Rev.Acad.Canar.Cienc., IV (núms. 1 y 2),45-53 (1992)

APLICACIONES DEL CAOS DETERMINISTA A LA DETECCION DE OUTLIERS

Fernando Fernández Rodríguez.
Departamento de Economía Aplicada.

José Miguel Pacheco Castelao.
Departamento de Matemáticas.

Universidad de Las Palmas de Gran Canaria.

ABSTRACT

In this paper the authors show that techniques employed in the prediction of chaotic time series can also be applied to detection of outliers. A definition of outlier is provided and a theorem on hypothesis testing is also proved.

Introducción : Conceptos y definiciones básicas

Una definición intuitiva de outlier en una serie temporal de datos es la siguiente: Una observación cuya discordancia con los restantes valores es excesiva con relación al modelo explicativo seleccionado para la serie. Véase Rosado (1990) y referencias en él.

Cuando se intentan efectuar predicciones sobre una serie temporal, la presencia de outliers es siempre un problema. Sin embargo, se probará que el desarrollo de métodos predictivos para series caóticas (Farmer y Sidorowich (1987)) permite formular

criterios de detección de outliers y de zonas turbulentas en las series.

El instrumento esencial en teoría de la predicción es el espacio de fases, construído de la siguiente manera:
Dados una serie temporal finita $(x_1,.....,x_N)$ y dos enteros positivos m y τ, para cada i en su rango de valores posibles se define la m-historia con retardo τ como :

$$x_i^{m,\tau} = (x_i , x_{i-\tau} , ... , x_{i-\tau(m-1)})$$

Normalmente, en las aplicaciones se tomará $\tau = 1$. El entero m se conoce como dimensión de inmersión . Las m-historias son vectores de un espacio m-dimensional real, que es precisamente el espacio de fases asociado a la serie dada.

Si la serie no está muestreada en un ruido, una selección adecuada de m y τ permitirá - por lo general - reconocer que las m-historias forman un atractor extraño , reflejo de alguna dinámica determinista subyacente en la serie. Tales series se denominan series caóticas .

El teorema de Takens (Takens (1981)) indica que si un sistema dinámico n-dimensional es estudiado a través de un único observable - la serie temporal -, entonces de modo genérico, la dinámica que se reconstruye en el espacio de fases es equivalente a la del sistema original siempre que $m > 2n$. En Fernández (1992) se presentan diversos procedimientos de encontrar la dimensión de inmersión adecuada para una serie temporal.

La reconstrucción de la ley dinámica no lineal que genera una serie se realiza por ajuste de un vector $\vec{\alpha}$ de parámetros y de una función f

$$x_{i+1} = f(x_i , x_{i-1} , ... , x_{i-(m-1)}, \vec{\alpha}) + e_{i+1} \quad (m-1 < i < N)$$

minimizando la suma de los errores cuadráticos e_i. En Casdagli

(1989) se estudian diferentes formas para la relación funcional f.

La reconstrucción global de una dinámica determinista puede fracasar según el número de datos disponibles. Ello es evidente si hay muy pocos, pero un número elevado producirá problemas de sobredeterminación del número de parámetros del modelo que en nada disminuyen los errores de predicción (Farmer y Sidorowich (1988)). Por esta razón se han introducido las técnicas de predicción por analogías mediante representaciones locales de la dinámica que permiten mejorar en muchos casos las predicciones a corto plazo.

Sean la serie temporal y su dimensión de inmersión m. Dada la m-historia del último elemento de la serie, x_N^m , se considera el conjunto de las k m-historias más próximas a ella dentro del espacio de fases (ver en Fernández (1992) los detalles y formalización):

$\{ x_{j_1}^m , \ldots\ldots , x_{j_k}^m \;/\; d_1 = \| x_{j_1}^m - x_N^m \|$ son los k menores valores obtenidos al calcular la distancia entre x_N^m y las restantes m-historias del espacio de fases $\}$

De esta forma, la predicción $x_N \longrightarrow \hat{x}_{N+1}$ se formula mediante alguna interpolación del tipo

$$\hat{x}_{N+1} = \Phi (x_{j_1+1}, \ldots\ldots, x_{j_k+1})$$

entre los valores consecutivos a los x_{j_1}. Tomando las precauciones pertinentes, esto es, $\forall k$, $N - j_k > s$, es posible formular la predicción $x_N \longrightarrow \hat{x}_{N+s}$ de modo análogo.

