f'(x)$, $f'_{\text{int}}(x)$ respectivamente. Como se puede apreciar, la reconstrucción de la función derivada por ambos métodos es similar y bastante aceptable, excepto de nuevo en los extremos del intervalo.

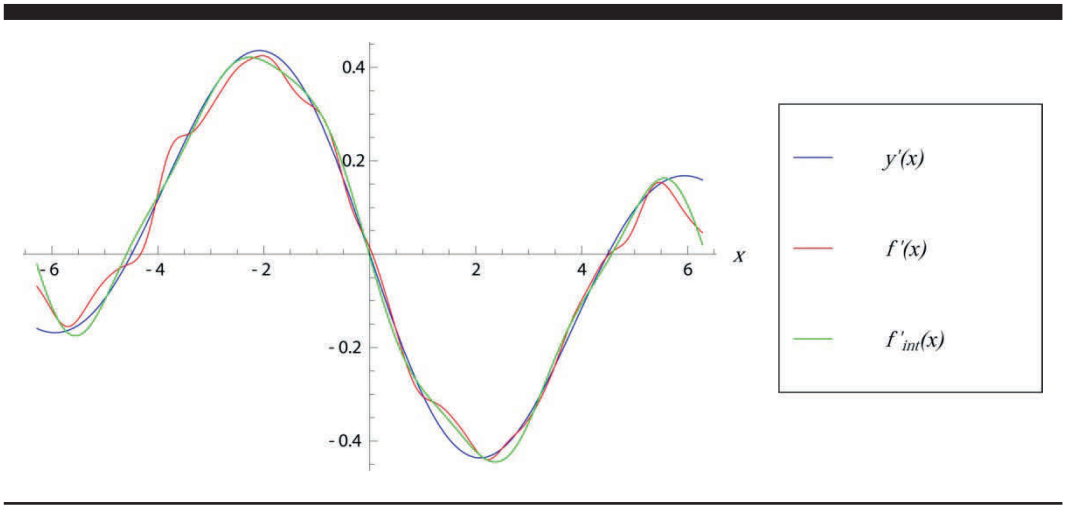

Figura 6. Derivada de la función $y(x)$ = sinc(x), junto con las derivadas de las funciones de ajuste e interpoladora: $f'(x)$, $f'_{\text{int}}(x)$ respectivamente.

Datos con error relativo constante

Para evaluar la bondad del ajuste descrito en la subsección "Datos con error relativo constante", hemos utilizado la siguiente función teórica

$$y(x) = x^2 + 10\sin 2x$$

Hemos tomado n = 100 puntos equiespaciados en las abscisas en el intervalo $x \in [4,10]$, con un error relativo normal en y_i de $\sigma_r = 0{,}05$. En la figura 7 se presenta la curva teórica $y(x)$ junto con los datos de la dispersión y_i y la curva de ajuste $f(x)$. En la figura 8 se presentan los datos y la función interpoladora de grado 3 entre los mismos.

