Protocolo para la identificación de modelos

ARIMA en series temporales

(según los pasos de Box-Jenkins)

1. **Representar** gráficamente la **serie**, su **función de autocorrelación simple** (**FAS**) y su **función de autocorrelación parcial** (**FAP**). La gráfica de la serie nos indica si la serie es estacionaria o no. Según los motivos por los que la serie no es estacionaria, tendremos que aplicar uno de los siguientes procedimientos hasta hacerla estacionaria.

a) Si tiene tendencia: Tomaremos **diferencias regulares** hasta que desaparezca. Normalmente el orden de la diferencia será 1, y raramente será mayor a 3.

b) Si es heterocedástica, es decir, no tiene varianza constante, habrá que transformar la serie. Con tomar el **logaritmo** es suficiente en nuestro caso, aunque existen algunas transformaciones más complejas, como las de Box-Cox.

c) Si es estacional: Tomaremos **diferencias estacionales** hasta que desaparezca el patrón que se repite. En la práctica es muy raro tener que aplicar más de una diferencia estacional.

Una vez que el gráfico de la nueva serie (una transformación de la original) indica que es estacionaria, podemos intentar deducir su estructura (¡no la de la serie original!) observando la FAS y la FAP, teniendo en cuenta que:

a) El **decrecimiento** de la altura **de las barras** (son los coeficientes de la función) debe ser **exponencial**. También suelen aparecer una **alternancia de los signos** o una **forma sinusoidal**; la alternancia depende de los signos de los parámetros, y la forma sinusoidal es debida a que hay raíces complejas en la ecuación característica del proceso. Si no hay decrecimiento exponencial, suele deberse a que la serie es integrada; para éstas el decrecimiento es lineal, no exponencial. (Hay un tipo de series estacionarias que tienen memoria mucho más larga que la de los AR, se las llama de «memoria larga», que muestran ese decrecimiento lineal; pero no se han visto y por tanto debéis pensar en las integradas cuando haya decrecimiento lineal.) Una serie es integrada cuando al tomar diferencias («diferenciarla») se convierte en una serie estacionaria; si esta serie estacionaria es un modelo ARMA(p,q) decimos que son modelos ARIMA(p,d,q).

b) Una de las dos funciones de autocorrelación va a mostrar una estructura como la descrita en a), mientras que la otra mostrará **algunas barras que salen de la banda de significatividad**. La posición de la última de estas barras, que suele coincidir con el número de barras que sobresalen, nos dicen el orden del proceso. Según sea la FAS la que presenta la estructura de a) o sea la FAP, diremos que el proceso es AR o MA, respectivamente.

c) Hay situaciones que no son muy claras, en estos casos puede tratarse de un modelo con las dos partes, la AR y la MA. Estos son los modelos ARMA(p,q). Sus funciones FAS y FAP son combinaciones de las de ambas partes, por lo que son difíciles de identificar a simple vista.

d) La FAS y la FAP de un ruido blanco (serie de datos procedentes de variables sin ninguna estructura común, es decir, independientes entre sí) tiene todos los coeficientes nulos en teoría, o no significativos en nuestras gráficas. Respecto a la banda de significatividad, se advierte en el libro «Análisis de series temporales», de Daniel Peña, que: *«como […] estos límites sobrestiman la varianza en los retardos pequeños, un valor próximo a los límites de confianza ±2/√T en los retardos iniciales debe considerarse un indicio claro de que el modelo es inadecuado»* (T es la longitud de la serie).

e) Tened en cuenta que las líneas de significatividad que aparecen en las gráficas de las funciones de autocorrelación suelen determinar una **zona de una significatividad del 95%**, es decir, el coeficiente que sobresale se puede considerar no nulo con un 95% de seguridad. Esto implica que **puede haber un 5% de barras que sobresalgan sin motivo real**. Debemos pensar esto cuando hay alguna barra aislada que no cuadra con el resto de la estructura de barras. Debido a los estimadores utilizados para esas barras/coeficientes, y para garantizar que las estimaciones se basan en un número suficiente de datos, el programa elige el número de barras que representa. Cuando hay 20, por ejemplo, puede suceder que 1 se haya salido erróneamente de la banda de significatividad.

f) Una opción que también se puede utilizar en este primer análisis, pero que no se ha visto en las prácticas, es aplicar el **contraste de aleatoridad de Box-Pierce** (marcándolo en la parte izquierda de la ventana). Este contraste está diseñado para estudiar si se puede considerar que una serie está formada por variables incorreladas o no (para variables con distribución normal, ser independientes y estar incorreladas son condiciones equivalentes). En el siguiente apartado, que consiste en otro tipo de análisis, se lo aplicaremos a los residuos; pero ahora podemos aplicárselo a nuestra serie original. Esto nos dirá si los datos originales se pueden considerar ruido blanco o no según si su **p-valor es mayor que 0,05** o no; no obstante, a no ser que este p-valor sea muy alto y por el resto de información estemos bastante seguros de que nuestra serie es efectivamente ruido, intentaremos ajustar un modelo, e interpretaremos este contraste solamente como una medida de la «estructura» que no se debe al azar: a mayor p-valor, menos estructura hay bajo nuestros datos. Observad que este contraste para una serie original produce exactamente el mismo resultado que cuando ajustamos un modelo ARIMA(0,0,0) a la serie original… ¡porque con este ajuste ella y sus residuos son la misma serie!