La expresión funcional más simple para Φ la da el cálculo del baricentro de los puntos $x_{j_1+s}^s$:

$$\hat{x}_{N+s}^{bar} = \frac{1}{k} \sum_{l=1}^{k} x_{j_l+s}$$

Se pueden generalizar este predictor considerando combinaciones lineales convexas . Nos referimos con ello a los predictores simpliciales

$$\hat{x}^{simp}_{N+s} = \sum_{l=1}^{k} \alpha_l x_{j_l+s} \quad , \quad \text{donde} \quad \sum_{l=1}^{k} \alpha_l = 1$$

Una primera alternativa en este sentido será suponer que los puntos más próximos al x^m_N en el espacio de fases deben influir más en la predicción. Estaremos considerando en este caso que

$$\alpha_l = (1/d_l) \; / \; \sum_{l=1}^{k} (1/d_l) \quad \text{donde} \quad d_l = \| x^m_N - x^m_{j_l} \|$$

Una segunda alternativa, interesante cuando la ley dinámica que genera los datos va cambiando paulatinamente con el tiempo, es considerar que α_l es función decreciente de $N - j_l$, es decir, que la influencia de cada m-historia del pasado tiene una influencia en el predictor que va disminuyendo con el transcurso del tiempo.

Para obtener una predicción independiente de cualquier consideración sobre los puntos del espacio de fases, se puede considerar un predictor simplicial con coeficientes aleatorios obtenido del modo siguiente: Se muestrean r veces k números aleatorios λ_1, ... , λ_k en una distribución uniforme y cada vez se consideran los coeficientes $\alpha_l = \lambda_l / \sum_{l=1}^{k} \lambda_l$. Con ello se construye el predictor simplicial siguiente promediando sobre las r relaciones del muestreo:

$$\hat{x}^{simp}_{N+s} = \frac{1}{r} \sum_{p=1}^{r} \left(\sum_{i_p=1}^{k} \alpha_{i_p} x_{j_{i_p}+s} \right)$$

Se plantean los siguientes problemas abiertos: ¿ influye decisivamente el valor r en los resultados de la predicción ? ¿ Es posible hallar un valor óptimo de k ?

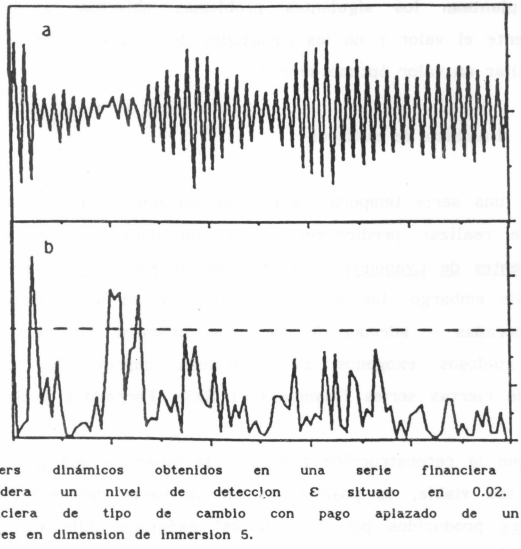
Detección dinámica de outliers

Para una serie temporal caótica el horizonte dentro del cual es posible realizar predicciones viene cuantificado a través de los exponentes de Lyapunov y la entropía de Kolmogorov (Schuster (1988)). Sin embargo, las series que aparecen en las aplicaciones están generadas - además de por una dinámica interna - por "shocks" ruidosos exógenos de naturaleza imprevisible. Es el ejemplo de ciertas series financieras (Bajo, Fernández y Sosvilla (1992a)).

Aunque la reconstrucción global de la dinámica que genera la serie no sea viable, es posible separar la parte determinista de los outliers producidos por los "shocks" exógenos. Ello es cierto independientemente del modelo seleccionado para explicar la dinámica determinista. Se dirá, por tanto, que el método de detección de outliers es robusto.

Definición : "Se dice que el dato x_1 de la serie temporal $\{x_1,....x_N\}$ es un outlier dinámico si la predicción por analogías a partir de la m-historia x^m_{1-1} se efectua con un error mayor que un valor $\varepsilon > 0$ prefijado".