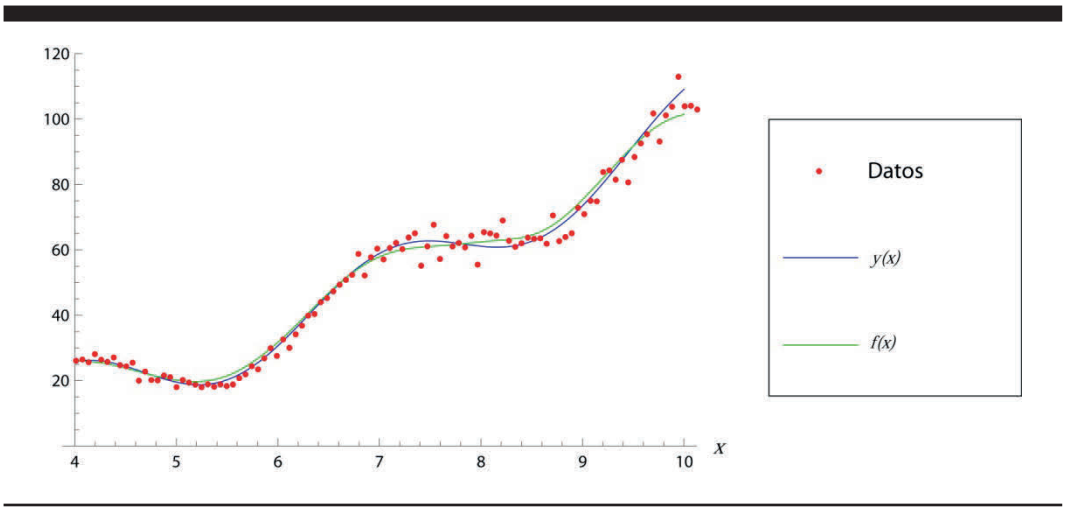

Figura 7. Representación de la curva de ajuste $f(x)$ de los datos y_i obtenidos a partir de la curva teórica $y(x)$.

Obsérvese que ahora no se puede aplicar la suavización de la curva mediante la transformada discreta de Fourier, porque en el error ya no es constante, sino que depende de la abscisa $\sigma(x)$. Como se puede observar, es difícil hacerse una idea de la curva teórica a partir de la curva interpoladora. En la figura 9 se ha representado gráficamente la curva de ajuste $f(x)$ para la misma dispersión de puntos que aparece en la figura 7, así como la banda de confianza, que contiene un 69% de los puntos. Una vez más, este resultado se corresponde bastante bien con el error normal que tienen los datos, pues teóricamente debería haber un 68,2689% de los datos dentro de la banda de confianza. El valor obtenido para s_{fit} a partir de la minimización numérica es

$$s_{fit} \approx 0,\, 267221.$$

En la figura 9 se puede apreciar la tendencia a que la banda de confianza se ensanche a medida que se incrementa el valor en la ordenada. Esto es debido a que el error relativo es el mismo para todos los datos, por lo que el error absoluto será tanto mayor cuanto mayor sea también el valor en la ordenada. En la figura 10 se ha representado la diferencia entre la curva teórica y la de ajuste, así como la banda del error $\pm\sigma(x)$. De nuevo, se observa que la discrepancia entre la curva teórica y la de ajuste se sale apreciablemente de la banda de error en el extremo derecho del intervalo. Esto es debido a que fuera del intervalo no tenemos puntos a los que ajustar. En la figura 11 se presentan las derivadas de la función teórica, de ajuste e interpoladora $y'(x)$, $f'(x)$, $f'_{\text{int}}(x)$, respectivamente. Como se puede apreciar, la reconstrucción de la función derivada es similar y bastante aceptable, excepto de nuevo en el extremo derecho del intervalo. La integración de la curva teórica en el intervalo en que se presentan los datos es

$$\int_4^{10} \left[x^2 + 10\sin 2x\right] dx \approx 309,672.$$

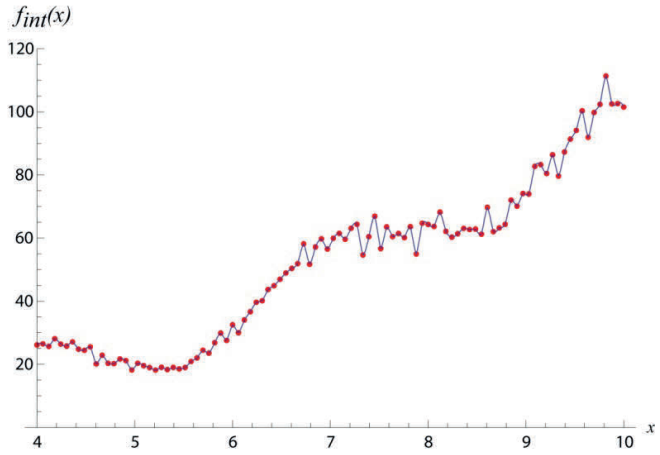

Figura 8. Interpolación de la misma dispersión de puntos que en la figura 7.

