Aspectos generales sobre la presencia de datos atípicos en series temporales

Pedro Galeano

Departamento de Estadística
Universidad Carlos III de Madrid

CAEPIA 2011 - I Workshop on Time Series
8 de Noviembre, 2011

Contenidos

1. Introducción

2. Atípicos en series temporales univariantes

   (a) Tipos de atípicos
   (b) Procedimientos habituales de detección basados en Chen y Liu (1993)
   (c) Procedimiento de detección de Galeano y Peña (2012)

3. Atípicos en series temporales multivariantes

   (a) Tipos de atípicos
   (b) Procedimiento de detección de Tsay, Peña y Pankratz (2000)
   (c) Procedimiento de detección de Galeano, Peña y Tsay (2006)

4. Conclusiones

Contenidos

1. **Introducción**

2. Atípicos en series temporales univariantes
   (a) Tipos de atípicos
   (b) Procedimientos habituales de detección basados en Chen y Liu (1993)
   (c) Procedimiento de detección de Galeano y Peña (2012)

3. Atípicos en series temporales multivariantes
   (a) Tipos de atípicos
   (b) Procedimiento de detección de Tsay, Peña y Pankratz (2000)
   (c) Procedimiento de detección de Galeano, Peña y Tsay (2006)

4. Conclusiones

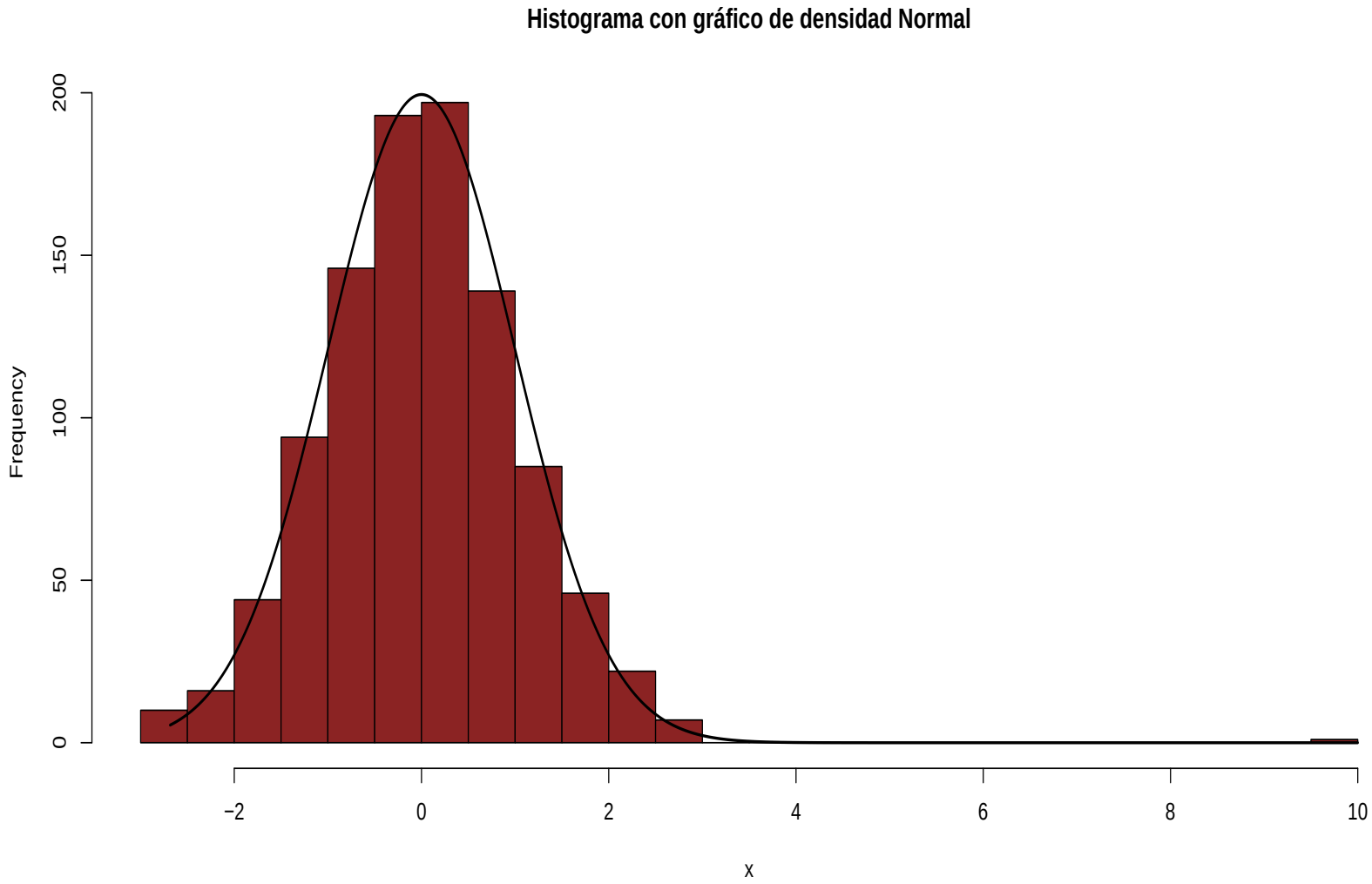
Observaciones anómalas

- Las técnicas estadísticas más habituales para el análisis de datos están basadas en **modelos** que suponen un cierto número de supuestos. Por ejemplo:
  - En un modelo de regresión múltiple, que entre respuesta y regresores existe una relación lineal y los errores del modelo son Gaussianos.
  - En problemas de clasificación, que los datos han sido generados por una mixtura de distribuciones Gaussianas.
  - En análisis de colas, que los tiempos entre llegadas son exponenciales.

- Sin embargo, en la práctica, puede ocurrir que el modelo supuesto es capaz de describir acertadamente una gran parte de los datos, pero un pequeño grupo de observaciones parecen seguir un comportamiento diferente.

- Estas observaciones anómalas han recibido diferentes nombres en diferentes áreas como, por ejemplo, observaciones atípicas, anómalas, discordantes, excepcionales, defectuosas, aberrantes, contaminantes, nocivas, etc...

Histograma de un conjunto de datos que incluyen un atípico



Datos atípicos

- Podemos ignorar la presencia de estos **datos atípicos**, pero entonces podemos perjudicar el análisis de múltiples formas dependiendo del tipo de datos con los que estamos trabajando, siendo los más habituales:

  1. introducir sesgo en los estimadores de los parámetros del modelo;
  2. influir en la potencia de contrastes basados en dichas estimaciones;
  3. ampliar los intervalos de confianza para los parámetros del modelo;
  4. influir en las predicciones.

Maneras de tratar con datos atípicos

1. Suponer un modelo más complejo:
   - Ventaja: los atípicos pueden indicar que estamos utilizando un modelo erróneo.
   - Desventaja: nos puede llevar a un modelo excesivamente complicado que explique bien un número muy reducido de observaciones y menos bien el resto.

2. Uso de estimadores robustos:
   - Ventaja: son estimaciones fiables de los parámetros del modelo supuesto.
   - Desventajas: (1) suelen ignorar la información proporcionada por los atípicos, y; (2) difíciles de obtener en modelos más complicados.

3. Uso de procedimientos de detección de atípicos:
   - Ventajas: (1) Permiten usar un modelo simple que incorpora la información proporcionada por los datos atípicos, y; (2) a veces utilizan estimaciones robustas.
   - Desventaja: Son algoritmos complicados y con diferentes problemas.

Los datos atípicos proporcionan información interesante

- Los atípicos suelen proporcionar información en una gran variedad de aplicaciones.

  1. Datos atípicos en registros médicos pueden indicar el empeoramiento de un paciente enfermo o la presencia de una enfermedad en una persona que estaba sana.
  2. Datos atípicos en movimientos con tarjetas de crédito pueden indicar robos o malos usos de dichas tarjetas.
  3. Lecturas anómalas en un avión pueden indicar el fallo de alguno de los componentes.
  4. Tráfico excesivo en un ordenador perteneciente a una red puede indicar que un ordenador pirateado está enviado datos a un destino no autorizado.

Datos atípicos en series temporales

- En **series temporales**, los datos atípicos son más complejos que en otro tipo de modelos debido a la estructura temporal. Por esto, se consideran diferentes tipos de datos atípicos con efectos diferentes.

- A continuación, se presentan diferentes tipos de datos atípicos en series temporales univariantes y multivariantes lineales, junto con algunos procedimientos de detección. Se dará especial atención a propuestas realizadas junto con Daniel Peña (Universidad Carlos III de Madrid) y Ruey S. Tsay (Universidad de Chicago).

Contenidos

1. Introducción

2. **Atípicos en series temporales univariantes**

   (a) Tipos de atípicos
   (b) Procedimientos habituales de detección basados en Chen y Liu (1993)
   (c) Procedimiento de detección de Galeano y Peña (2012)

3. Atípicos en series temporales multivariantes

   (a) Tipos de atípicos
   (b) Procedimiento de detección de Tsay, Peña y Pankratz (2000)
   (c) Procedimiento de detección de Galeano, Peña y Tsay (2006)

4. Conclusiones

Modelo ARIMA(p, d, q)

- La mayor parte de la literatura sobre datos atípicos se centra en modelos lineales.
- Modelo ARIMA(p, d, q) univariante:

$$\phi(B)\,(1-B)^d\,x_t = c + \theta(B)a_t, \qquad a_t \sim N\left(0, \sigma_a^2\right)$$

donde B es el operador retardo tal que $Bx_t = x_{t-1}$, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ y $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$.

- Las representaciones autorregresiva y de media móvil están dadas por:

$$\pi(B)x_t = c_\pi + a_t, \qquad x_t = c_\psi + \psi(B)a_t$$

donde $\pi(B) = \phi(B)\,(1-B)^d\,\theta(B)^{-1}$ y $\psi(B) = \theta(B)\phi(B)^{-1}\,(1-B)^{-d}$.

Datos atípicos en modelos ARIMA(p, d, q)

- Debido a la dependencia temporal, se han considerado diferentes datos atípicos en modelos ARIMA(p, d, q).

- Los cuatro efectos más usuales son:

  1. Atípico aditivo.
  2. Atípico innovativo.
  3. Cambio de nivel.
  4. Cambio transitorio.

- Otros posibles efectos son:

  1. Rachas de atípicos aditivos.
  2. Cambios de tendencia.

Atípico aditivo

- Fox (1972) introduce los atípicos aditivo e innovativo.
- Un **atípico aditivo** (**AO**) es un atípico que afecta exclusivamente a una observación.
- Una serie afectada por un AO en $t = h$ se define como sigue:

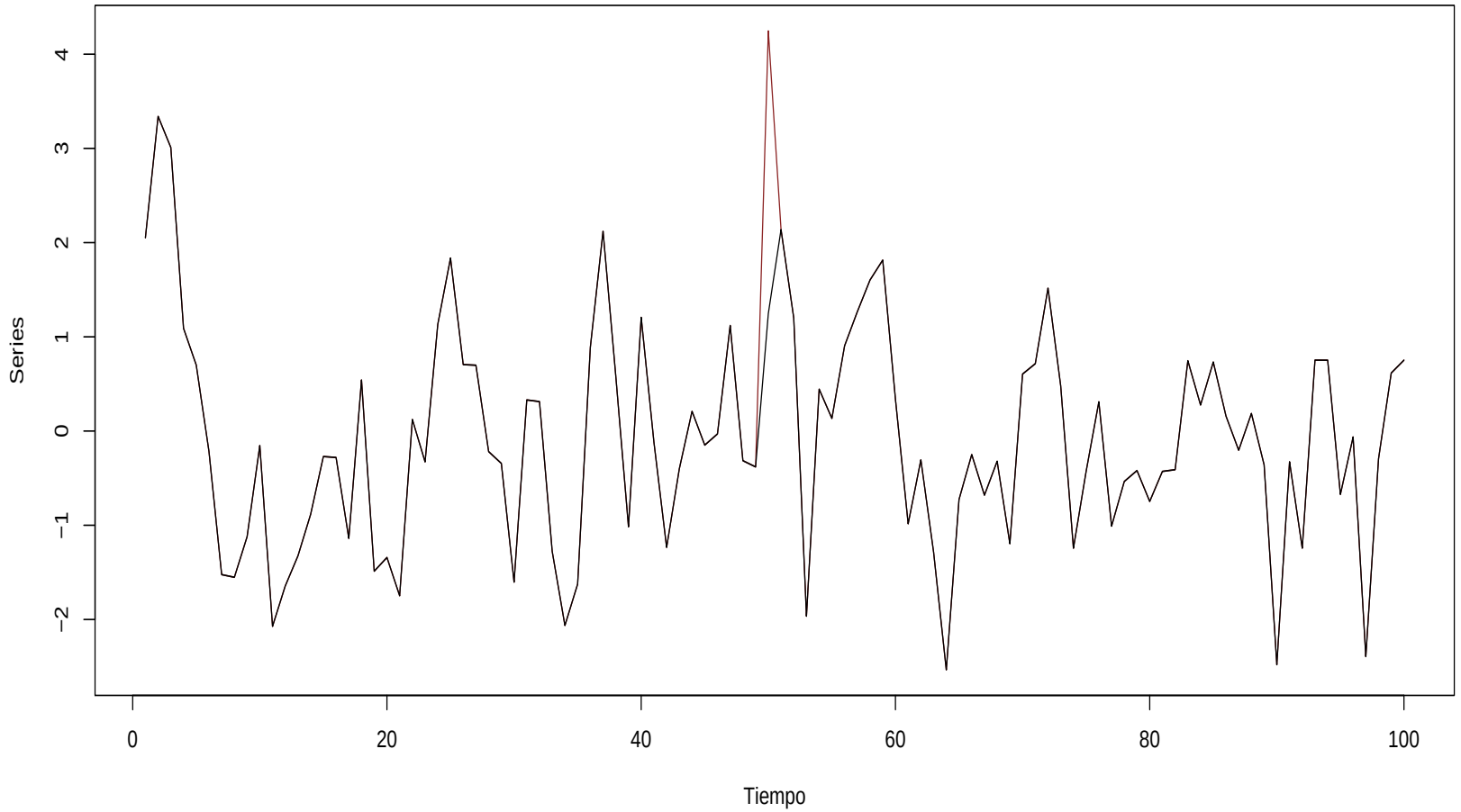
$$y_t = x_t + wI_t^{(h)}$$

donde x_t es una serie temporal que sigue un modelo ARIMA(p, d, q), $I_t^{(h)} = 1$, si $t = h$ e $I_t^{(h)} = 0$, si $t \neq h$ y w es el tamaño del atípico.

