

# Capítulo 11

## Diagnosis del modelo y Predicción

**Hirotsugu Akaike** (1919 -)

Estadístico japonés. Ha investigado extensamente en las relaciones entre la teoría de información y la estadística y fue pionero en el desarrollo de criterios de selección de modelos en series temporales. Su criterio AIC se ha utilizado extensamente en todas las ramas de la ciencia. Ha hecho también contribuciones destacadas en la formulación de los modelos de series temporales y ha tenido un papel muy destacado en el desarrollo de la estadística en Japón.

### 11.1 Introducción

La diagnosis del modelo requiere comprobar que las hipótesis básicas realizadas respecto a los residuos son ciertas. Estos deben tener: (1) media cero; (2) varianza constante; (3) falta de correlación para cualquier retardo; (4) distribución normal.

La diagnosis esta relacionada, pero es distinta de la selección de modelos estudiada en el capítulo anterior. Es posible que el mejor modelo seleccionado no sea compatible con las hipótesis realizadas, y entonces tendremos que buscar nuevos modelos. También es posible que muchos modelos pasen los contrastes diagnósticos, y entonces podemos seleccionar el mejor entre ellos con los criterios de selección. Vemos por tanto que la diagnosis es complementaria a los criterios de selección.

Además de comprobar las hipótesis básicas, la etapa de diagnosis del modelo incluye la detección de términos deterministas, si existen. Estos términos van a detectarse por la presencia de términos de media móvil no invertibles. Finalmente, en este capítulo estudiaremos el cálculo de las predicciones con un modelo ARIMA cuando los parámetros se han estimado en una muestra y existe además incertidumbre respecto al modelo correcto que siguen los datos.

### 11.2 Contrastes de autocorrelación

#### 11.2.1 Propiedades de las autocorrelaciones estimadas

El primer contraste a realizar es si los residuos estimados están incorrelados. Para ello se calcula su *fas* mediante:

$$\hat{r}_k = \frac{\sum_{t=1}^{T-k} (\hat{a}_t - \bar{a}) (\hat{a}_{t+k} - \bar{a})}{\sum_{t=1}^T (\hat{a}_t - \bar{a})^2} \quad (11.1)$$

donde  $\bar{a}$  es la media de los  $T$  residuos. Si los residuos son independientes, los coeficientes,  $\hat{r}_k$ , para  $k$  no muy pequeño, serán aproximadamente variables aleatorias con media cero, varianza asintótica  $1/T$  y distribución normal. La varianza asintótica es válida para  $k$  grande, pero no para los primeros retardos.

Por ejemplo, puede demostrarse que si la serie sigue un proceso AR(1) la desviación típica asintótica de la autocorrelación de orden uno para los residuos,  $\hat{r}_1$ , es  $\sqrt{(1 - \phi^2)/T}$ , que puede ser mucho menor que  $1/\sqrt{T}$ . En consecuencia, el valor  $1/\sqrt{T}$  debe considerarse como un límite máximo de la desviación típica de las autocorrelaciones residuales.

El procedimiento habitual de verificar la incorrelación de los residuos es dibujar en sus funciones de autocorrelación simple o parcial estimadas dos líneas paralelas a distancia  $2/\sqrt{T}$  del origen y comprobar si todos los coeficientes  $\hat{r}_k$  están dentro de estos límites de confianza. Como estos límites son, aproximadamente, del 95%, en promedio uno de cada veinte coeficientes de autocorrelación estimados saldrá fuera, por lo que la aparición de un valor significativo en un retardo elevado es esperable. Si embargo, como según lo anterior estos límites sobreestiman la varianza en los retardos pequeños, un valor próximo a los límites de confianza  $\pm 2/\sqrt{T}$  en los retardos iniciales debe considerarse un indicio claro de que el modelo es inadecuado.

### 11.2.2 El contraste de Ljung-Box sobre las autocorrelaciones

Un contraste global de que los primeros  $h$  coeficientes son cero ( $h$  debe ser grande) es el contraste de Ljung-Box. Si los residuos son realmente ruido blanco, hemos visto en el capítulo 3 que los coeficientes de correlación estimados son asintóticamente normales, con media cero y varianza  $(T - k)/T(T + 2)$ . Por tanto, el estadístico

$$Q(h) = T(T + 2) \sum_{j=1}^h \frac{\hat{r}_j^2}{T - j} \quad (11.2)$$

se distribuye, asintóticamente, como una  $\chi^2$  con grados de libertad igual al número de coeficientes en la suma ( $h$ ) menos el número de parámetros estimados,  $n$ . Para modelos no estacionales  $n = p + q + 1$ , o  $n = p + q$ , según que el modelo tenga o no constante, y para los estacionales, que habitualmente no tienen constante,  $n = P + p + Q + q$ . Concluiremos que el modelo es inadecuado si el valor de  $Q(h)$  obtenido con (11.2) es mayor que el percentil 0.95 de la distribución  $\chi^2$  con  $h - n$  grados de libertad, que denotaremos por  $\chi_{.95}^2(h - n)$ . En general rechazaremos el test siempre que la probabilidad :

$$P(\chi^2(h - n) > Q(h))$$

sea pequeña (menor que 0.05 o 0.01).