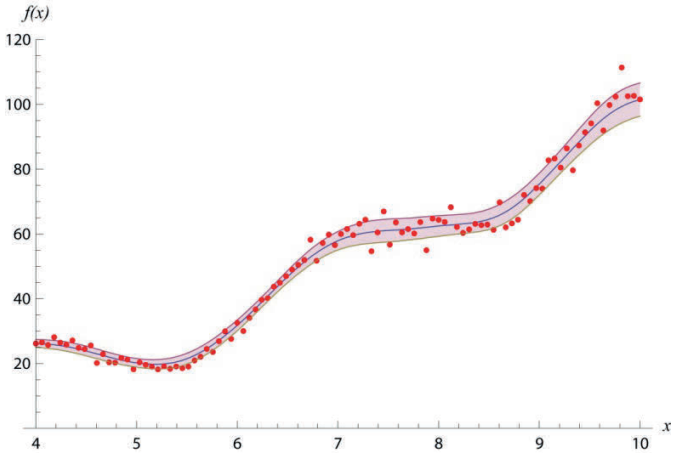

Figura 9. Función de ajuste y banda de confianza de los puntos obtenidos en la figura 7.

Asimismo, la integración numérica de la función interpoladora es

$$\int_4^{10} f_{\mathrm{int}(x)}\,dx \approx 310,672.$$

Mientras que la función de ajuste es

$$\int_4^{10} f(x)dx \pm \int_4^{10} \Delta f(x)dx = 310 \pm 16. \tag{48}$$

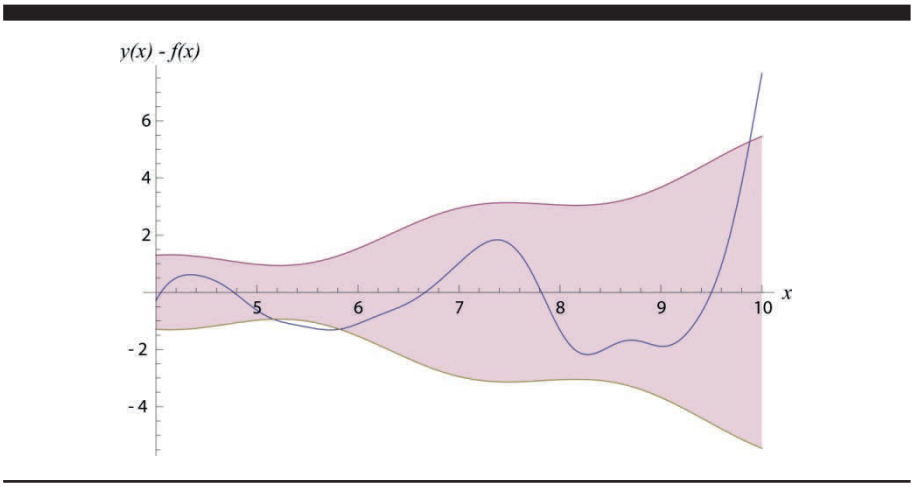

Figura 10. Representación gráfica de la diferencia entre la curva teórica $y(x)$ y la curva de ajuste $f(x)$. La parte sombreada corresponde a la banda de error $(-\sigma(x), \sigma(x))$.

Se puede observar que el error obtenido en (48) coincide con el error relativo utilizado en la simulación numérica

$$\frac{\int_4^{10} \Delta f(x)dx}{\int_4^{10} f(x)dx} = 0,0509997 \approx \sigma_r = 0,05.$$