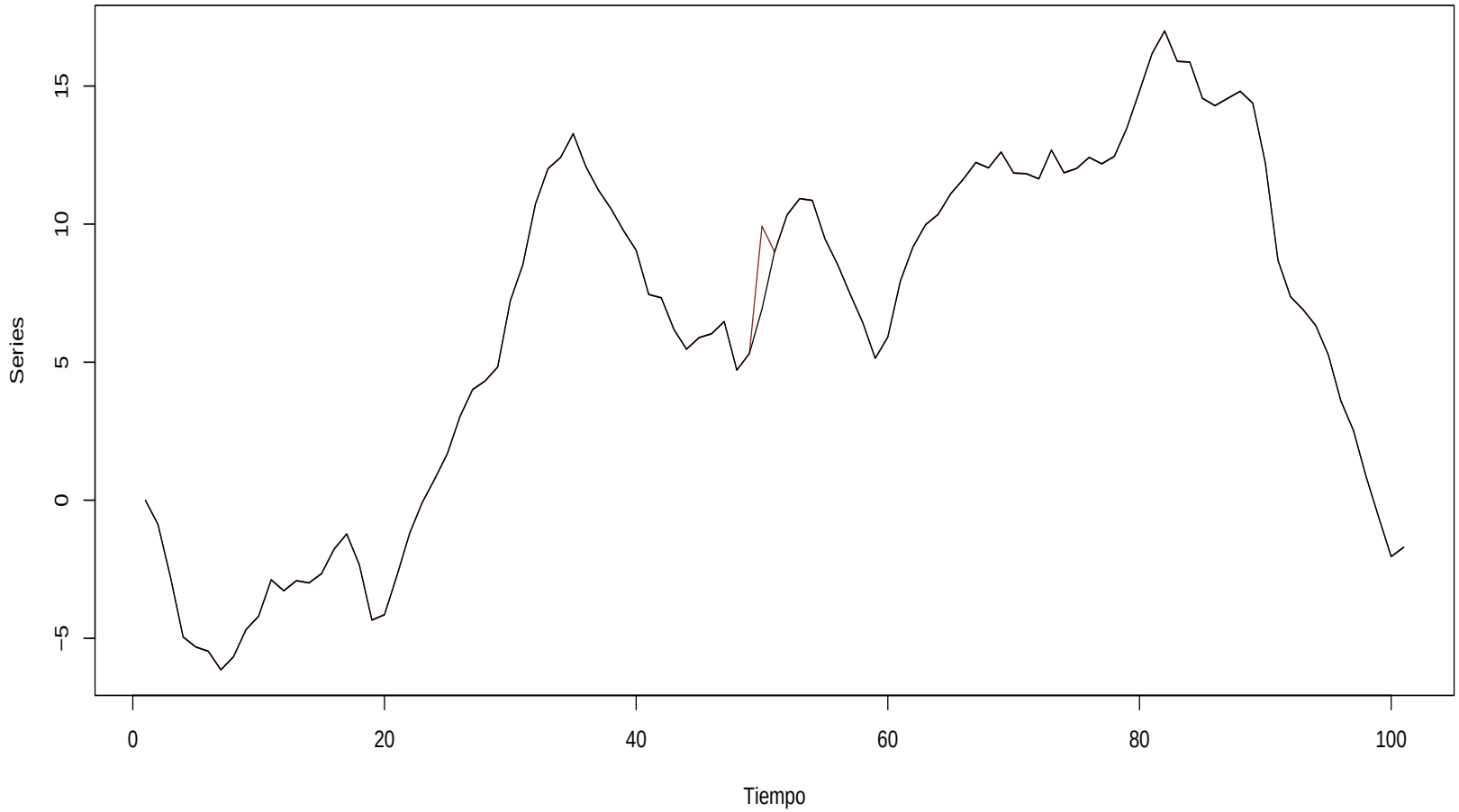
- Notar que un AO modifica seriamente las innovaciones del modelo como sigue:

$$e_t = a_t + \pi(B)wI_t^{(h)}$$

Atípico aditivo en un modelo AR(1) estacionario



Atípico aditivo en un modelo ARIMA(1,1,0) no estacionario



Atípico innovativo

- Un **atípico innovativo** (**IO**) es un atípico que afecta exclusivamente a una innovación.

- Una serie afectada por un IO en $t = h$ se define como sigue:

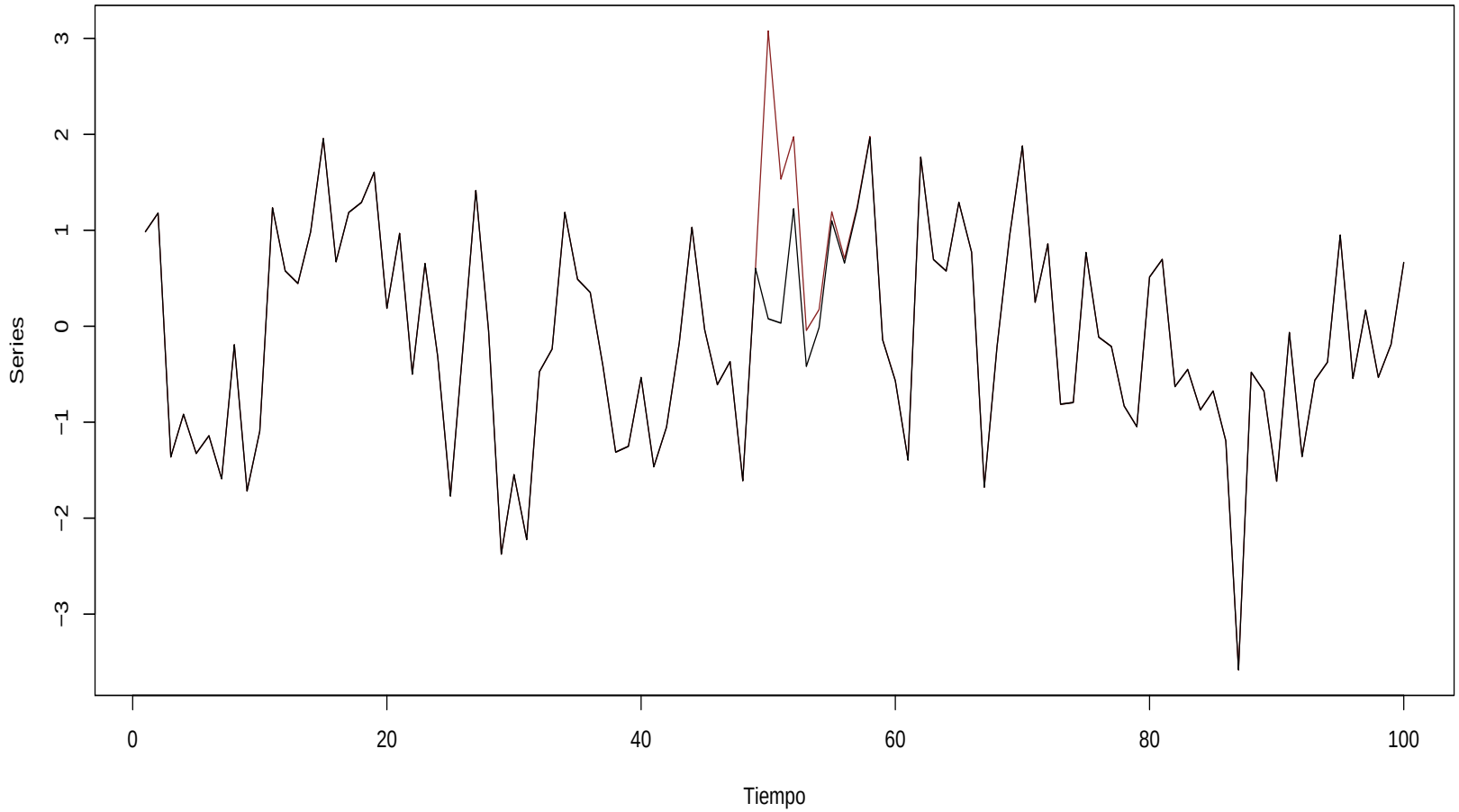
$$y_t = x_t + \psi(B) w I_t^{(h)}$$

  donde x_t es una serie temporal que sigue un modelo ARIMA(p, d, q) y $\psi(B)$ es la representación de media móvil del modelo.

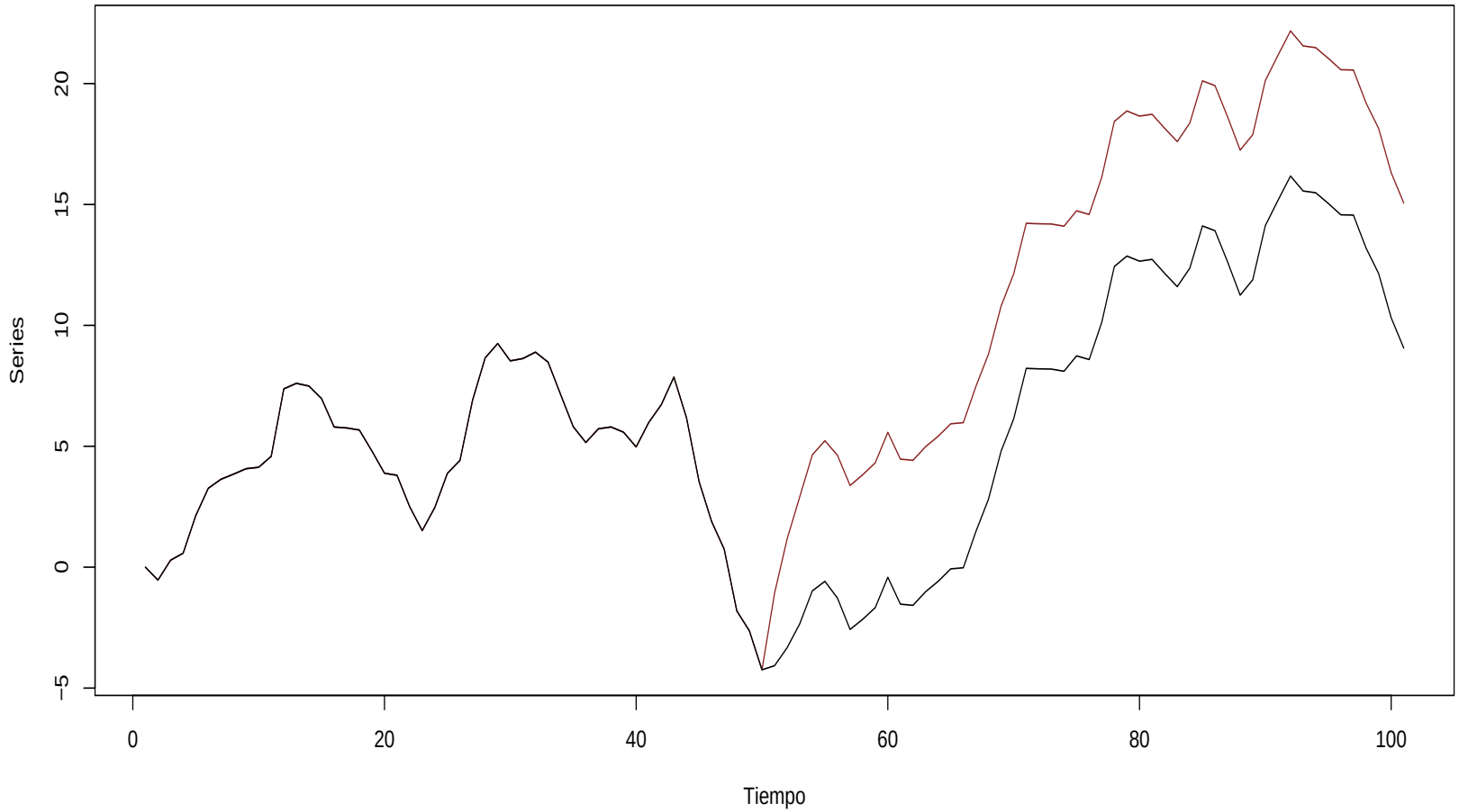
- Notar que un IO modifica únicamente una innovación del modelo:

$$e_t = a_t + w I_t^{(h)}$$

Atípico innovativo en un modelo AR(1) estacionario



Atípico innovativo en un modelo ARIMA(1,1,0) no estacionario



Cambio de nivel

- Los cambios de nivel y los cambios transitorios fueron introducidos por Tsay (1988).
- Un **cambio de nivel** (**LS**) es un cambio en el nivel de la serie. Por lo tanto, afecta a todas las observaciones a partir del instante en el que ocurre.
- Una serie afectada por un LS en $t = h$ se define como sigue:

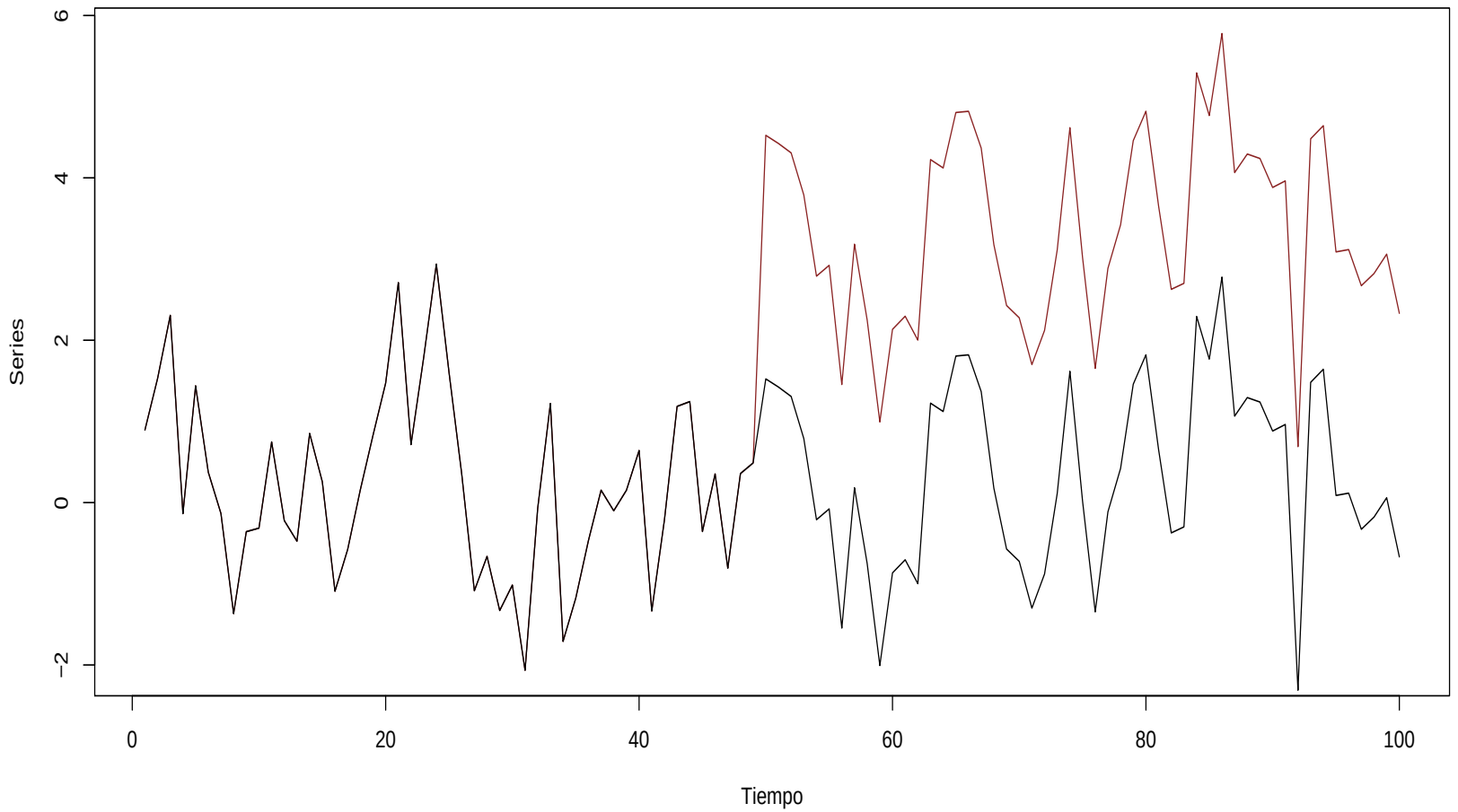
$$y_t = x_t + wS_t^{(h)}$$

  donde x_t es una serie temporal que sigue un modelo ARIMA(p, d, q) y $S_t^{(h)} = 0$ si $t < h$ y $S_t^{(h)} = 1$ si $t \geq h$.

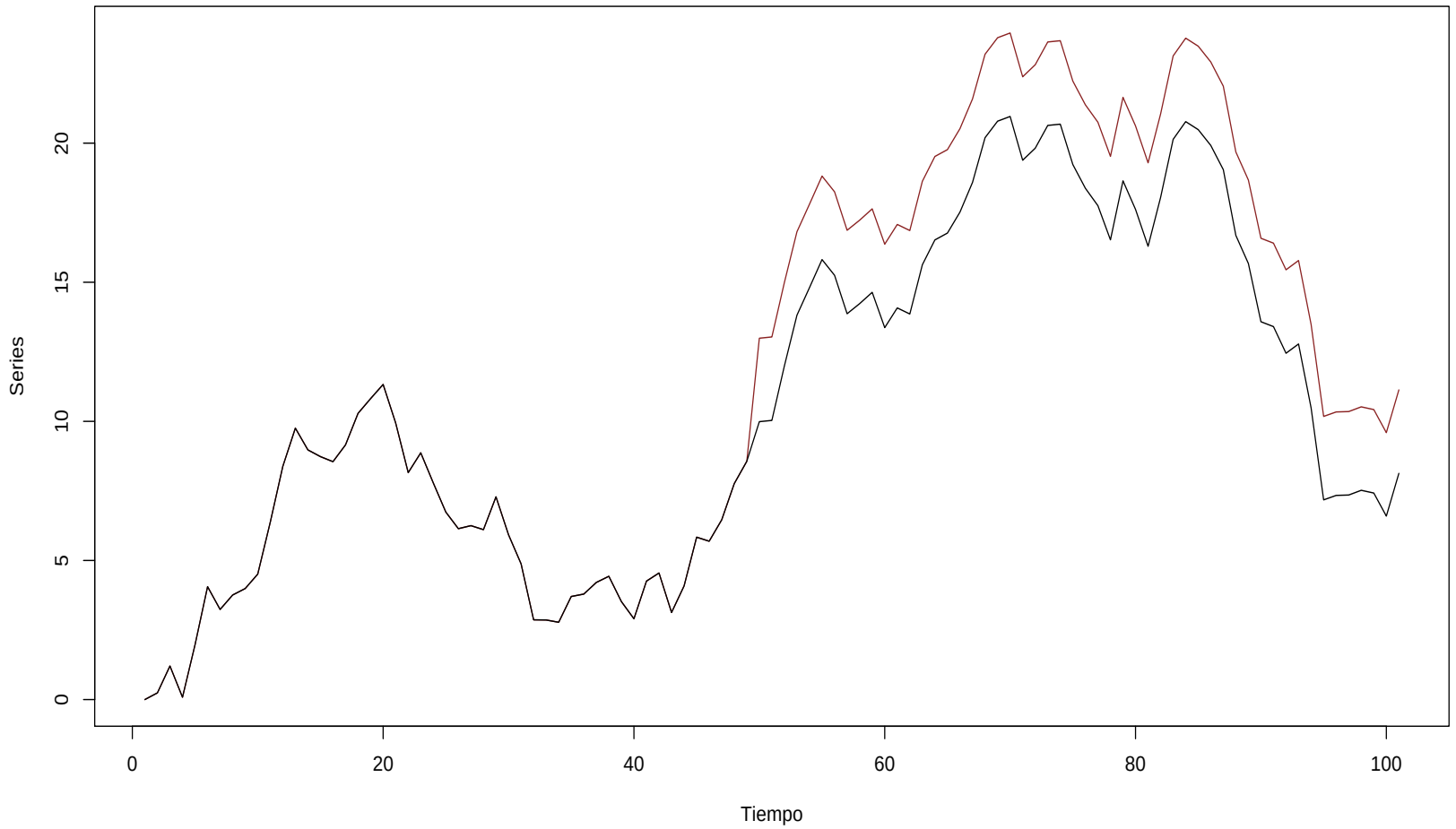
- Notar que un LS modifica todas las innovaciones del modelo desde $t = h$ como sigue:

$$e_t = a_t + \pi(B)wS_t^{(h)}$$

Cambio de nivel en un modelo AR(1) estacionario



Cambio de nivel en un modelo ARIMA(1,1,0) no estacionario



Cambio transitorio

- Un **cambio transitorio** (**TC**) es un cambio que decrece exponencialmente. Por lo tanto, afecta a todas las observaciones a partir del instante en el que ocurre, si bien el efecto se diluye.

- Una serie afectada por un TC en $t = h$ se define como sigue:

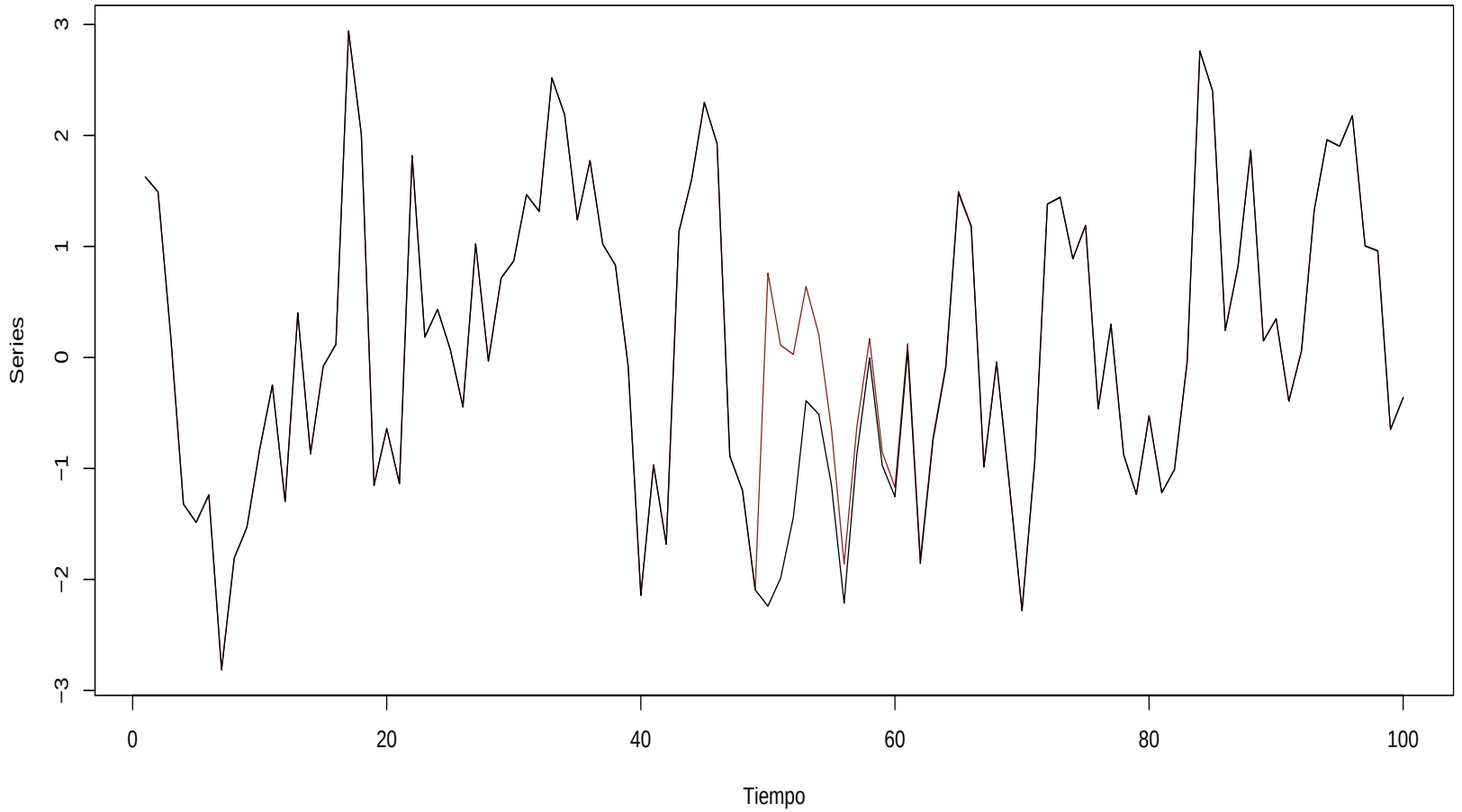
$$y_t = x_t + \frac{w}{1 - \delta B} I_t^{(h)}$$

  donde x_t es una serie temporal que sigue un modelo ARIMA(p, d, q) y $0 < \delta < 1$.