### 11.2.3 El contraste del determinante

El contraste de Ljung-Box tiene el inconveniente de que da aproximadamente el mismo peso a todos los coeficientes de autocorrelación y es invariante ante permutaciones de los coeficientes. Sin embargo, intuitivamente deberíamos dar más peso a los coeficientes de orden bajo que a los de orden alto. Peña y Rodríguez (2003) han propuesto un contraste más potente que el de Ljung-Box que tiene esta propiedad. El contraste se basa en la matriz de autocorrelación

$$\mathbf{R}_m = \begin{vmatrix} 1 & \hat{r}_1 & \dots & \hat{r}_{m-1} \\ \hat{r}_1 & 1 & \dots & \hat{r}_{m-2} \\ \dots & \dots & 1 & \dots \\ \hat{r}_{m-1} & \hat{r}_{m-2} & \dots & 1 \end{vmatrix}$$

y el estadístico es

$$D_m = -\frac{T}{m+1} \log |\hat{\mathbf{R}}_m|,$$

que sigue asintóticamente una distribución gamma con parámetros

$$\alpha = \frac{3(m+1)\{m-2(p+q)\}^2}{2\{2m(2m+1)-12(m+1)(p+q)\}} \quad \text{y} \quad \beta = \frac{3(m+1)\{m-2(p+q)\}}{2m(2m+1)-12(m+1)(p+q)},$$

La distribución tiene media  $\alpha/\beta = (m+1)/2 - (p+q)$  y varianza  $\alpha/\beta^2 = (m+1)(2m+1)/3m - 2(p+q)$ . Los percentiles del estadístico  $D_m$  para el contraste se obtienen fácilmente calculando los parámetros de la gamma con las fórmulas anteriores y utilizando las tablas de esta distribución (muchos programas de ordenador incluyen esta función). Alternativamente, puede utilizarse que la variable

$$ND_m^* = (\alpha/\beta)^{-1/4} (4/\sqrt{\alpha}) \left( (D_m)^{1/4} - (\alpha/\beta)^{1/4} \left( 1 - \frac{1}{6\alpha} \right) \right), \quad (11.3)$$

se distribuye aproximadamente como una variable normal estandar.

Puede demostrarse que este estadístico puede escribirse como:

$$D_m = T \sum_{i=1}^m \frac{(m+1-i)}{(m+1)} \hat{\pi}_i^2, \quad (11.4)$$

donde  $\hat{\pi}_i^2$  es el cuadrado del coeficiente de correlación parcial de orden  $i$ . Este contraste puede verse como un contraste Ljung-Box modificado, donde en lugar de utilizar los coeficientes de correlación múltiple utilizamos los parciales, pero con ponderaciones  $\frac{(m+1-i)}{m}$ . Estas ponderaciones decrecen linealmente con el retardo, de manera que  $\hat{\pi}_1^2$  tiene peso uno y  $\hat{\pi}_m^2$  peso  $1/m$ .

### 11.2.4 Sobreajuste: contrastes con el criterio BIC

Una técnica complementaria a los contrastes anteriores es la técnica del sobreajuste, que consiste en estimar un modelo de orden mayor que el analizado y comprobar si se obtienen coeficientes estimados significativos. Con esto es posible captar pequeñas estructuras remanentes que pueden mejorar las predicciones, pero que quizás no se detectan claramente en el análisis de los residuos. En general, si hemos ajustado un  $\text{ARIMA}(p, d, q)$  que parece adecuado, el sobreajuste se aplica estimando los modelos  $\text{ARIMA}(p+r, d, q)$  y  $\text{ARIMA}(p, d, q+r)$ , para un valor de  $r$  bajo, normalmente 1 o 2, y comprobando si los parámetros adicionales son significativos.

No conviene ampliar el orden de las partes AR y MA a la vez, ya que puede producirse una compensación de efectos. Si el modelo ajustado inicialmente es,

$$\phi(B) z_t = \theta(B) a_t \quad (11.5)$$

obtendremos un ajuste igualmente bueno con

$$\phi^*(B) z_t = \theta^*(B) a_t \quad (11.6)$$

siendo  $\phi^*(B) = \phi(B)(1 - \phi B)$  y  $\theta^*(B) = \theta(B)(1 - \theta B)$  y  $\phi$  aproximadamente igual a  $\theta$ . Por tanto, si el modelo adecuado es (11.5) y estimamos (11.6) obtendremos todos los parámetros significativos y sólo notaremos la sobreparametrización al factorizar los operadores AR y MA. En consecuencia, siempre es conveniente obtener las raíces de los operadores AR y MA en modelos mixtos y comprobar que no hay factores que se cancelan en ambos lados.

Una manera automática de realizar el sobreajuste es ajustar modelos AR hasta un orden  $p_{\max}$  prefijado a los residuos del modelo y seleccionar el mejor modelo AR mediante el criterio BIC. Si el mejor modelo seleccionado es un  $\text{AR}(0)$ , es decir, un proceso de ruido blanco, aceptamos que los residuos están incorrelados. En caso contrario, rechazamos que el modelo sea adecuado y el grado del mejor modelo seleccionado nos indica como debemos modificar el modelo actual. Como el criterio BIC es consistente, es decir para tamaños muestrales grandes tiende a seleccionar el modelo adecuado con probabilidad que tiende hacia uno, este procedimiento funcionará bien en la práctica si tenemos muestras grandes. Con muestras pequeñas disponemos de menos garantías, y es conveniente realizar siempre el contraste del determinante sobre los residuos estimados. Observemos que este procedimiento no puede aplicarse con el criterio AIC, ya que puede comprobarse que cuando los residuos son ruido blanco todavía existe una probabilidad apreciable, del orden del 15%, de que el AIC seleccione un modelo AR para los residuos (véase Peña y Rodríguez, 2004).