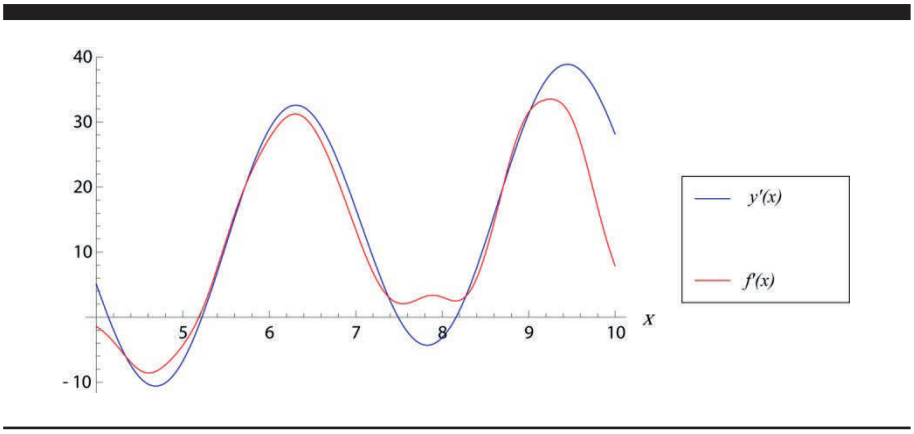

Figura 11. Derivada de la función $y(x) = x^2 + 10\sin(2x)$, junto a las derivadas de las funciones de ajuste e interpoladora $f'(x)$, $f'_{\text{int}}(x)$, respectivamente.

Datos con error desconocido

Para evaluar la bondad del ajuste descrito en la sección correspondiente, hemos utilizado la siguiente función teórica

$$y(x) = \log x + \cos x. \tag{49}$$

Hemos tomado n = 100 puntos equiespaciados en las abscisas en el intervalo $x \in [3,11]$, con un error normal que tiene un perfil lorentziano,

$$\sigma(x) = \sigma\left(\frac{3}{(x-x_o)^2+\delta}+1\right), \tag{50}$$

donde σ = 0,05, x_0 = 6 y δ = 0,3. En la figura 12 se presenta la curva teórica $y(x)$ junto a los datos de la dispersión y_i y la curva de ajuste $f(x)$, así como la función del error $\sigma(x)$. Se puede observar que los puntos presentan una mayor desviación con respecto a la curva teórica en la zona donde el error alcanza el máximo. En la figura 13 se ha representado gráficamente la curva de ajuste $f(x)$, junto a la banda de confianza, que incluye un 69% de los datos (lo cual concuerda con la distribución normal del error). El valor obtenido para s_{fit} a partir de la minimización numérica de (44) es

$$s_{fit} \approx 0,278353.$$

En la figura 13 se puede apreciar que la banda de confianza se ensancha precisamente en la zona donde se tiene un mayor error, es decir, en torno al punto x_0 = 6, donde según (50) $\sigma(x)$ presenta el máximo. En la figura 14 se han representado la curva teórica $\sigma(x)$ y el error de ajuste $\Delta f(x)$.

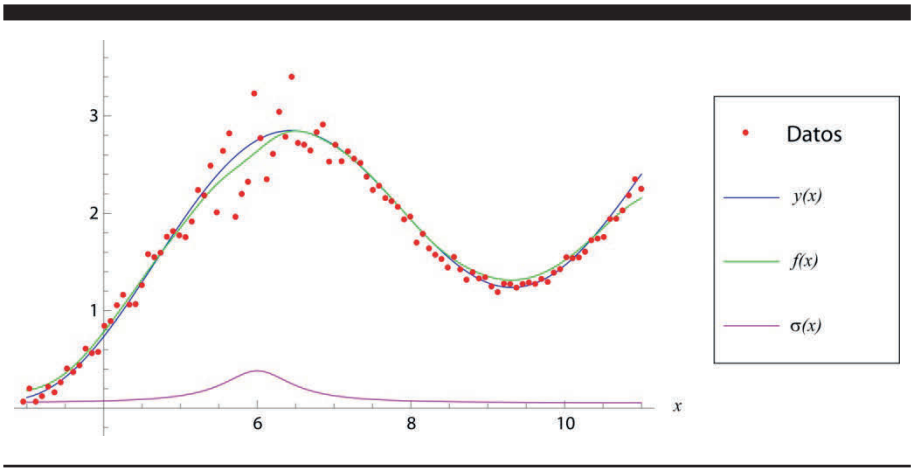

Figura 12. Representación de la curva de ajuste $f(x)$ de los datos y_i obtenidos a partir de la curva teórica $y(x)$, así como la función del error de los datos $\sigma(x)$.