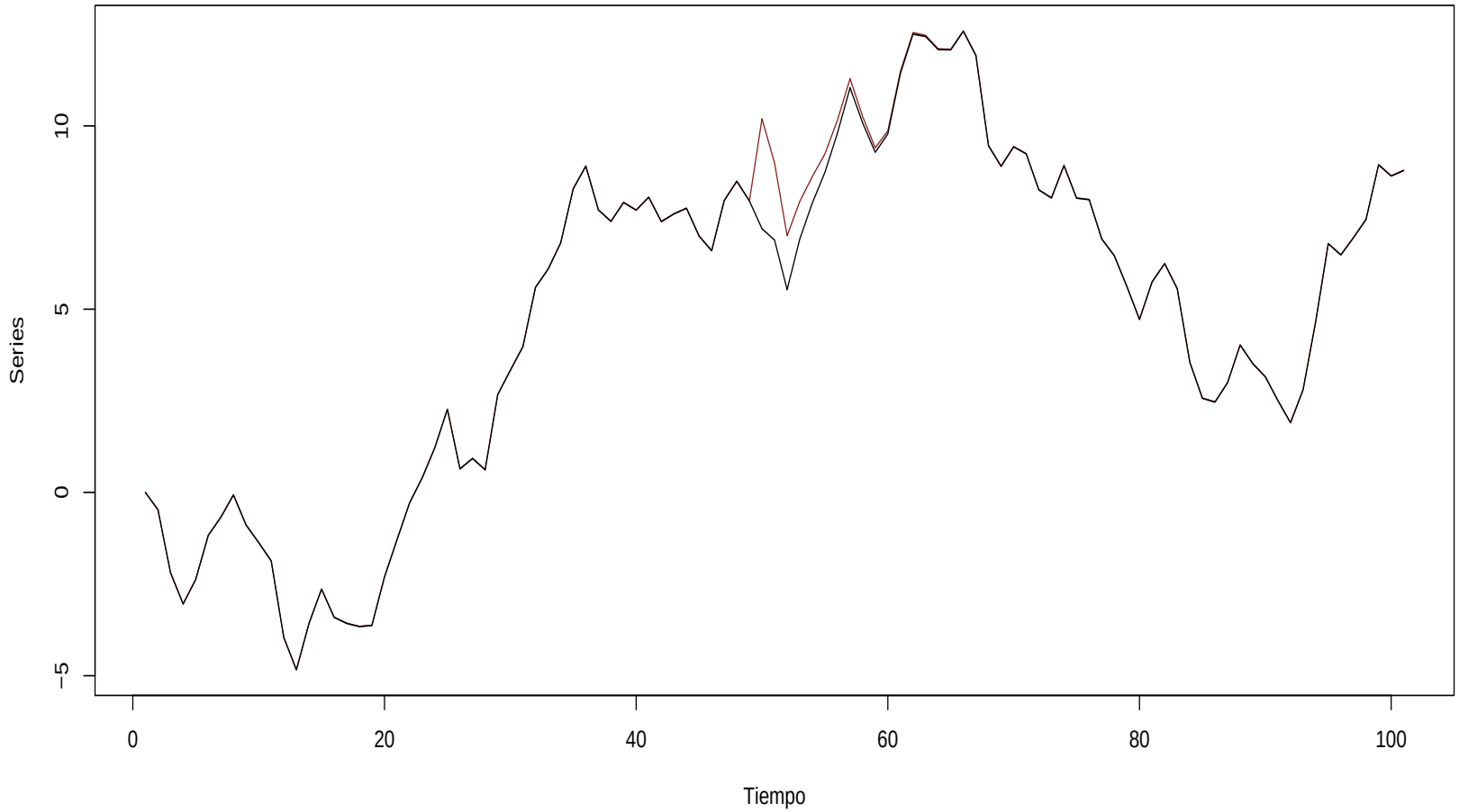
- Notar que un TC modifica todas las innovaciones del modelo desde $t = h$ como sigue:

$$e_t = a_t + \pi(B) \frac{w}{1 - \delta B} I_t^{(h)}$$

Cambio transitorio en un modelo AR(1) estacionario



Cambio transitorio en un modelo ARIMA(1,1,0) no estacionario



Cambio de tendencia en modelos ARIMA$(p, 1, q)$

- Un **cambio de tendencia** (**RS**) es un cambio en la tendencia lineal de la serie. Por lo tanto, afecta a todas las observaciones a partir del instante en el que ocurre.

- Una serie afectada por una racha de RS en $t = h$ se define como sigue:

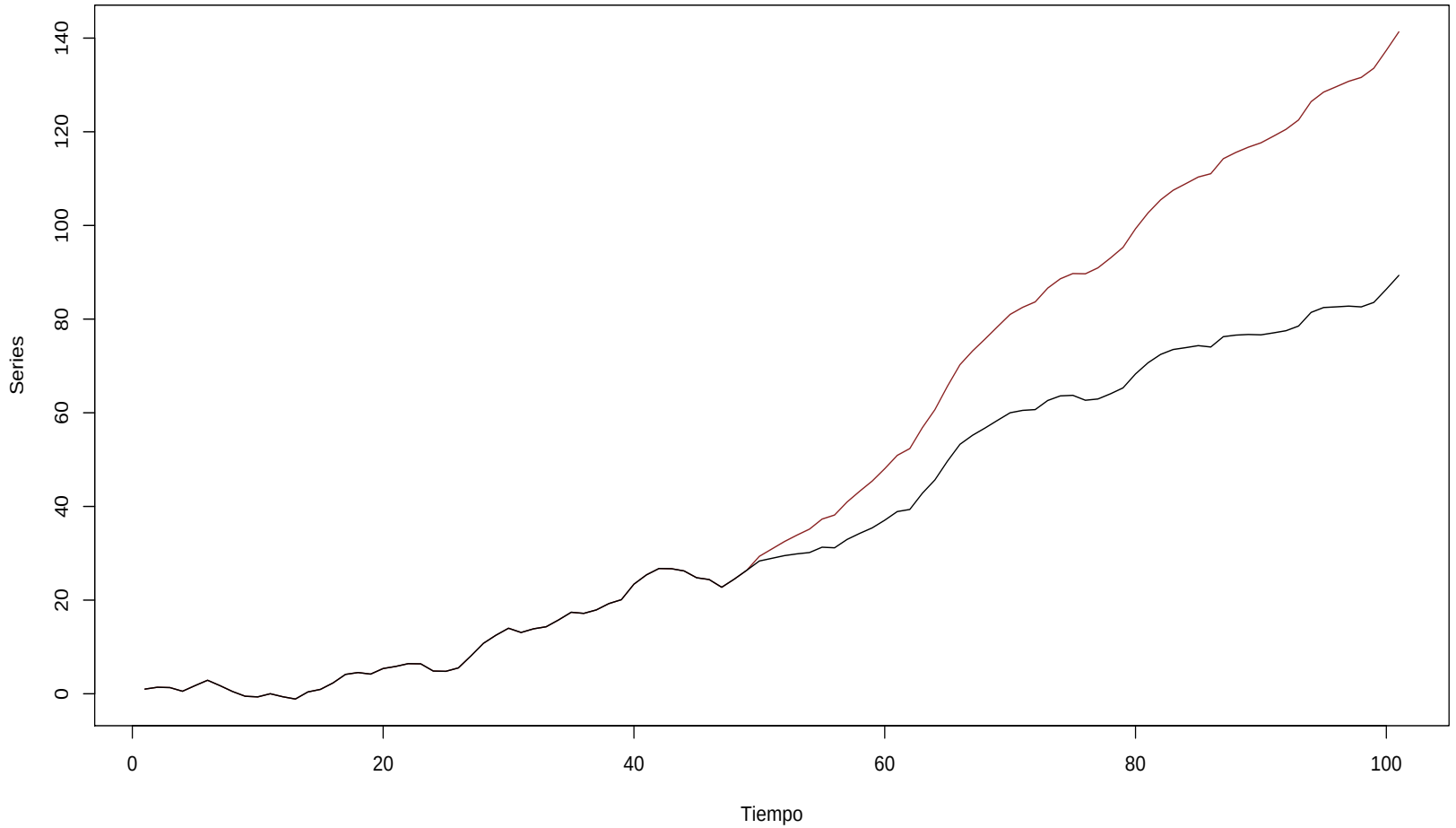
$$y_t = x_t + wR_t^{(h)}$$

  done x_t es una serie temporal que sigue un modelo ARIMA$(p, 1, q)$ y $R_t^{(h)} = 0$ si $t < h$ y $R_t^{(h)} = t - h + 1$ si $t \geq h$.

- Notar que un RS modifica todas las innovaciones del modelo desde $t = h$ como sigue:

$$e_t = a_t + \pi(B)wR_t^{(h)}$$

Cambio de tendencia en un modelo ARIMA(1,1,0) no estacionario



Efectos de la presencia de datos atípicos

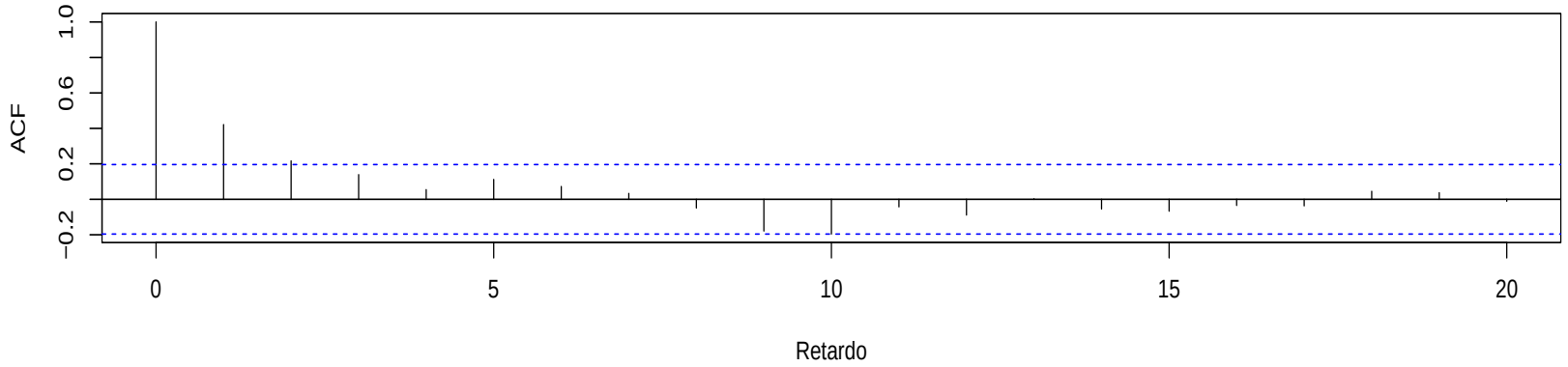
- Los atípicos tienen diferentes efectos sobre la identificación, la estimación y la predicción.

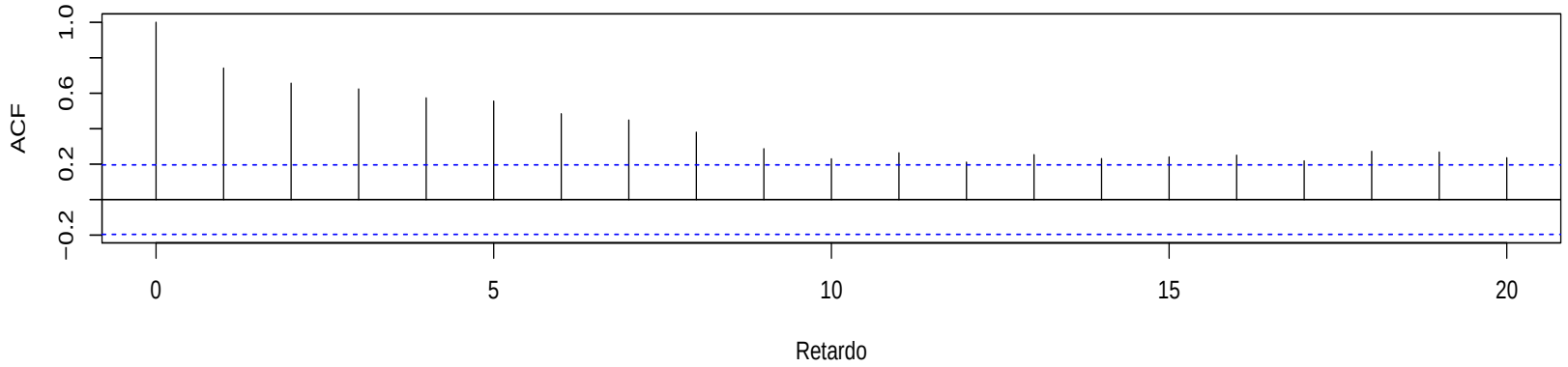
- Se pueden derivar estos efectos analiticamente en casos particulares. Mejor veamos un ejemplo simulado.

- Generamos una serie de tamaño 105 de un modelo AR(1) de parámetros $\phi = 0.5$ y $\sigma_a^2 = 1$, a la que denominados x_t. A continuación, definimos una serie y_t que es similar a la serie x_t pero le incluimos un cambio de nivel de tamaño $w = 3$ en $t = 50$.

- Obtenemos las funciones de autocorrelación (ACF) para las primeras 100 observaciones de las dos series, que se pueden ver en la siguiente transparencia. Como se puede comprobar, las diferencias son notables.

Funciones de autocorrelación (ACF)





Efectos de la presencia de datos atípicos

- A continuación, estimamos un modelo AR(1) para los 100 primeros valores de la serie y_t ignorando la presencia del cambio de nivel. Las estimaciones de ϕ y de σ_a^2 son 0.9011 y 1.344, respectivamente.

- Entonces, estimamos un modelo AR(1) para los 100 primeros valores de la serie y_t teniendo en cuenta la presencia del cambio de nivel. La estimacienes de ϕ y de σ_a^2 son 0.5254 y 0.9604, respectivamente.

- Las predicciones de las siguientes cinco observaciones con ambos modelos comparadas con las observaciones reales son:

	y_{101}	y_{102}	y_{103}	y_{104}	y_{105}
Verdaderos	5.13	4.24	4.96	4.45	3.52
Predichos con el primer modelo	3.26	2.94	2.65	2.38	2.15
Predichos con el segundo modelo	3.51	3.46	3.43	3.41	3.40

Procedimientos para la detección de atípicos en modelos ARIMA(p, d, q)

- Se han propuesto una gran cantidad de procedimientos para la detección de datos atípicos y para la estimación conjunta de los parámetros del modelo y los tamaños de los atípicos.

- Estos procedimientos son algoritmos iterativos basados o bien en múltiples contrastes de hipótesis o bien en procedimientos bayesianos. Los más populares son los siguientes:

  1. Métodos basados en **estadísticos de razón de verosimilitudes**: Fox (1972), Tsay (1986 y 1988), Chang, Tiao y Chen (1988), Chen y Liu (1993) y Sánchez y Peña (2003), entre otros.
  2. Métodos basados en **medidas de influencia**: Bruce y Martin (1989), Peña (1990), Ledolter (1990), Lefrançois (1991) y Ljung (1993), entre otros.
  3. Métodos **Bayesianos**: Abraham y Box (1979), McCulloch y Tsay (1993), McCulloch y Tsay (1994) y Justel, Peña y Tsay (2001), entre otros.