## 11.3 Otros contrastes

### 11.3.1 Contrastes de media cero

Los residuos estimados de un modelo ARIMA no están en principio sujetos a la restricción  $\sum_{t=1}^T \hat{a}_t = 0$ . Esta condición sólo se impone cuando tenemos un modelo AR y sus parámetros se han estimado por mínimos cuadrados, pero en la estimación por máxima verosimilitud exacta esta restricción no existe. Para contrastar la hipótesis de que la perturbación tiene esperanza nula en el caso general, suponiendo  $T$  residuos y  $p + q$  parámetros, calcularemos su media.

$$\bar{a} = \frac{\sum \hat{a}_t}{T}$$

y su varianza

$$\hat{s}_a^2 = \frac{\sum_{t=1}^T (\hat{a}_t - \bar{a})^2}{T - p - q}$$

y concluiremos que  $E[\hat{a}_t] \neq 0$ , si

$$\frac{\bar{a}}{\hat{s}_a / \sqrt{T}}$$

es significativamente grande con relación a la distribución  $N(0, 1)$ . Este contraste debe de aplicarse después de comprobar que los residuos están incorrelados, para asegurar que  $\hat{s}_a$  es un estimador razonable de la varianza.

### 11.3.2 Contrastes de homocedasticidad

La estabilidad de la varianza se comprueba estudiando el gráfico de los residuos a lo largo del tiempo. En los capítulos siguientes estudiaremos contrastes adicionales para ciertas formas de heterocedasticidad. Si a la vista de los residuos estimados parece existir un cambio de varianza a partir de un punto,  $t = n_1$ , podemos dividir el intervalo muestral en dos partes y aplicar un contraste de varianzas. En la hipótesis de que ambos tramos tienen la misma varianza, el estadístico

$$F = \frac{\sum_{t=1}^{n_1} \hat{a}_t^2 / n_1}{\sum_{t=n_1+1}^T \hat{a}_t^2 / (T - n_1)} = \frac{s_1^2}{s_2^2}$$

se distribuirá aproximadamente como una F con  $n_1$  y  $T - n_1$  grados de libertad. Para aplicar este contraste conviene que, al menos, dispongamos de 10 observaciones en cada tramo. De la misma forma, si sospechamos  $h$  cambios de varianza en los periodos  $n_1, \dots, n_h$  el contraste de igualdad de varianzas es

$$\lambda = T \log \hat{\sigma}^2 - \sum_{i=1}^h n_i \log s_i^2$$

donde  $\hat{\sigma}^2$  es la varianza de los residuos en toda la muestra y  $s_i^2$  la varianza en el tramo  $i$  de longitud  $n_i$  observaciones. Bajo la hipótesis de que la varianza es la misma en todos los tramos se demuestra que este estadístico será, asintóticamente, una ji cuadrado con  $h - 1$  grados de libertad.

### 11.3.3 Contrastes de normalidad

La normalidad se comprueba con cualquiera de los test de normalidad. Un contraste simple es calcular los coeficientes de asimetría,  $cs$ , y kurtosis,  $ck$ , de los residuos y utilizar que, bajo la hipótesis de normalidad, la variable

$$X = \frac{Tcs^2}{6} + \frac{T(ck - 3)^2}{24}$$

es una  $\chi^2$  con dos grados de libertad.

Finalmente, conviene *siempre* estudiar el gráfico de los residuos estimados  $\hat{a}_t$  a lo largo del tiempo. Su inspección se facilita situando en él los límites  $\pm 2\hat{\sigma}$  y  $\pm 3\hat{\sigma}$ , y estudiando los puntos que salen fuera de estos límites de control. Con esto podemos detectar valores atípicos cuyo tratamiento se hará en los capítulos siguientes.

### Ejemplo 11.1

La figura 11.1 presenta el gráfico de los residuos de la serie de matriculación estimada con el modelo  $(0,1,1)\times(1,1,1)_{12}$ . Se observan algunos valores atípicos bastante destacados. La *fas* de los residuos se presenta en la figura 11.2. Ningún coeficiente es significativo y el estadístico  $Q$  calculado con 24 coeficientes es 29.5 y con 39 coeficientes es 41. Estos valores deben compararse con una  $\chi^2$  con 21 y 36 grados de libertad respectivamente y sus percentiles al 0.95 son 32.7 y 43.8, con lo que concluimos que, con este contraste, no encontramos ninguna evidencia de dependencia serial en los residuos.

La media de estos residuos es  $-.0017$  y su desviación típica  $.0497$ , que supone un error de predicción porcentual del 4.97%. La media de los residuos no es significativamente distinta de cero, ya que

$$t = \frac{-.0017}{.0497/\sqrt{479}} = -.74$$

y podemos concluir que la media poblacional de los residuos es cero. La variabilidad de los residuos, salvo por unos posible valores atípicos, parece constante en el tiempo. De hecho, si dividimos la serie en dos partes y calculamos la varianza en cada una obtenemos valores muy similares. Concluimos por tanto que, aparte de algunos valores atípicos, los residuos parecen ruido blanco