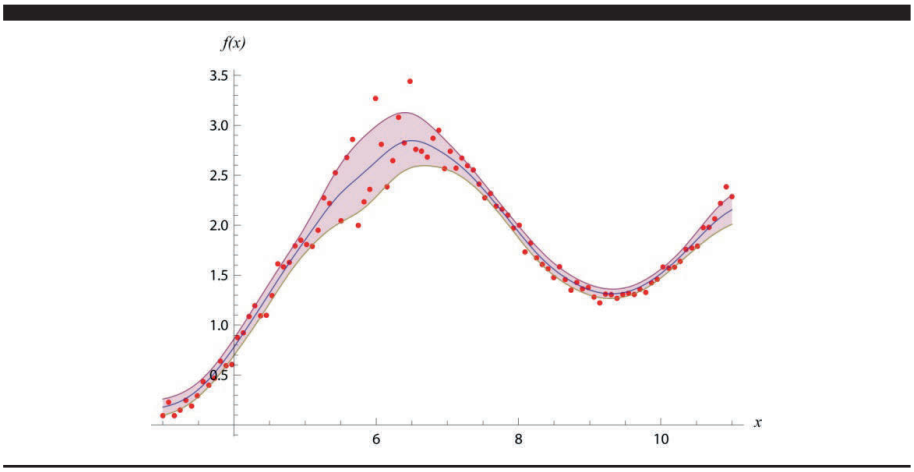

Figura 13. Función de ajuste y banda de confianza de los puntos obtenidos en la figura 12.

Como se puede observar, el error en el ajuste reproduce bastante bien el perfil del error introducido en la simulación de los datos, excepto en el extremo derecho del intervalo. En la figura 15 se presentan las derivadas de la función teórica y de ajuste: $y'(x)$, $f'_{\text{int}}(x)$, respectivamente. Como se puede apreciar, la reconstrucción de la función derivada es peor en la zona donde el error experimental que presentan los datos es mayor; y también, una vez más, en el extremo derecho del intervalo.

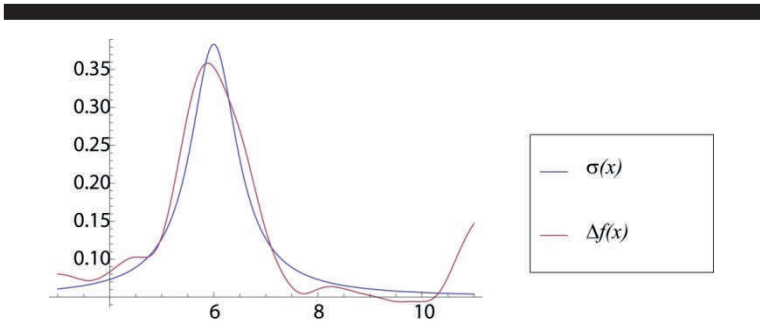

Figura 14. Comparación del error teórico $\sigma(x)$ y del error del ajuste $\Delta f(x)$.

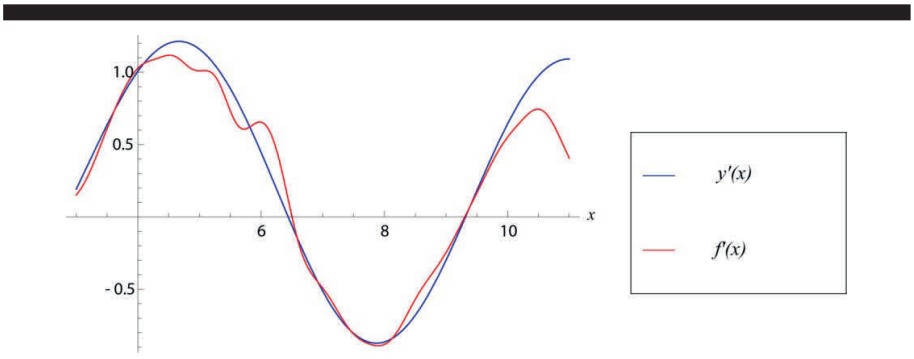

Figura 15. Derivada de la función $y(x) = \log x + \cos x$, junto a las derivadas de las funciones de ajuste e interpoladora: $f'(x)$, $f'_{\text{int}}(x)$, respectivamente.