Estadísticos de razón de verosimilitudes

- Los métodos que han tenido más exito son los básados en estadísticos de razón de verosimilitudes (**LRT**).

- Notar que las innovaciones, bajo cualquier tipo de atípico en $t = h$, se puede escribir como sigue:

$$e_t = a_t + \varphi(B)\, w I_t^{(h)}$$

  donde $\varphi(B)$ es un polinomio en el operador retardo B, dado por:

  - Para un AO, $\varphi^{AO}(B) = \pi(B)$
  - Para un IO, $\varphi^{IO}(B) = 1$
  - Para un LS, $\varphi^{LS}(B) = (1 - B)^{-1} \pi(B)$
  - Para un TC, $\varphi^{TC}(B) = (1 - \delta B)^{-1} \pi(B)$

- Por lo tanto, si conocemos $\pi(B)$, el tipo y el momento de aparición $t = h$, podemos estimar el tamaño del atípico, ω por máxima verosimilitud (mínimos cuadrados), utilizando el modelo de regresión anterior y contrastar su efecto mediante el LRT.

Estadísticos de razón de verosimilitudes

- Los estadísticos de razón de verosimilitudes están dados por:

$$\lambda_{AO,h} = \frac{\widehat{w}_{AO}}{\rho_{AO}\sigma_a} \qquad \lambda_{IO,h} = \frac{\widehat{w}_{IO}}{\sigma_a} \qquad \lambda_{LS,h} = \frac{\widehat{w}_{LS}}{\rho_{LS}\sigma_a} \qquad \lambda_{TC,h} = \frac{\widehat{w}_{TC}}{\rho_{TC}\sigma_a},$$

  respectivamente, donde $\widehat{w}_{AO}$, $\widehat{w}_{IO}$, $\widehat{w}_{LS}$ y $\widehat{w}_{TC}$ son las estimaciones de los tamaños de los atípicos y ρ_{AO}, ρ_{IO}, ρ_{LS} y ρ_{TC}, dependen de los parámetros del modelo.

- Es importante notar que estamos suponiendo que conocemos los parámetros del modelo, el tipo de atípico y el momento de aparición del atípico.

- En la práctica, no conocemos ni los parámetros del modelo, ni el número de atípicos, ni sus tipos y ni, por supuesto, los momentos de aparición.

Procedimiento estándar (Chen y Liu, 1993)

1. Identificar y estimar un modelo ARIMA para la serie observada.

2. Obtener los estadísticos **LRT** para cada punto y tipo, $\lambda_{AO,t}$, $\lambda_{IO,t}$, $\lambda_{LS,t}$ y $\lambda_{TC,t}$, $t = 1, \ldots, n$.

3. Obtener:

$$(i_{max}, h_{max}) = argmax \left\{ |\lambda_{AO,t}|, \ |\lambda_{IO,t}|, \ |\lambda_{LS,t}|, \ |\lambda_{TC,t}| \right\},$$

   donde (i_{max}, h_{max}) es la estimación del tipo y de la localización del atípico.

4. Si $|\lambda_{i_{max},h_{max}}|$ es significativo, suponemos un atípico de tipo i_{max} en $t = h_{max}$. Eliminar su efecto mediante $\widehat{w}_{i_{max},h_{max}}$, reestimar el modelo y repetir los pasos anteriores hasta que no se detectan más atípicos.

5. Ajustar un modelo conjunto para los parámetros y los atípicos. Si alguno no es significativo, se elimina y se vuelve a estimar el modelo.

Algunos aspectos importantes del procedimiento

- En el primer paso hay que identificar y estimar el modelo:
  1. Utilizando identificación y estimación robusta.
  2. Aplicar el algoritmo para modelos ARIMA con diferentes ordenes y seleccionar el más adecuado mediante algún criterio de selección, como el criterio de Akaike (**AIC**) o criterio Bayesiano (**BIC**).

- Los valores críticos para aplicar los contrastes se obtienen por simulación.

- Diferentes versiones de este algoritmo están implementados en:
  1. **TRAMO** del Banco de España desarrollado por Agustín Maravall.
  2. **X-12-ARIMA** del U.S. Bureau of Census desarrollado por David F. Findley y William R. Bell.
  3. **SCA** de Scientific computing associates Corp. desarrollado por George C. Tiao y Lon-Mu Liu.

Problemas importantes de procedimientos iterativos

- Notar que bajo la presencia de varios atípicos calculamos los estadísticos LRT y estimamos efectos influenciados por su presencia. Esto produce dos efectos adversos:

  1. Efecto de **enmascaramiento**: La presencia de algunos atípicos enmascara la presencia de otros.
  2. Efecto de **propagación**: La presencia de algunos atípicos produce que observaciones no atípicas parezcan serlo.

- El algoritmo tiende a confundir **IO**'s y **LS**'s: bajo la presencia de un **LS**, el estadístico para un **IO** suele ser mayor que el estadístico para un **LS**.

- No se buscan cambios en la tendencia en series no estacionarias.

Procedimiento de detección de Galeano y Peña

- Sería mucho mejor si en lugar de utilizar un algoritmo iterativo detectaramos todos los atípicos en el mismo paso. Galeano y Peña (2012) proponen un procedimiento para el caso de atípicos aditivos basado en esta idea para modelos ARIMA.

- Supongamos que la serie observada $y_1, \ldots, y_n$ contiene m atípicos aditivos. Entonces:

$$y_t = x_t + w_{t_1} I_t^{(t_1)} + \cdots + w_{t_m} I_t^{(t_m)}$$

  donde:

  - x_t sigue un modelo ARIMA(p, d, q);
  - $\tau_m = (t_1, \ldots, t_m)$ es el vector que contiene las localizaciones de los atípicos;
  - $w_{t_1}, \ldots, w_{t_m}$ son los tamaños de los atípicos.

Procedimiento de detección de Galeano y Peña

- Consecuentemente, la serie y_t sigue un modelo de regresión con errores ARIMA dado por:

$$\phi_p(B)(1-B)^d\left(y_t - w_{t_1}I_t^{(t_1)} - \cdots - w_{t_m}I_t^{(t_m)}\right) = \theta_q(B)\,a_t,$$

  en la que los regresores son variables indicadoras y los parámetros asociados con los regresores son los tamaños de los atípicos.

- Denotamos este modelo por M_{τ_m}. Los p_m parámetros de este modelo son:

$$\rho_{\tau_m} = \left(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, w_{t_1}, \ldots, w_{t_m}, \sigma^2\right)$$

  where $p_m = p + q + m + 1$.

Procedimiento de detección de Galeano y Peña

- Entonces, dada y, el número y la localización de los atípicos aditivos, m y τ_m, y el vector de parámetros ρ_{τ_m} son desconocidos y deben ser estimados.

- Determinar el número y la localización de los atípicos en y es equivalente a seleccionar el modelo M_{τ_m} con los verdaderos atípicos de entre todo el conjunto de modelos candidatos.

- Notar que los modelos candidatos incluyen el modelo sin atípicos, M_{τ_0}, los n modelos con un atípico, M_{τ_1}, etc. . . En total, existen $\binom{n}{m}$ modelos candidatos con m atípicos que cubren todos las posibles localizaciones de los atípicos.

- Suponiendo que el número de atípicos tiene una cota superior, $m_{\max} < n$, el número total de modelos candidatos está dado por:

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{m_{\max}}$$

Procedimiento de detección de Galeano y Peña

- Para calcular la probabilidad de que la serie y pertenezca al modelo M_{τ_m} podemos utilizar el Teorema de Bayes:

$$p\left(M_{\tau_m}|y\right) = \frac{p\left(M_{\tau_m}\right)\mathcal{L}\left(M_{\tau_m}|y\right)}{f\left(y\right)},$$

siendo:

1. $p\left(M_{\tau_m}\right)$ es la probabilidad a priori del modelo M_{τ_m}.
2. $\mathcal{L}\left(M_{\tau_m}|y\right)$ es la verosimilitud marginal para el modelo M_{τ_m}.
3. $f\left(y\right)$ es la verosimilitud incondicional de y dada por,

$$f\left(y\right) = \sum_{j=0}^{m_{\max}} \sum_{\tau_j} p\left(M_{\tau_j}\right)\mathcal{L}\left(M_{\tau_j}|y\right).$$

- El problema es que estas cantidades dependen de los parámetros del modelo, que son desconocidos.

Procedimiento de detección de Galeano y Peña

- Utilizando una expansión de segundo orden de la verosimilitud $\mathcal{L}(M_{\tau_m}|y)$, es posible demostrar que:

$$\log p(M_{\tau_m}|y) \simeq \ell_{\tau_m}(\widehat{\rho}_{\tau_m}) - \frac{p_m}{2}\log(n) + \log p(M_{\tau_m}) + C$$

donde $\ell_{\tau_m}(\widehat{\rho}_{\tau_m})$ es el logaritmo de la verosimilitud en $\widehat{\rho}_{\tau_m}$ y C es constante.

- Eliminando la constante, se define un criterio de selección de modelos dado por:

$$BICUP(M_{\tau_m}) = -2\ell_{\tau_m}(\widehat{\rho}_{\tau_m}) + p_m \log n + \log p(M_{\tau_m})$$

donde:

$$p(M_{\tau_m}) = \frac{1}{1+m_{\max}}\frac{1}{\binom{n}{m}}$$

es decir, a priori tomamos una distribución uniforme sobre el número de atípicos.

Procedimiento de detección de Galeano y Peña

- Sin embargo, calcular el criterio para todos los modelos candidatos es prácticamente imposible incluso para valores pequeños de n y de m_{max}.

- La idea es dividir las observaciones de la serie temporal en dos grupos: en el primero se incluyen que tienen un alto potencial de ser atípicas, y en el segundo se incluyen las observaciones que se puede descartar que lo sean.

- De esta manera, si n_1 es el número de observaciones en el primer grupo, entonces, en lugar de calcular el valor de criterio para todos los modelos candidatos, se incluyen todos los modelos que consideran como atípicas cualquier subconjunto de las n_1 observaciones del primer grupo. El número de modelos candidatos se reduce enormemente.

- Para realizar la división se han propuesto un procedimiento basado en la comparación del valor del criterio para grupos de una y dos observaciones (ver artículo).

Procedimiento de detección de Galeano y Peña

- Se genera una serie temporal de tamaño $n = 100$ que sigue un modelo estacional ARIMA$(0,1,1) \times (0,1,1)_{12}$:

$$\left(1 - B^{12}\right)(1 - B)\, x_t = \left(1 + 0.5B^{12}\right)(1 + 0.4B)\, a_t,$$

  donde a_t sigue una distribución Gaussiana de media 0 y desviación típica $\sigma = 0.7071$.

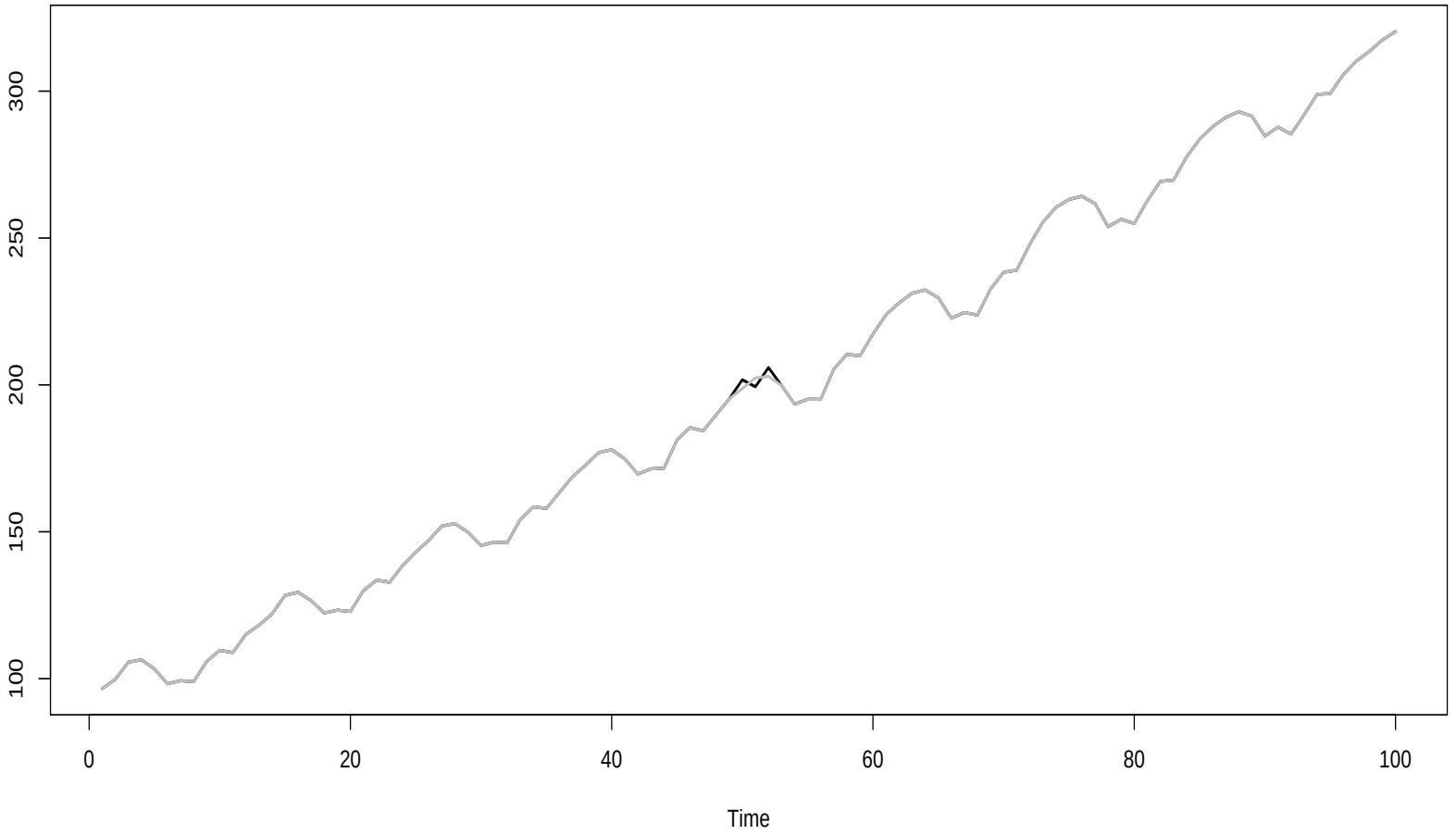
- Esta serie se contamina con tres atípicos aditivos en $t_1 = 50$, $t_2 = 51$ y $t_3 = 52$ de tamaños $w_{50} = 3$, $w_{51} = -3$ and $w_{52} = 3$, respectivamente. Entonces, la serie contaminada es:

$$y_t = x_t + 3I_t^{(50)} - 3I_t^{(51)} + 3I_t^{(52)},$$

  para $t = 1, \ldots, T$.