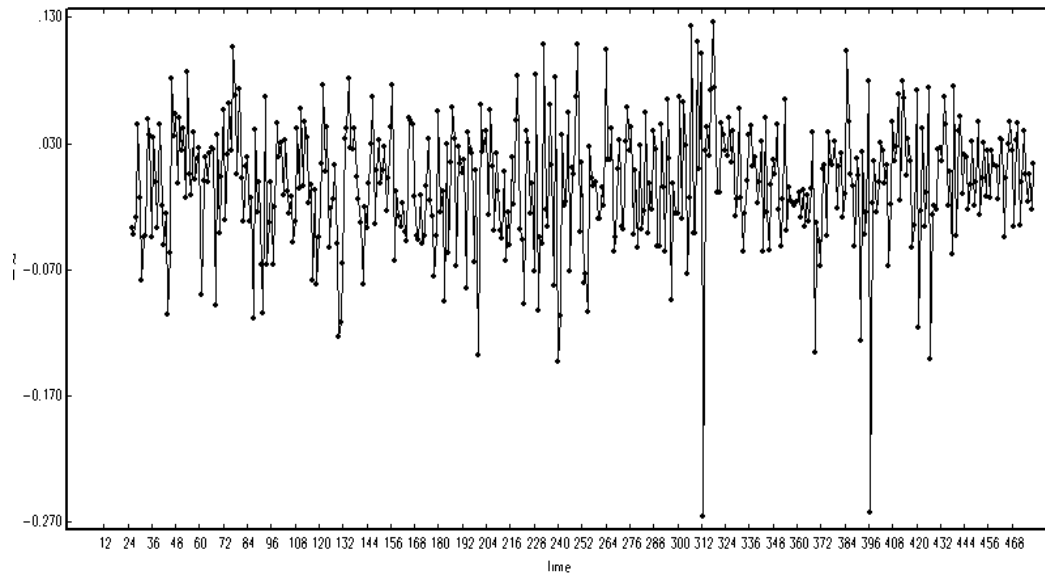


Figura 11.1: Residuos de la serie de matriculación de vehículos con el modelo  $(0,1,1)\times(1,1,1)$ .

## 11.4 Contraste de estabilidad del modelo

Si el modelo es adecuado, los errores de predicción para un período por delante son variables aleatorias normales de media cero, varianza constante e independientes. En consecuencia, si hemos estimado el modelo con datos hasta  $t = T$  y a continuación generamos predicciones que vamos readaptando al obtener los valores observados, y llamamos  $\hat{a}_{T+1}, \dots, \hat{a}_{T+h}$  los errores de predicción de un período adelante, la variable

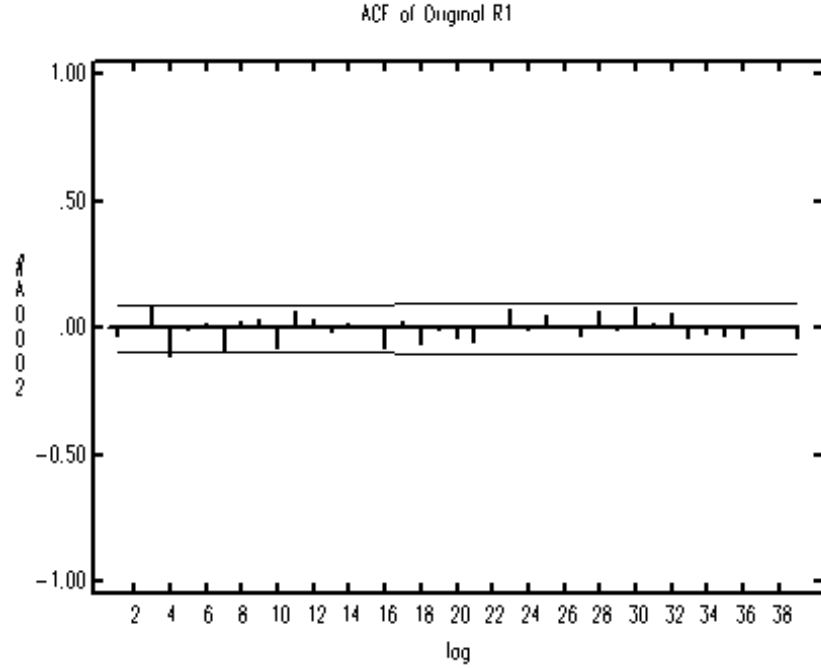


Figura 11.2: ACF de la serie de residuos de matriculación de vehículos

$$Q = \sum_{j=1}^h \hat{a}_{T+j}^2 / \sigma^2$$

será una  $\chi^2$  con  $h$  grados de libertad. Como  $\sigma^2$  se estima mediante  $\hat{\sigma}^2$ , la variable

$$\frac{Q}{h} = \frac{\sum_{j=1}^h \hat{a}_{T+j}^2 / h}{\hat{\sigma}^2} \quad (11.7)$$

será una  $F$  con  $h$  y  $T - p - q$  grados de libertad, siendo  $T$  el número de datos iniciales y  $p + q$  el número de parámetros estimados.

En consecuencia, si  $Q/h$  es significativamente grande, esto sugiere que el modelo no es adecuado.

## 11.5 Interpretación del modelo. Componentes deterministas

Hemos visto que los modelos ARIMA engloban como caso particular modelos con componentes deterministas, como tendencias polinómicas y estacionalidad constante. En la diagnosis del modelo conviene investigar si el modelo ajustado contiene este tipo de componentes. Estos componentes deterministas sólo pueden aparecer cuando el modelo contenga simultaneamente diferencias, regulares o estacionales, y estructura MA regular o estacional con alguna raíz próxima al círculo unidad que pueda cancelarse en ambos miembros. Vamos a analizar estos aspectos.