La integración de la curva teórica en el intervalo en que se presentan los datos es

$$\int_3^{11}[\log x+\cos x]dx\approx 13,9399.$$

Asimismo, la integración numérica de la función interpoladora es

$$\int_3^{11} f_{\text{int}}(x)dx\approx 13,946.$$

Mientras que la con la función de ajuste es

$$\int_3^{11} f(x)dx\pm\int_3^{11}\Delta f(x)dx=13,93\pm 0,97.$$

Ajuste del IBEX35

Los resultados obtenidos en el apartado anterior permiten tener confianza en el método de ajuste propuesto anteriormente aplicado ahora a una serie temporal de datos en la que se desconozca su variabilidad respecto a una cierta tendencia, también desconocida. Este es el caso, por ejemplo, de los índices de bolsa. En la figura 16 se ha representado gráficamente la banda de confianza para la serie temporal de los datos diarios a cierre de bolsa del índice IBEX35, desde el 18 de enero hasta el 8 de junio de 2010.

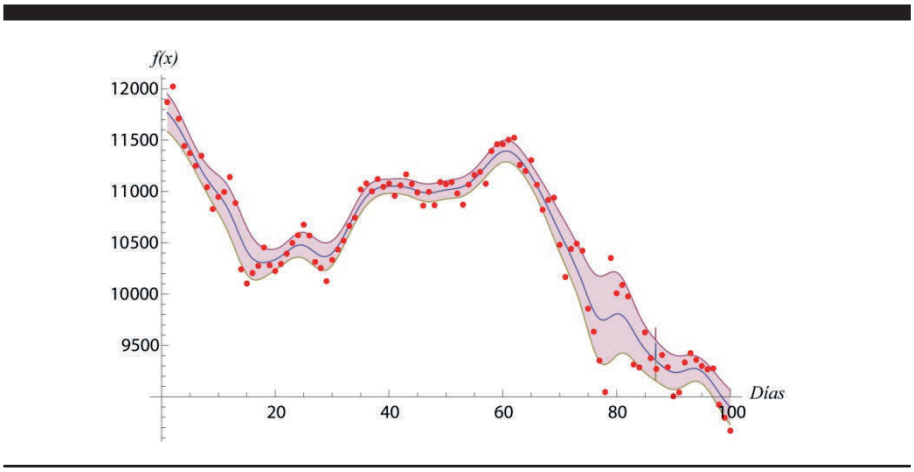

Figura 16. Banda de confianza para el IBEX35.

Dentro de la banda de confianza se encuentra el 64% de los datos, lo cual indica que el error es cercano al normal. Como se puede apreciar, la banda de confianza ofrece una medida de la volatilidad en función del tiempo.

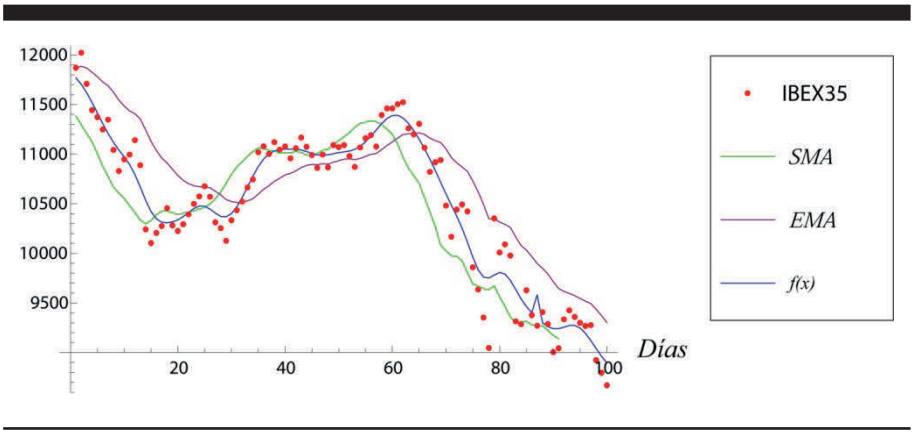

Figura 17. Comparación de las medias móviles simple SMA y exponencial EMA con la función de ajuste propuesta *f*.