- Las dos series se muestran en la figura de la siguiente transparencia.

Procedimiento de detección de Galeano y Peña



Procedimiento de detección de Galeano y Peña

- Comparación de los parámetros estimados con el modelo ARIMA$(0,1,1)\times(0,1,1)_{12}$ sin y con los atípicos incluidos en la serie.

	Parámetros estimados	
Verdaderos parámetros	Modelo sin atípicos	Modelo son atípicos
$\theta_1 = -0.4$	0.378 (0.092)	-0.456 (0.125)
$\Theta_1 = -0.5$	0.220 (0.135)	-0.408 (0.153)
$w_{50} = 3$	$-$	2.814 (0.58)
$w_{51} = -3$	$-$	-2.973 (0.785)
$w_{52} = 3$	$-$	2.84 (0.568)
$\sigma = 0.707$	1.373	0.722

Procedimiento de detección de Galeano y Peña

- Aplicamos el procedimiento propuesto. En primer lugar, buscamos datos con alta probabilidad de ser atípicos. Encontramos las tres observaciones incluidas como atípicas.

- A continuación, calculamos el valor del BICUP para los 8 modelos candidatos. El criterio selecciona el modelo verdadero.

τ_m	$(-)$	(50)	(51)	(52)	$(50,51)$	$(50,52)$	$(51,52)$	$(50,51,52)$
BICUP	321.3	316.6	267.6	315.9	268.5	248.5	269	**246.8**

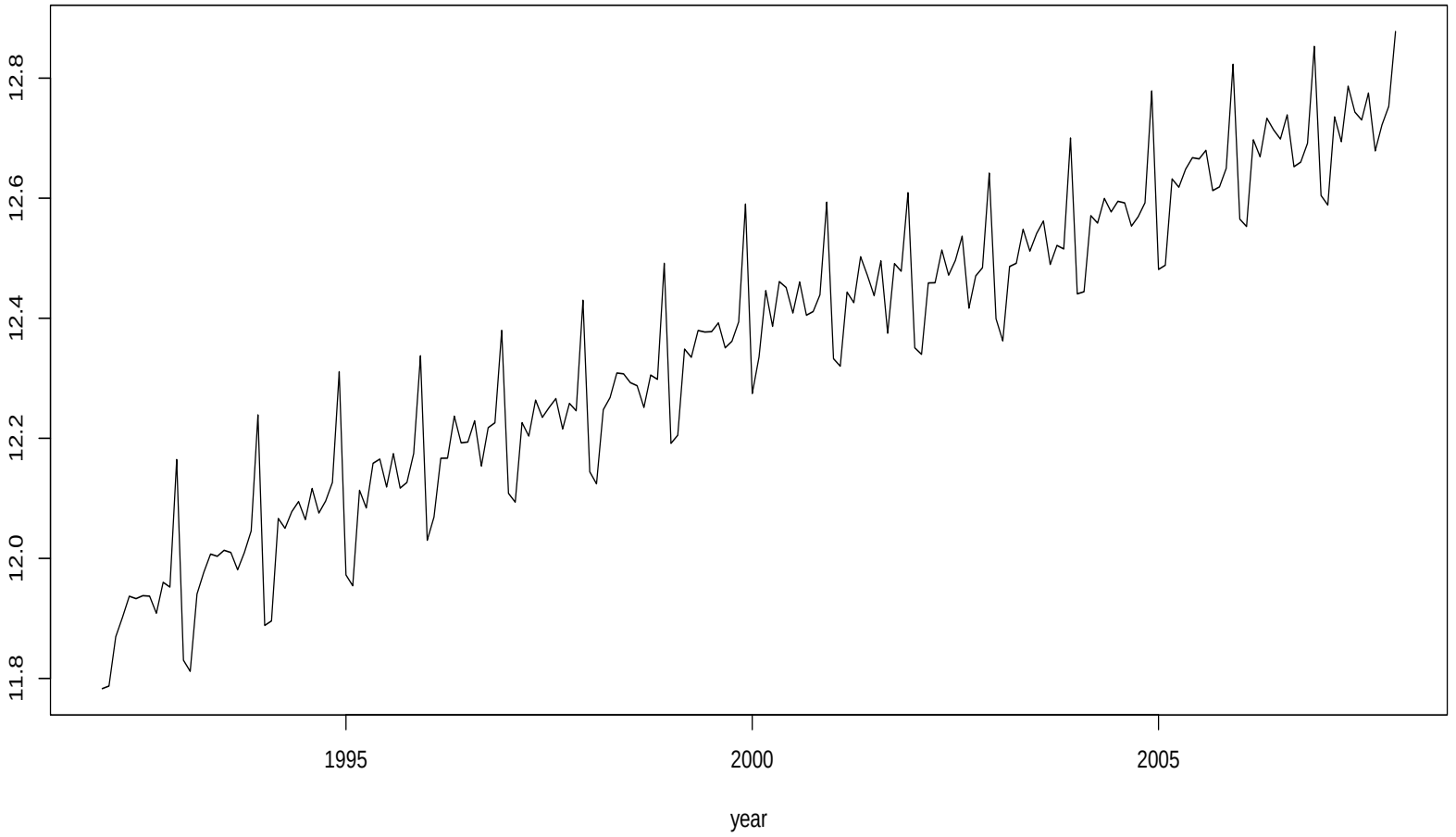
Procedimiento de detección de Galeano y Peña

- A continuación, aplicamos el procedimiento propuesto al logaritmo de la serie mensual de ventas minoristas en Estados Unidos desde Enero de 1992 a Diciembre de 2007. La serie se muestra en el siguiente transparencia y muestra un claro comportamiento estacional.

- Para tener en cuenta el efecto de los días laborables se incluyen siete regresores. Las primeras seis son r_{1t}=$(no.\ de\ Lunes)-(no.\ de\ Domingos)$ en mes t,. . . ,r_{6t}=$(no.\ de\ Sábados)\ -(no.\ de\ Domingos)$ en mes t, mientras que la septima variable es r_{7t}=$días\ en\ el\ mes\ t$.

- Entonces, se ajusta a la serie un modelo estacional ARIMA$(0,1,1)\times(0,1,1)_{12}$ junto con los siete regresores. Los residuos del modelo indican que el ajuste es adecuado.

Procedimiento de detección de Galeano y Peña



Procedimiento de detección de Galeano y Peña

- A continuación, aplicamos el procedimiento propuesto para seleccionar atípicos en la serie.

- El procedimiento para detectar potenciales atípicos selecciona las observaciones en Octubre de 2001 y Mayo de 2005.

- A continuación, calculamos el valor del BICUP para los 4 modelos candidatos. El criterio selecciona el modelo con las dos observaciones como atípicas.

τ_m	$(-)$	(118)	(161)	$(118, 161)$
BICUP	-1015.2	-1026.2	-1016.7	$\mathbf{-1034.5}$

Atípicos en otros modelos univariantes

- Modelos **ARIMA** con algoritmos genéticos: Baragona, Battaglia y Poli (2011).
- Modelos **AR** con errores Poisson: Fokianos y Fried (2011).
- Modelos **TAR**: Battaglia y Orfei (2005).
- Modelos **GARCH**: Doornik y Ooms (2005) y Hotta y Tsay (2012).
- Modelos **INGARCH**: Fokianos y Fried (2010).
- Procedimientos robustos para series temporales ARIMA se pueden encontrar en el libro de Maronna, Martin y Yohai (2006), en Muler, Peña y Yohai (2010).

Contenidos

1. Introducción

2. Atípicos en series temporales univariantes

   (a) Tipos de atípicos
   (b) Procedimientos habituales de detección basados en Chen y Liu (1993)
   (c) Procedimiento de detección de Galeano y Peña (2012)

3. **Atípicos en series temporales multivariantes**

   (a) Tipos de atípicos
   (b) Procedimiento de detección de Tsay, Peña y Pankratz (2000)
   (c) Procedimiento de detección de Galeano, Peña y Tsay (2006)

4. Conclusiones

Atípicos en series temporales multivariantes

- Una serie multivariante $X_t = (X_{1,t}, \ldots, X_{k,t})'$ sigue un modelo ARMA(p, q) vectorial si:

$$(1 - \Phi_1 B - \cdots - \Phi_p B^p) X_t = C + (1 - \Theta_1 B - \cdots - \Theta_q B^q) A_t, \qquad A_t \sim N_k(0, \Sigma)$$

donde Φ_i y Θ_j son matrices.

- De esta formula podemos obtener las representaciones autorregresivas y de media móvil dadas por:

$$\Pi(B) C_t = C_\Pi + A_t, \qquad X_t = C_\Psi + \Psi(B) A_t$$

donde $\Pi(B) = \Phi(B)\Theta(B)^{-1}$ y $\Psi(B) = \Theta(B)\Phi(B)^{-1}$.

Atípicos en series temporales multivariantes

- Tsay, Peña and Pankratz (2000) introducen atípicos en modelos vectoriales ARMA(p, q).

1. Atípico aditivo multivariante (**MAO**): $Y_t = X_t + W I_t^{(h)}$, donde $W = (w_1, \ldots, w_k)'$ es el tamaño del atípico.

2. Atípico innovativo multivariante (**MIO**): $Y_t = X_t + \Psi(B) W I_t^{(h)}$.

3. Cambio de nivel multivariante (**MLS**): $Y_t = X_t + W(1-B)^{-1} I_t^{(h)} = X_t + W S_t^{(h)}$

4. Cambio transitorio multivariante (**MTC**): $Y_t = X_t + W(1-\delta B)^{-1} I_t^{(h)}$

Procedimiento de Tsay, Peña y Pankratz

- Proponen el uso de los estadísticos **LRT** para contrastar la presencia de un dato atípico:

$$J_{i,h} = W_{i,h}{}'\Sigma_{i,h}^{-1}W_{i,h},$$

donde $W_{i,h}$ es la estimación de W que tiene una cierta matriz de covarianzas $\Sigma_{i,h}$.

- Proponen un algoritmo similar al propuesto por Chen y Liu (1993) para series univariantes: Obtener

$$(i_{max}, h_{max}) = argmax \left\{J_{AO,t}, J_{IO,t}, J_{LS,t}, J_{TC,t}\right\},$$

(i_{max}, h_{max}) es la estimación del tipo y de la localización del presumible atípico. Si es significativo, estimar su efecto, eliminar y repetir la búsqueda.

- Mismos problemas que en el caso univariante: Especificación del modelo, confusión entre **MIO** y **MLS** y falta de efectos no estacionarios.

Procedimiento de Galeano, Peña y Tsay

- Galeano, Peña y Tsay (2006) proponen un procedimiento basado en proyecciones del vector de series de direcciones adecuadas. La idea fundamental es proyectar en las direcciones que maximizan y minimizan el **coeficiente de curtosis** de las series proyectadas.