2. Observando las tres gráficas anteriores podemos hacernos una idea de qué modelo subyace a nuestra serie, o al menos de cuáles son los primeros candidatos que debemos probar. Para comprobar analíticamente (no visualmente) un modelo, vamos a intentar **ajustar un modelo ARIMA(p,d,q)** y observaremos dos cosas que miden lo bien que ajusta el modelo: lo necesarios que son todos sus términos, y lo que se parecen al ruido blanco los residuos que nos quedan. Estos son los dos análisis que hemos visto en la Práctica 4. Ajustamos el modelo ARIMA(p,d,q) según lo que hayamos deducido de todos

los apartados anteriores: el valor de «d» es el número de diferencias regulares que hemos tenido que tomar en «1.» para hacer la serie estacionaria, «p» es el orden de la parte AR y «q» el de la parte MA. Frecuentemente vamos a tener que probar varios modelos ARIMA(p,d,q) hasta que encontremos el modelo que nos convence, incluso cuando se cumplan las condiciones que se exigen en los siguientes análisis:

a) Las condiciones en el primero son que, para cada coeficiente del modelo propuesto, **el valor absoluto del estadístico «t» sea mayor que 2**, a la vez que el **p-valor** de ese contraste (que mide si se puede considerar cero o no) es **menor que 0,05**. Los coeficientes que no cumplan estas dos condiciones no son necesarios en el modelo, porque no hay evidencia de que deban ser no nulos. Esto nos ayuda en la búsqueda de nuestro modelo lo más sencillo posible.

b) El otro análisis nos proporciona varios contrastes de hipótesis. Vamos a fijarnos en el de **Box-Pierce**, aplicado a los residuos después de ajustar el modelo. En este caso el contraste confirma con más fuerza que los residuos son ruido blanco cuanto mayor es su p-valor. El nivel mínimo es 0,05, es decir, debe suceder que **el p-valor sea mayor que 0,05**; y cuando mayor sea, más evidencia hay de que los residuos son ruido blanco (y mejor hemos captado en el modelo la parte que no se debe al azar).

c) Otro recurso (gráfico) es el de observar, tras ajustar el modelo ARIMA, las funciones **FAS y FAP de los residuos**, lo que nos dará una información análoga a la del contraste de aleatoriedad de Box-Pierce, pero de forma gráfica. Hacemos aparecer estas gráficas marcándolas tras apretar el botón *«Graphical options»*.

Finalmente, recordad que si habéis transformado la serie original, el modelo **encontrado es el de la transformación, no el de la serie original**. Esto se puede indicar diciendo que, después de hacer las transformaciones «tal y tal», la serie resultante tiene estructura «cual». En los modelos ARIMA(p,d,q) esto se aclara automáticamente, pues debemos entender su notación como: «Después de tomar una diferencia regular de orden 'd', la serie estacionaria que queda tiene una estructura de ARMA(p,q)»

Puesto que no hay un único modelo correcto (que explica suficientemente bien los datos), de la misma manera que no hay una única función que pase por varios puntos que tracemos en un sistema de ejes XY, algunos criterios que se pueden seguir para elegir un modelo satisfactorio son:

- Debe verificar las condiciones gráficas y numéricas necesarias.
- Sus residuos después de ajustar el modelo deben ser ruido, os recuerdo de nuevo lo que se advierte en el libro «Análisis de series temporales», de Daniel Peña: *«como […] estos límites sobrestiman 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»* (T es la longitud de la serie).
- Entre dos modelos que sólo difieran en el p-valor del contraste de Box-Pierce, preferimos el que tenga un valor mayor, puesto que da más garantías de que los residuos que proporciona el modelo son ruido.

- Entre los que verifican las mismas condiciones numéricas más o menos, queremos tomar el más sencillo posible, es decir, el que necesite menos parámetros. Los parámetros que necesita un modelo ARMA(p,q) son «p+q».
- Cuando el orden de una parte, AR ó MA, es muy alto, podemos probar a añadir algún término de la otra parte y restarle orden a ésta.
- En último término, la elección del modelo final debe ser un equilibrio entre simplicidad y eficacia, porque siempre sería posible añadir parámetros indefinidamente para obtener mejoras…

Observación: Aquí se han descrito los pasos que habría que seguir para el caso de series que no tienen parte estacional. Cuando la tuvieran, la metodología sería la misma, con la excepción de que habría que ajustar un modelo *ARIMA(p,d,q)xARIMA(P,D,Q)*$_s$.

Bibliografía citada:

[1] Peña, Daniel. *Análisis de series temporales*. Madrid: Alianza Editorial, S.A., 2005

David Casado de Lucas: http://www.est.uc3m.es/dcasado/, david.casado@uc3m.es
Departamento de Estadística
Universidad Carlos III de Madrid
C/ Madrid 126
28903 Getafe (Madrid)

27/01/06