En la figura 17 se pueden comparar las suavizaciones típicas que se utilizan en economía [6] (media móvil simple SMA y media móvil exponencial EMA) con la función de ajuste propuesta *f*. En la media móvil simple se ha tomado un intervalo de $N = 10$ días, mientras que en la media móvil exponencial se ha tomado un factor $\alpha = 0{,}1$. Como se aprecia en la figura 17, la media móvil simple tiende a subestimar los datos cuando la tendencia es a la baja y a sobreestimar cuando la tendencia es al alza; mientras que en la media móvil exponencial ocurre lo contrario. La función de ajuste propuesta *f* tiende a ajustarse a los datos, quedando entre la media móvil simple y exponencial. Un inconveniente que presentan las medias móviles es que no existe un método general que determine los parámetros N o α óptimos a partir de los datos.

Además, la suavización que ofrecen las medias móviles consiste en una sucesión de puntos que requieren de la interpolación para obtener una curva suave.

Por último, cabe señalar que para la evaluación numérica y la representación gráfica de todos los resultados obtenidos en esta sección se ha utilizado MATHEMATICA.

CONCLUSIONES

Se ha propuesto la forma de una curva de ajuste para una dispersión cualquiera de una gran cantidad de puntos experimentales, así como la banda de confianza en la que se debe encontrar dicha curva de ajuste. Para comprobar la bondad de la curva de ajuste propuesta, se ha simulado numéricamente una dispersión de puntos a partir de una curva teórica y un cierto ruido aleatorio.

Por un lado, se han considerado los dos tipos de ruido más típicos en las medidas experimentales, error absoluto y error relativo constantes, ambos con una distribución normal. Los puntos obtenidos de esta forma se han ajustado según el método propuesto y se ha comprobado que se reproduce la curva teórica dentro del error introducido en los datos. Se ha comprobado que el método de suavización basado en la transformada discreta de Fourier ofrece unos resultados similares a los de la curva de ajuste propuesta, pero sólo se puede utilizar cuando los datos presentan un error absoluto constante. Asimismo, la integración numérica de los datos a partir de la suavización de Fourier no ofrece el error cometido en la evaluación de la integral. Sin embargo, esto es posible a partir de integración de la banda de confianza de la curva de ajuste propuesta.

Por otro lado, tomando un ruido no constante en la simulación de los datos, no solo se ha podido reproducir la curva teórica, sino que también se ha reproducido, a partir de los datos, el perfil del ruido introducido. Este resultado ha permitido comparar nuestra propuesta de ajuste a una serie temporal de un índice de bolsa (IBEX35), con las suavizaciones que normalmente se utilizan en Economía. El método propuesto no sólo se ajusta mejor, sino que además permite medir la volatilidad en función del tiempo a partir de la banda de confianza. Además, el método propuesto establece el parámetro óptimo de ajuste a partir de los mismos datos, lo cual no ocurre en las suavizaciones típicas utilizadas en Economía, como la media móvil y la media móvil exponencial.

BIBLIOGRAFÍA

[1] J. Stoer and R. Bulirsch 1980. Introduction to Numerical Analysis, SpringerVerlag, New York.
[2] N. Ahmed, T. Natarajan, K. R. Rao, Discrete Cosine Transform, IEEE Trans. Computers C-23 (1974)90-93.
[3] J. Xiuyu Cong 2003. Historical Development of Central Limit Theorem.
[4] J. Spanier, K. B. Oldham1987. An atlas of functions, Hemisphere Publising Corporation.
[5] E. W. Weisstein, Sinc Function. MathWorld-A Wolfram Web Resource. http://mathworld.wolfram.com/SincFunction.html
[6] J. F. Kenney, E. S. Keeping 1962. Moving Averages §14.2 en Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, 221-223.
[7] A. Korganoff 1961. Méthodes de Calcul Numérique, Tome 1, Dunod, París.
[8] A. Korganoff and M. Pavel-Parvu 1961. Méthodes de Calcul Numérique, Tome 2, Dunod, París.