- Sean $v'Y_t = y_t$ y $v'X_t = x_t$, donde v es un vector. Entonces:

$$Y_t = X_t + \alpha(B)\, wI_t^{(h)} \Longrightarrow y_t = x_t + v'\alpha(B)\, wI_t^{(h)}$$

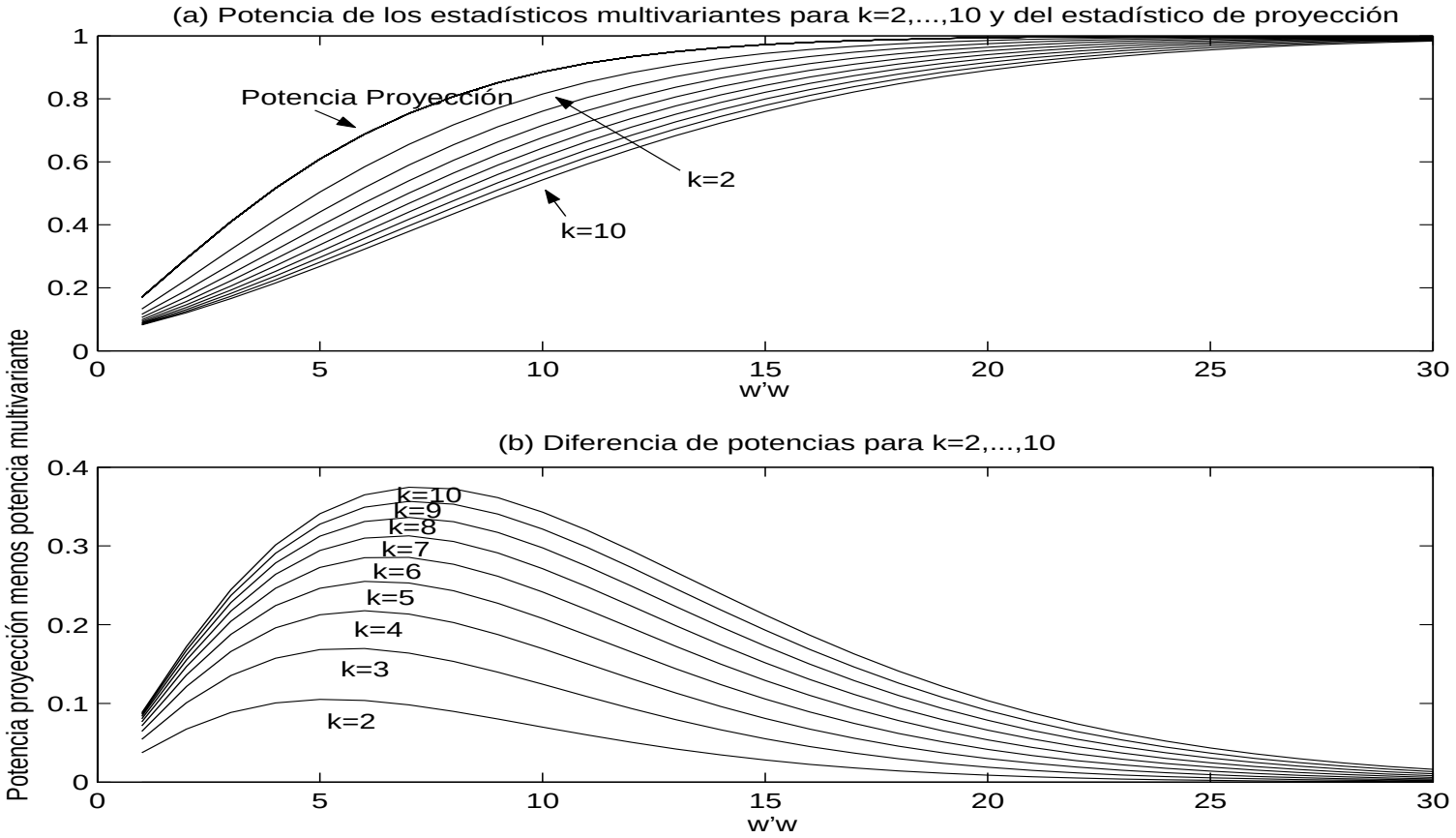
  donde el tipo de atípico se mantiene en la proyección excepto para un **MIO**, que produce una racha de atípicos.

- El procedimiento propuesto tiene las siguientes ventajas:

  1. No es necesario especificar a priori los ordenes del modelo vector ARMA.
  2. No confunde **MIO** y **MLS**.
  3. Tiene en cuenta efectos no estacionarios si la serie es no estacionaria.

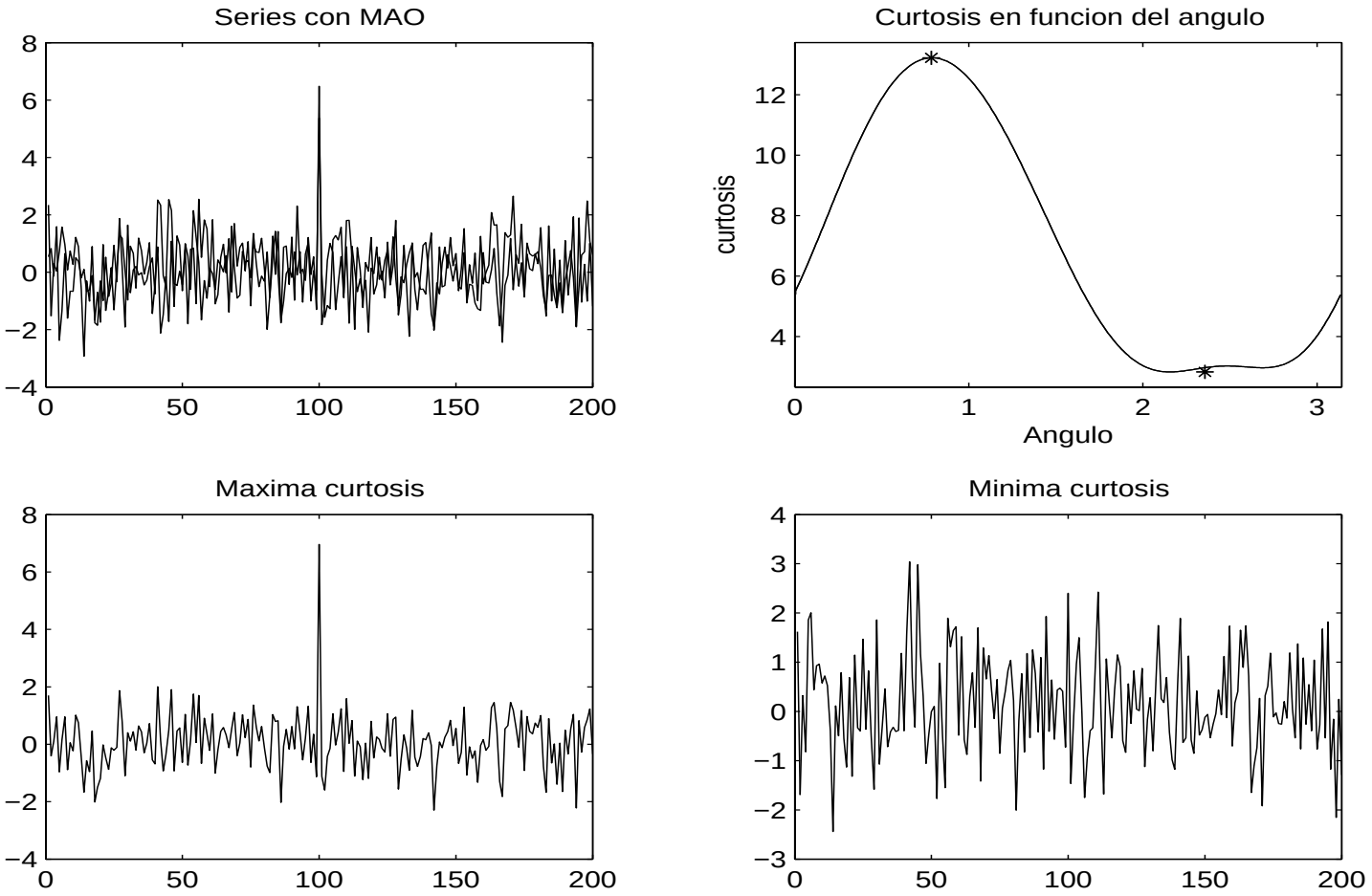
Procedimiento de Galeano, Peña y Tsay

Potencia de los estadísticos multivariantes en series de ruido blanco para $k = 2, \ldots, 10$ para un **MAO** como función de $w'w$ comparada con la potencia de los estadísticos de la proyección cuando v es la dirección que maximiza el coeficiente de curtosis.



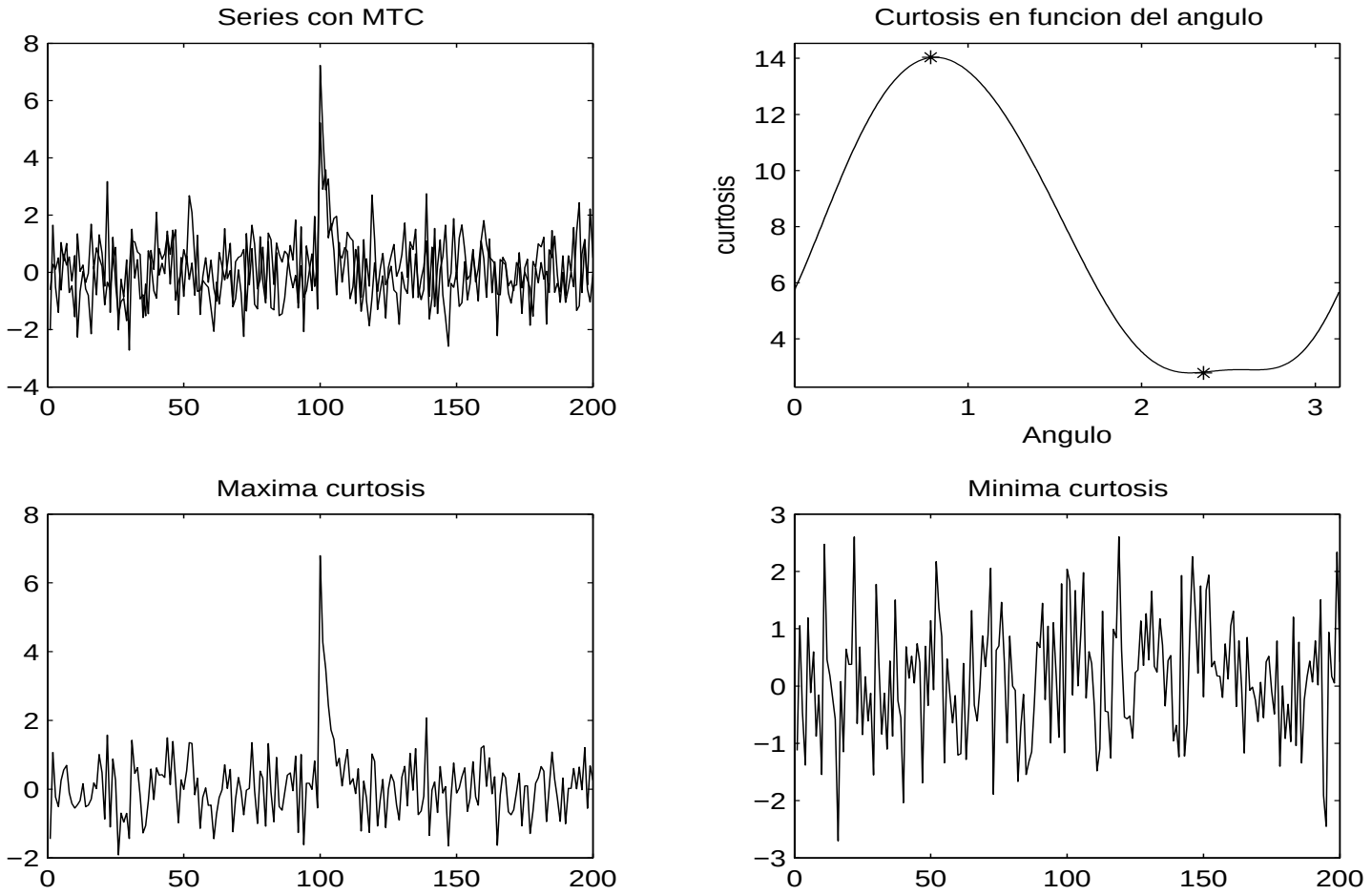
Procedimiento de Galeano, Peña y Tsay

Serie de ruido blanco con $k = 2$ y $n = 200$, con un **MAO** en $h = 100$. Proyectamos en direcciones v tales que $v = (cos(\theta), sen(\theta))$ para $\theta = 0, 0.01, \ldots, \pi$.



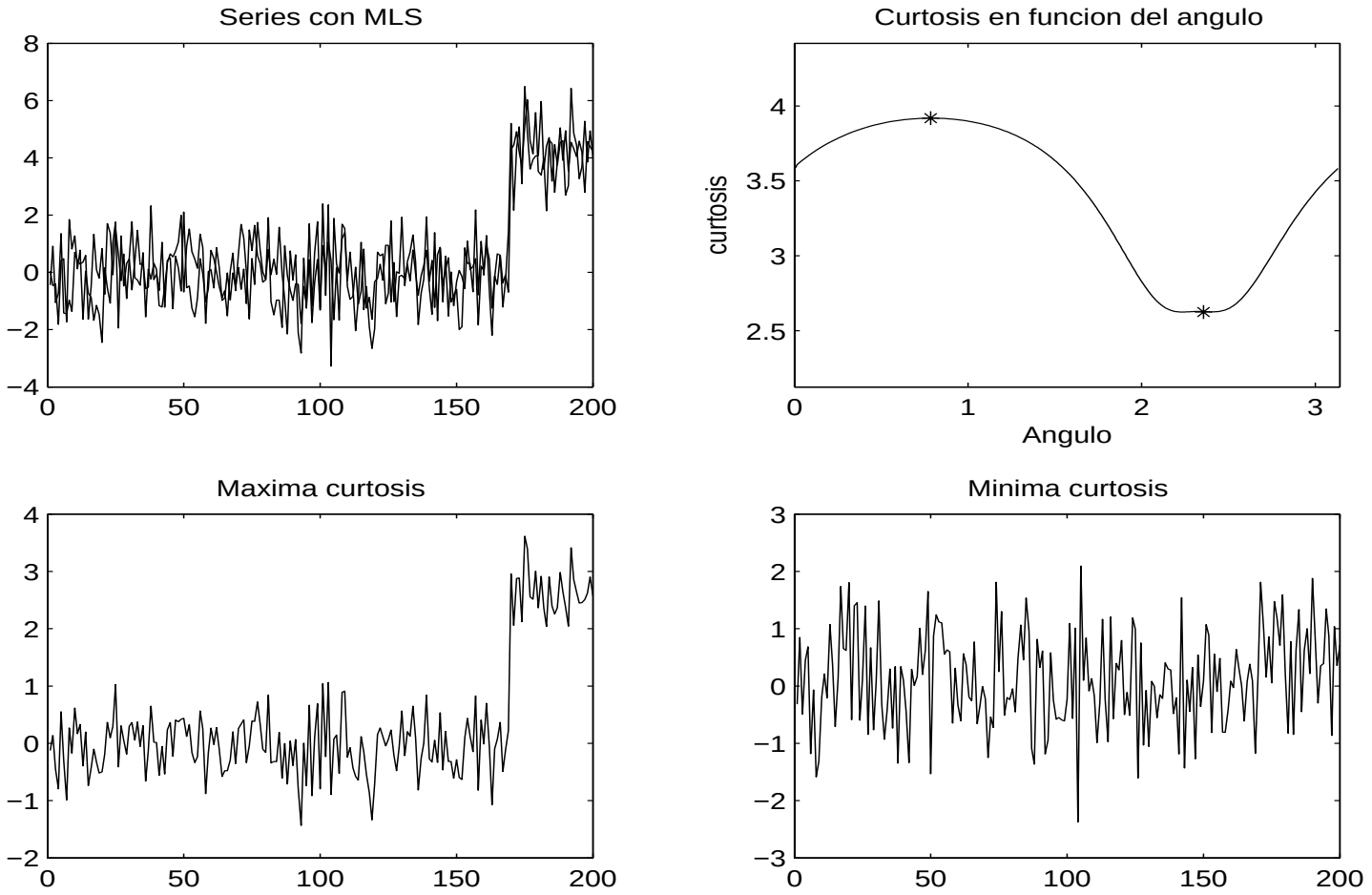
Procedimiento de Galeano, Peña y Tsay

Mismo experimento para un **MTC** en $h = 100$.