### 11.5.1 Tendencias

Un caso importante de cancelación de operadores es la producida entre una diferencia regular y una media móvil regular no invertible. Por ejemplo, el modelo

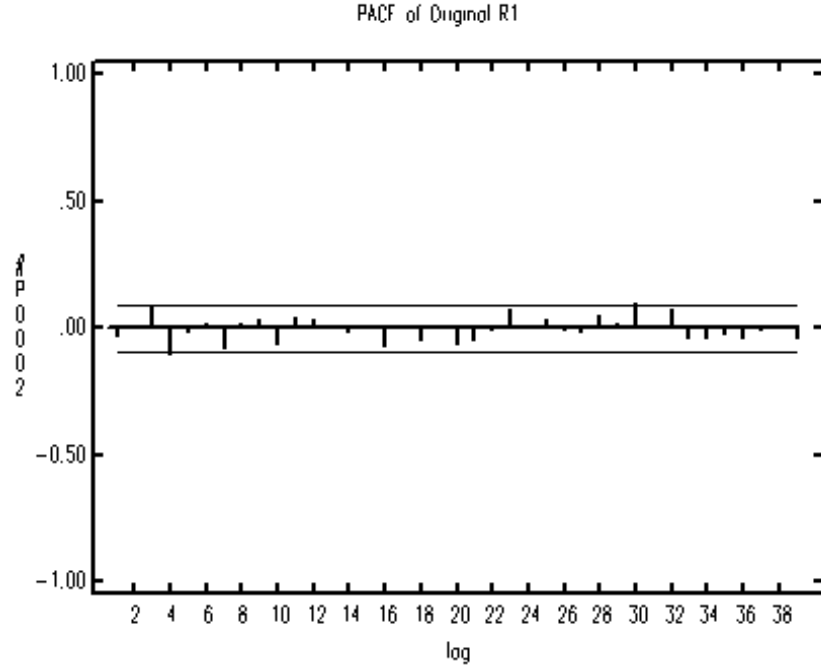


Figura 11.3: PACF de los residuos de la serie de matriculación de vehículos

$$\nabla z_t = (1 - \theta B) a_t \quad (11.8)$$

cuando  $\theta \simeq 1$  es equivalente al modelo determinista:

$$z_t = \beta_0 + a_t \quad (11.9)$$

donde,  $\beta_0$  es una constante, como puede comprobarse tomando diferencias en (11.9), ya que  $\nabla \beta_0 = 0$ . En este caso la aparición de una raíz unitaria en la parte MA es debida a una sobrediferenciación del proceso (11.9) que no requiere diferencias de ningún tipo. Al cancelar las diferencias aparece un término determinista que es simplemente la constante del modelo.

Análogamente el modelo:

$$\nabla z_t = \beta_1 + (1 - \theta B) a_t \quad \text{con } \theta \simeq 1 \quad (11.10)$$

se obtiene al diferenciar el modelo con tendencia lineal determinista:

$$z_t = \beta_0 + \beta_1 t + a_t \quad (11.11)$$

En efecto, al tomar diferencias en (11.11) resulta

$$\nabla z_t = \beta_1 + \nabla a_t$$

ya que  $\nabla \beta_1 t = \beta_1 t - \beta_1 (t - 1) = \beta_1$ . En el modelo (11.10) cuando  $\theta \simeq 1$  la cancelación de las diferencias produce una tendencia lineal.

En general, el modelo

$$\nabla^d z_t = (1 - B)^d a_t \quad (11.12)$$

es equivalente al modelo con tendencia determinista polinómico:

$$z_t = \beta_0 + \beta_1 t + \dots + \beta_d t^{d-1} + a_t \quad (11.13)$$

La demostración de este resultado se obtiene aplicando el operador diferencias  $d$  veces en (11.13). Como caso particular si  $d = 1$  tenemos el modelo (11.8) y si  $d = 2$  un modelo con tendencia lineal (11.12).

### 11.5.2 Estacionalidad

La cancelación de operadores estacionales produce un resultado similar a la cancelación de los operadores regulares, pero la interpretación es diferente porque ahora las constantes los son con el periodo estacional. El modelo que resulta al cancelar las diferencias estacionales en:

$$(1 - B^{12}) z_t = (1 - B^{12}) a_t \quad (11.14)$$

es

$$z_t = S_t^{(12)} + a_t \quad (11.15)$$

donde  $S_t^{(12)}$  es cualquier función periódica cada 12 meses definida por

$$S_t^{(12)} = S_{t-12}^{(12)} \quad \forall t$$

Por ejemplo, si modelamos una serie mensual de temperatura podemos obtener que sigue el modelo periódico:

$$z_t = \cos\left(\frac{2\pi}{12}t + \delta\right) + a_t$$

y tomando una diferencia estacional se reduce al (11.14). Este modelo es compatible con cualquier forma de estacionalidad determinista. Por ejemplo, si una serie mensual tiene una estacionalidad que podemos representar mediante variables ficticias del tipo

$$z_t = \sum_{j=1}^{12} w_j I_t^j + a_t \quad (11.16)$$

donde  $I_t^j = 1$  si  $t = j + 12$  y cero en otro caso, al tomar una diferencia estacional en (11.16) se obtiene de nuevo (11.14).