Nótese que en esta definición existe un componente subjetivo en la elección del valor ε. Esta selección se hará de acuerdo con el contenido y significado de los datos que conforman la serie.



Outliers dinámicos obtenidos en una serie financiera donde se considera un nivel de detección ε situado en 0.02. a) Serie financiera de tipo de cambio con pago aplazado de un mes. b) Errores en dimensión de inmersión 5.

La distribución de los errores e_i es desconocida de antemano y la detección de outliers se traduce, por tanto, en estudiar la distribución de los errores e_i y formular los contrastes de hipótesis adecuados.

Debido a la sencillez matemática del predictor simplicial las hipótesis sobre la distribución de la serie original de datos x_i pueden traducirse facilmente a hipótesis sobre los errores e_i; ello permite establecer contrastes acerca de la distribución original de los elementos de la serie.

El caso de más interés, por sus implicaciones prácticas, consiste en distinguir un ruido blanco de una dinámica determinista endógena no lineal y por tanto predecible a corto plazo. Tal es el caso de muchas series financieras (Bajo, Fernández y Sosvilla (1992b)).

Sea la serie $\{x_1, \ldots, x_N\}$, con un espacio de fases m-dimensional; se considera la predicción baricéntrica con k

m-historias de x_{N+1}

$$\hat{x}^{bar}_{N+1} = \frac{1}{k} \sum_{l=1}^{k} x_{j_l+1}$$

Si la serie está generada por variables aleatorias independientes idénticamente distribuidas N(0,σ), el error

$$\varepsilon_{N+1} = x_{N+1} - \hat{x}^{bar}_{N+1} = x_{N+1} - \frac{1}{k} \sum_{l=1}^{k} x_{j_l+1}$$

es una variable aleatoria que sigue una distribución N(0,σ_1), donde la desviación típica es

$$\sigma_1 = (\sigma^2 + \frac{1}{k^2} \sum_{l=1}^{k} \sigma^2)^{1/2} = (\sigma^2 + \frac{k\sigma^2}{k^2})^{1/2} = \sigma(1 + \frac{1}{k})^{1/2}$$

Del razonamiento anterior se deduce inmediatamente el siguiente

Teorema

"Dada una serie temporal $\{x_1, \ldots, x_N\}$, el siguiente contraste de hipótesis permite discriminar entre la procedencia puramente aleatoria de la serie o la existencia de una dinámica no lineal endógena determinista:

H_0 : La variable ε_{N+1} sigue una distribución N(0,σ_1)

H_1 : La variable ε_{n+1} sigue una distribución N(0,σ^*) con $\sigma^* < \sigma_1$"

Para investigar, en la práctica, si la distribución de los errores de un número n de predicciones se corresponde con la de la hipótesis H_0, se considera un nivel α de significación y se define una región crítica R para una χ^2 con n-1 grados de libertad.

REFERENCIAS

Bajo, O.; F. Fernández y S. Sosvilla

(1992a) "Chaotic Behaviour in Exchange-Rate Series: First Result for the Peseta-U.S. Dollar Case". De próxima aparición en Economics Letters.

(1992b) "Volatilidad y predecibilidad en las series del tipo de cambio peseta-dolar: Un enfoque basado en el caos determinista". De próxima aparición en Revista Española de Economía.

Casdagli, M. (1989) "Nonlinear Prediction of Chaotic Time Series" Physica D 35, 335-356.

Farmer,J.D. y J. Sidorowich

(1987) "Predicting Chaotic Time Series", Physical Review Letters. Vol 59, 8, 845-848.

(1988) "Exploiting Chaos to Predict the Future and Reduce Noise". In Evolution, Learning and Cognition. Editado por Y.C. Lee, World Scientific Press, pp 27.

Fernández, F. (1992) "El problema de la predicción en series temporales: Aplicaciones del caos determinista". Tesis Doctoral. Departamento de Economía Aplicada. Universidad de Las Palmas de G.C.

Rosado, F. (1990) "Selection of multivariate influencial observations". COMPSTAT-9th Symposium on Computational Statistic.

Schuster, H.G. (1988) Deterministic Chaos. An Introduction. VCH. Weinheim.

Takens, F. (1981) "Detecting strange attractors in turbulence". In Lect. Notes Math. 898. Editado por D.A. Raud and L.S. Young. Springer-Verlag. New York.

Recibido: 20 de Julio de 1992