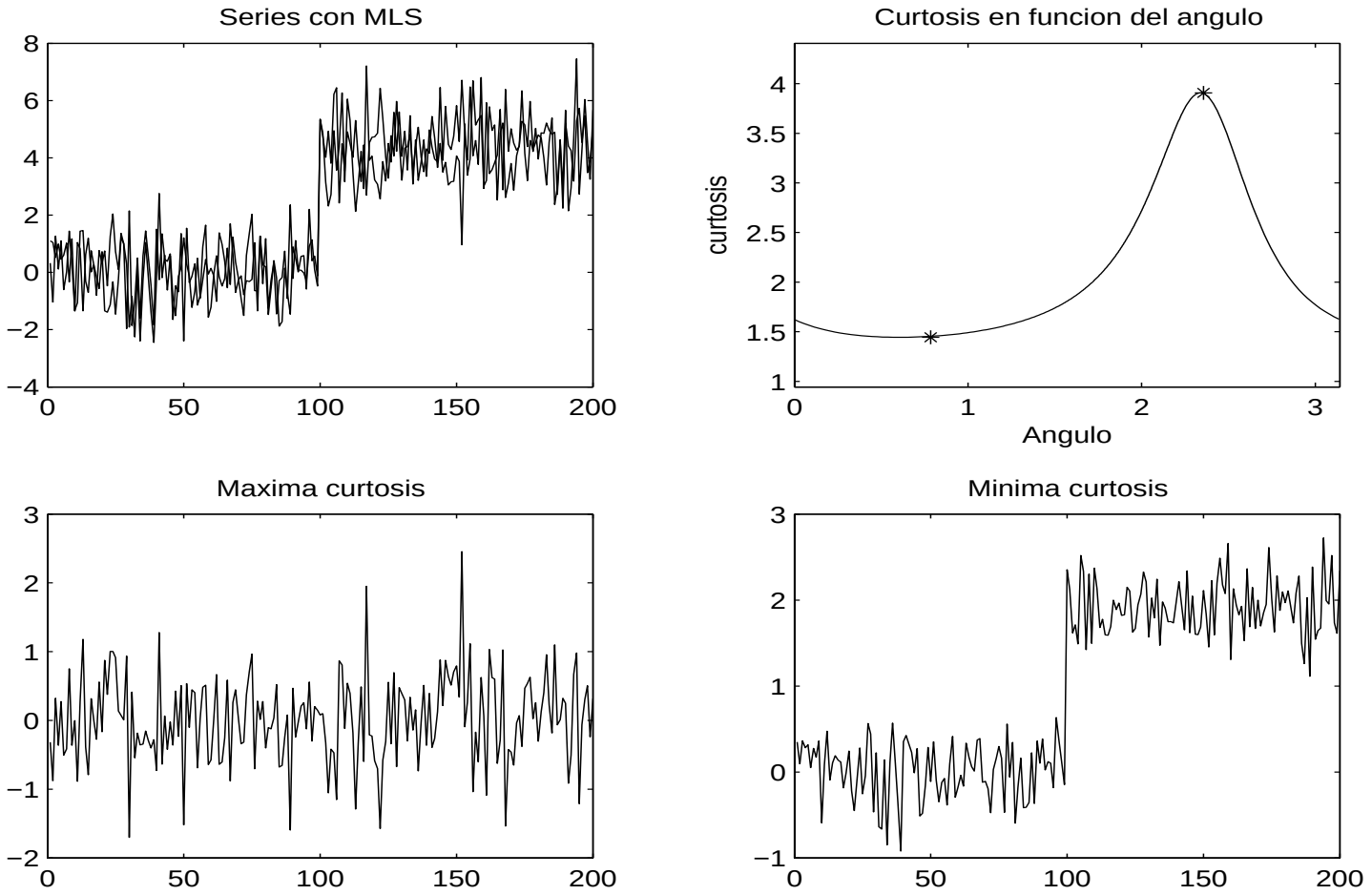
Procedimiento de Galeano, Peña y Tsay

Mismo experimento para un **MLS** en $h = 170$.



Procedimiento de Galeano, Peña y Tsay

Mismo experimento para un **MLS** en $h = 100$.



Procedimiento de Galeano, Peña y Tsay

- Si la serie es estacionaria, proyectamos en las direcciones que maximizan y minimizan el coeficiente de curtosis.

- Entonces, buscamos cambios de nivel mediante **estadísticos de tipo CUSUM** que son más potentes que los LRT para detectar cambios de nivel y son robustos a la presencia del resto de atípicos.

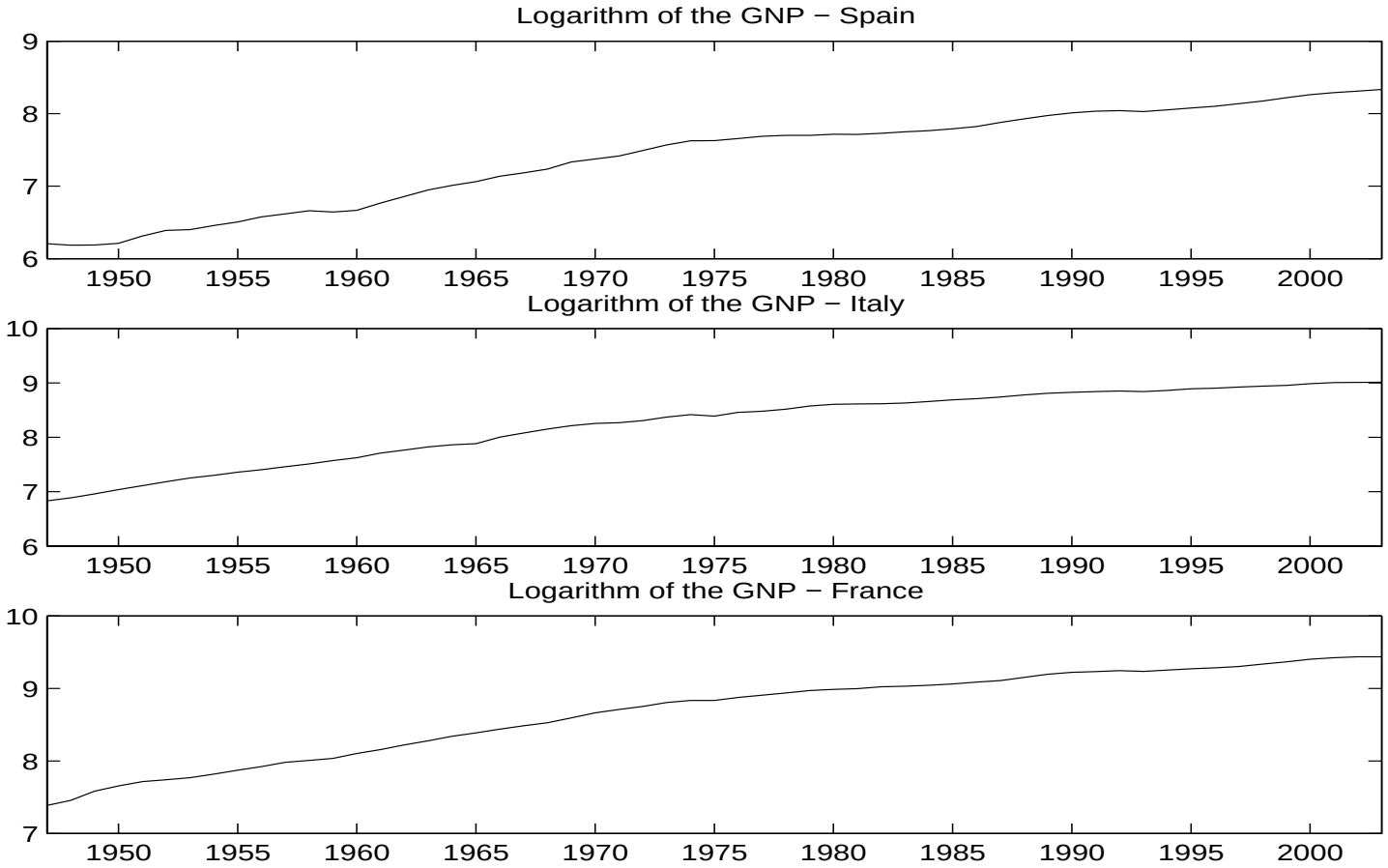
- A continuación, buscamos el resto de atípicos ajustando un autorregresivo de orden alto.

- Si la serie no es estacionaria, podemos tomar diferencias para convertir la serie en estacionaria. Entonces, un cambio de tendencia se convierte en un cambio de nivel en la serie diferenciada, etc. . . El resto del procedimiento es similar.

- Existen múltiples aspectos técnicos que se pueden encontrar en el artículo.

Procedimiento de Galeano, Peña y Tsay

Logaritmo del producto interior bruto (PIB) de España, Italia y Francia de 1947 a 2003.



Procedimiento de Galeano, Peña y Tsay

- Aplicamos el procedimiento para detección de cambios de nivel a las primeras diferencias de la serie. Se detecta un **MRS** en 1975 que afecta fundamentalmente al PIB de España.

- Entonces, aplicamos el procedimiento para detectar el resto de atípicos.

Procedimiento Propuesto						
Iterations	(Λ_I, h_I)	(Λ_A, h_A)	(Λ_L, h_L)	(Λ_T, h_T)	Time	Type
1	(4.11,1966)	(4.05,1965)	(4.78,1966)	(4.22,1966)	1966	MLS
2	(3.37,1976)	(4.77,1975)	(4.15,1975)	(4.45,1975)	1975	MAO
3	(3.14,1960)	(3.74,1960)	(3.49,1960)	(3.68,1960)	—	—
Procedimiento propuesto por Tsay, Peña y Pankratz						
Iterations	(J_I, h_I)	(J_A, h_A)	(J_L, h_L)	(J_T, h_T)	Time	Type
1	(15.08,1966)	(15.54,1965)	(11.39,1975)	(14.11,1966)	—	—

Atípicos en series multivariantes

- Atípicos en series multivariantes mediante componentes independientes en Baragona y Battaglia (2007).
- Atípicos mediante algoritmos genéticos en Baragona, Battaglia y Poli (2011).
- Estimación robusta de modelos para series multivariantes en Croux, Gelper y Mahieu (2010) y Gelper, Fried y Croux (2010).

Contenidos

1. Introducción

2. Atípicos en series temporales univariantes
   (a) Tipos de atípicos
   (b) Procedimientos habituales de detección basados en Chen y Liu (1993)
   (c) Procedimiento de detección de Galeano y Peña (2012)

3. Atípicos en series temporales multivariantes
   (a) Tipos de atípicos
   (b) Procedimiento de detección de Tsay, Peña y Pankratz (2000)
   (c) Procedimiento de detección de Galeano, Peña y Tsay (2006)

4. **Conclusiones**

Conclusiones

- En esta presentación se ha revisado la detección de datos atípicos en series temporales lineales, tanto univariantes como multivariantes.

- En particular, se han presentado varios tipos de atípicos usuales en estos tipos de modelos y diferentes algoritmos, incluyendo algunas propuestas propias.

- Todavía existen problemas en este tipo de algoritmos. Además la extensión a modelos no lineales se está realizando de manera lenta por la dificultad intrinseca de estos procedimientos.

Referencias

Box, G. E. P. y Tiao, G. (1975) "Intervention Analysis with Applications to Economic and Environmental Problems," *Journal of the American Statistical Association*, 70, 70-79.

Chen, C. y Liu, L. (1993) "Joint Estimation of Model Parameters and Outlier Effects in Time Series," *Journal of the American Statistical Association*, 88, 284-297.

Fox, A. J. (1972) "Outliers in Time Series," *Journal of the Royal Statistical Society B*, 34, 350-363.

Galeano, P. y Peña, D. (2012) "Additive outlier detection in seasonal ARIMA models by a modified Bayesian Information Criterion" In *Economic Time Series: Modeling and Seasonality*, Chapman and Hall.

Galeano, P. y Peña, D., and Tsay, R. S. (2006) "Outliers Detection in Multivariate Time Series By Projection Pursuit." *Journal of the American Statistical Association*, 101, 654-669.

Tsay, R. S., Peña, D. and Pankratz, A.E. (2000) "Outliers in Multivariate Time Series," *Biometrika*, 87, 789-804.