### 11.5.3 Caso general

Cuando el modelo contiene diferencias estacionales y regulares y términos de media móvil de ambos tipos podemos tener simultáneamente tendencias y estacionalidad determinista. Por ejemplo el modelo de pasajeros de avión con coeficientes próximos a la unidad

$$\nabla \nabla_{12} z_t = (1 - \theta B)(1 - \Theta B) a_t, \quad \theta \approx \Theta \approx 1$$

Es equivalente al modelo

$$z_t = \beta_0 + \beta_1 t + S_t^{(12)} + a_t \quad (11.17)$$

que contiene una tendencia lineal determinista y un estacionalidad también determinista. La comprobación es inmediata tomando diferencias en (11.18).

En consecuencia, cuando tengamos diferencias en el modelo la aparición de un componente MA(1) regular o estacional con parámetro muy próximo a uno, suele indicar un componente estacional determinista. Observemos que estas propiedades son coincidentes con la forma de la ecuación de previsión de un modelo ARIMA estudiadas en el capítulo 8.



## 11.6 Predicción

Las predicciones del modelo estimado pueden realizarse como en el capítulo 8, utilizando los parámetros estimados como si fuesen los verdaderos. Sin embargo, los intervalos de predicción en ese capítulo se calcularon suponiendo los parámetros conocidos y teniendo sólo en cuenta la incertidumbre por el desconocimiento de las innovaciones futuras. En general, cuando el modelo se estima con los datos, tendremos cuatro tipos de incertidumbre en la predicción, asociadas al desconocimiento de :

1. Las innovaciones futuras.
2. La distribución de las innovaciones.
3. Los verdaderos valores de los parámetros.
4. El modelo que ha generado los datos.

La primera fuente de incertidumbre es inevitable y no depende del tamaño muestral. Como los valores futuros de la serie dependen de innovaciones futuras desconocidas, tendremos una incertidumbre creciente con el horizonte de predicción. Ya vimos que esta incertidumbre es acotada para procesos estacionarios, pero crece sin límite para procesos no estacionarios.

Las tres restantes fuentes de incertidumbre dependen del tamaño muestral y aunque son pequeñas para tamaños muestrales grandes, pueden ser importantes en muestras pequeñas (menores de 50 datos). Si tenemos muestras grandes podemos comprobar si la distribución de las innovaciones es normal pero con una muestra pequeña la potencia de un contraste de normalidad es muy baja, con lo que existe siempre cierta incertidumbre respecto a la forma de la distribución de las innovaciones. También tendremos incertidumbre respecto a los valores reales de los parámetros, ya que se han estimado y tienen cierto error de estimación. Finalmente, como el modelo se ha seleccionado a partir de los datos es posible que haya varios modelos compatibles con los datos y aunque hayamos seleccionado alguno con un criterio no podemos ignorar que este modelo puede ser erróneo.

### 11.6.1 Muestras grandes

Si el tamaño muestral es grande, la incertidumbre dominante es la debida a las innovaciones y la otras tres pueden despreciarse. Suponiendo que la distribución de las innovaciones es normal, se justifica en el apéndice 11.1 que la incertidumbre respecto a los parámetros será pequeña. Analogamente, como veremos en la sección siguiente, cuando el tamaño muestral es grande la incertidumbre respecto al modelo será también pequeña.

Suponiendo normalidad, podemos construir intervalos de confianza en muestras grandes para la predicción del 95% tomando los estimadores como parámetros y calculando el intervalo como en el caso en que los parámetros son conocidos. El intervalo será:

$$z_{T+k} \in \hat{z}_T(k) \pm 1,96 \hat{\sigma} \left( 1 + \hat{\psi}_1^2 + \dots + \hat{\psi}_{k-1}^2 \right)^{1/2}$$

### 11.6.2 Intervalos mediante Remuestro

Si las innovaciones no son normales, o si el tamaño muestral no es muy grande y no queremos despreciar la incertidumbre debida a la estimación de los parámetros, podemos utilizar técnicas de remuestro (bootstrap) para generar las predicciones y construir a partir de ellas los intervalos de confianza. Un procedimiento simple de utilizar la verdadera distribución de las innovaciones y tener en cuenta la incertidumbre por la estimación de los parámetros es el siguiente (véase Tombs y Chucany, 1993 y Pascual, Romo y Ruiz, 2000):

1. Estimar el modelo y obtener los parámetros,  $\hat{\phi}$  y  $\hat{\theta}$  y los residuos.
2. Obtener una muestra con reemplazamiento de  $T$  residuos de la distribución de los residuos estimados. Sean  $a_1^*, \dots, a_T^*$  la muestra obtenida.
3. Reconstruir una nueva serie con los parámetros estimados y estos residuos mediante:

$$z_t^* = \hat{\phi}_1 z_{t-1}^* + \dots + \hat{\phi}_h z_{t-h}^* + a_t^* - \hat{\theta}_1^* a_{t-1}^* - \dots - \hat{\theta}_q^* a_{t-q}^*$$

tomando como valores iniciales  $z_1 = \dots = z_h = \bar{z}$  y  $a_{-q} = \dots = a_0 = 0$ .

4. Estimar los parámetros en la serie anterior. Sean  $\hat{\phi}^*$  y  $\hat{\theta}^*$  los nuevos parámetros estimados.
5. Calcular predicciones de valores futuros de la serie,  $\hat{z}_T(k)$  utilizando los parámetros estimados en la etapa anterior,  $\hat{\phi}^*$  y  $\hat{\theta}^*$ , pero con valores de  $z$  iguales a los últimos verdaderos valores observados  $z_{T-1}, \dots, z_{T-h}$  e innovaciones iguales a las reales que hemos estimado para la serie  $a_{t-1}, \dots, a_{t-q}$ .
6. Repetir los pasos 2 a 5 un número grande de veces,  $N$ . Obtener los intervalos de confianza de la distribución de  $\hat{z}_T(k)$  a partir de los  $N$  valores obtenidos.

Este enfoque puede generalizarse para tener en cuenta también la incertidumbre en la estimación del modelo. Si podemos encontrar una distribución para los posibles ordenes del modelo, podemos generar realizaciones del proceso con distintos órdenes y en cada una calcular los errores de predicción. De esta manera obtenemos una distribución generada de errores de predicción que podemos utilizar para construir intervalos de confianza. (véase Alonso, Peña y Romo, 2003, para una implantación de este enfoque).

### 11.6.3 Predicción mediante promedio de modelos

Supongamos que se han estimado  $k$  modelos posibles para una serie y que disponemos de los valores del BIC para cada modelo. Según vimos en el capítulo anterior, los valores del BIC son, salvo constantes aditivas,  $-2\log P(M_i|D)$ , donde  $P(M_i|D)$  es la probabilidad del modelo  $i$  dados los datos  $D$ . Entonces, como

$$BIC_i = -2\log P(M_i|D) + c$$

tenemos que

$$P(M_i|D) = c_1 e^{-\frac{1}{2}BIC_i}$$

Para calcular la constante  $c_1$  utilizamos que la suma de las probabilidades de todos los modelos debe ser la unidad. Por tanto

$$\sum_{j=1}^k P(M_j|D) = 1 = c_1 \sum_{j=1}^k e^{-\frac{1}{2}BIC_j}$$

Con estos resultados, podemos transformar los valores del BIC en probabilidades a posteriori de los distintos modelos mediante

$$P(M_i|D) = \frac{e^{-\frac{1}{2}BIC_i}}{\sum_{j=1}^k e^{-\frac{1}{2}BIC_j}}$$

La distribución de probabilidad de una nueva observación será entonces una distribución mezclada

$$\sum P(M_i|D)f(z|M_i) \tag{11.18}$$

donde  $f(z|M_i)$  la función de densidad de la nueva observación de acuerdo con el modelo  $i$ . Esta distribución será normal, con media igual a la predicción con dicho modelo y varianza igual a la de esta predicción. Por ejemplo, para un periodo adelante si generamos predicciones con cada modelo y llamamos  $\hat{z}_T^{(i)}(1)$  a la predicción para el periodo  $T+1$  con el modelo  $i$ , la esperanza de la distribución (11.18), que llamaremos  $\hat{z}_T(1)$  será:

$$\hat{z}_T(1) = \sum_{i=1}^k \hat{z}_T^{(i)}(1)P(M_i|D)$$

que es el resultado de combinar las predicciones de todos los modelos para obtener una predicción agregada única. Esta forma de calcular las predicciones se conoce como promedio Bayesiano de modelos (Bayesian Model Averaging). En general esta predicción será más precisa, en promedio, que la generada por un modelo único. Además permitirá construir intervalos de predicción más realistas que los obtenidos ignorando la incertidumbre de los parámetros y del modelo. Llamando  $\hat{\sigma}_i^2$  a la varianza de las innovaciones del modelo  $i$ , que coincidirá con el error de predicción a un paso con este modelo, la varianza de la combinación de modelos será:

$$var(\hat{z}_T(1)) = E[(z_{T+1} - \hat{z}_T(1))^2]$$

Modelo	BIC	P(M/Datos)
ARIMA(0,1,1)×(0, 1, 1) <sub>12</sub>	8.08	.913
ARIMA(0,1,1)×(1, 1, 1) <sub>12</sub>	14.14	.044
ARIMA(0,1,1)×(2, 1, 0) <sub>12</sub>	14.22	.043

Tabla 11.1: Valores de los criterios de seleccion con la serie de matriculación

donde esta esperanza se calcula respecto a la distribución (11.18). Puede demostrarse (véase ejercicio 11.3) que la varianza de la distribución mezclada es

$$\text{var}(\hat{z}_T(1)) = \sum_{i=1}^k \hat{\sigma}_i^2 P(M_i|D) + \sum_{i=1}^k (\hat{z}_T^{(i)}(1) - \hat{z}_T(1))^2 P(M_i|D)$$

que permitirá construir intervalos más realistas para las predicciones. Estos resultados se generalizan para cualquier intervalo de predicción.

### Ejemplo 11.1

Vamos a ver como se combinarían las predicciones para la serie de matriculación de vehículo. Supongamos que los tres modelos posibles son los indicado en la Tabla 11.1, que se estimaron en el capítulo anterior.

A partir de los datos del BIC de cada modelo podemos calcular las probabilidades a posteriori de cada modelo mediante

$$p(M_1|D) = \frac{\exp(-\frac{1}{2}8.08)}{\exp(-\frac{1}{2}8.08) + \exp(-\frac{1}{2}14.14) + \exp(-\frac{1}{2}14.22)} = .913$$

y análogamente para los otros modelos. En este caso la incertidumbre respecto al modelo es pequeña y las predicciones obtenidas mediante la combinación serán similares a las del mejor modelo. Sin embargo, los intervalos de predicción serán más grandes, debido a la incertidumbre adicional entre los modelos.

### Ejercicios 11

11.1 Ajustar un modelo a los datos del paro en España y analizar los residuos del modelo para comprobar que son adecuados.

11.2 Demostrar que la media de la distribución  $\alpha f(\mu_1, \sigma_1^2) + (1 - \alpha)f(\mu_2, \sigma_2^2)$  es  $\mu = \alpha\mu_1 + (1 - \alpha)\mu_2$  y la varianza  $\alpha\sigma_1^2 + (1 - \alpha)\sigma_2^2 + \alpha(\mu_1 - \mu)^2 + (1 - \alpha)(\mu_2 - \mu)^2$ .

11.3 Utilizar el resultado anterior para demostrar la fórmula de la media y la varianza cuando combinamos predicciones con distintos modelos

11.4 Demostrar que si el modelo más probable es el de menor varianza residual los intervalos contruidos por el promedio de modelos serán más amplios que los de un modelo único.

## 11.7 Apéndice 11: predicción con incertidumbre de los parámetros y del modelo

Para ilustrar el error debido a la estimación de los parámetros, supongamos que hemos decidido que el modelo es AR(1) y generamos predicciones con :

$$\hat{z}_T(1) = \hat{\phi}z_T$$

donde suponemos que el parámetro  $\phi$  se estima por mínimos cuadrados como en un modelo de regresión. Entonces su estimación es  $\hat{\phi} = \sum_{t=2}^T z_t z_{t-1} / \sum_{t=2}^T z_{t-1}^2$  y la varianza de la estimación es  $\text{var}(\hat{\phi}) = \sigma^2 / \sum_{t=2}^T z_{t-1}^2$ . Llamando  $e_T(1) = a_{T+1}$ , tenemos

$$e_T^*(1) = a_{T+1} + (\phi - \hat{\phi})z_T.$$

Elevando al cuadrado esta expresión y tomamos esperanzas respecto a la distribución conjunta de las observaciones, se obtiene

$$E(e_T^*(1)^2) = E(a_{T+1}^2) + E[(\phi - \hat{\phi})^2 z_T^2]$$

ya que el doble producto se anula al ser  $a_{T+1}$  independiente de los valores muestrales observados y por tanto independiente de  $\hat{\phi}$ , que depende de los valores pasados, y de  $z_T$ . Para calcular la esperanza del segundo utilizaremos la propiedad del Apéndice 8.1 que nos dice que esta esperanza respecto a la distribución conjunta de los datos puede calcularse en dos etapas. Podemos calcular primero la esperanza respecto a la distribución de  $(z_1, \dots, z_{T-1} | z_T)$ , es decir condicionar al valor  $z_T$ , y después la esperanza del resultado respecto a la distribución de  $z_T$ . De esta manera, en la primera esperanza  $z_T$  se toma como constante y se considera la variabilidad debida a la estimación de  $\phi$  en distintas realizaciones del proceso que producen distintos valores del estimador, aunque suponemos que todas estas realizaciones tienen el mismo valor de  $z_T$ . Al condicionar en  $z_T$  :

$$E \left[ (\phi - \hat{\phi})^2 z_T^2 | z_T \right] = z_T^2 E(\phi - \hat{\phi})^2 = z_T^2 \text{var}(\hat{\phi})$$

donde hemos supuesto que la  $E \left[ (\phi - \hat{\phi})^2 | z_T \right] \approx E(\phi - \hat{\phi})^2$ , la varianza del estimador. Entonces sustituyendo la expresión de la varianza,  $\text{var}(\hat{\phi}) = \sigma^2 / T s_z^2$ , donde hemos llamado  $s_z^2 = \sum z_{t-1}^2 / T$ , tenemos que

$$\text{var}(e_T^*(1) | z_T) = \sigma^2 (1 + z_T^2 / T s_z^2)$$

Si ahora tomamos esperanzas respecto  $z_T$ , como el cociente  $z_T^2 / s_z^2$  tiene esperanza uno obtenemos que  $\text{var}(e_T^*(1)) = \sigma^2 (1 + 1/T)$ . Si  $T$  es grande podemos despreciar el segundo término frente al primero, lo que implica despreciar la incertidumbre asociada a la estimación de los parámetros del modelo. En general, en muestras grandes la incertidumbre adicional por los parámetros estimados es despreciable.

Para ilustrar el error debido a la incertidumbre en el modelo, supongamos que hemos decidido que el modelo es AR(1) cuando el modelo es AR(p). El error real de predicción será

$$e_T^*(1) = z_{T+1} - \hat{\phi} z_T = a_{T+1} + (\phi_1 - \hat{\phi}) z_T + \dots + \phi_p z_{T-p} \quad (11.19)$$

Podemos calcular la incertidumbre si el proceso fuese en realidad AR(p) elevando al cuadrado en (11.19) y tomando esperanzas condicionadas a los valores  $z_T, \dots, z_{T-p}$ , con lo que tenemos que

$$\text{var}(e_T^*(1) | z_T, \dots, z_{T-p}) = \sigma^2 (1 + z_T^2 / T s_z^2) + \sum \phi_j \phi_i z_{T-i} z_{T-j}$$

pero esta expresión no es útil porque desconocemos los valores de los parámetros. Si seleccionamos el modelo con el criterio BIC para tamaños muestrales grandes las probabilidades del modelo correcto tenderán a uno, con lo que si hacemos predicciones combinando modelos tendremos el mismo resultado que si utilizamos el modelo seleccionado con el BIC.