Procesos ARMA

El tema central de este capítulo es el de introducir los procesos ARMA que son bastamente utilizados en todas las ramas de la ciencia para modelar series de tiempo. El esquema de este capítulo es el siguiente. En la sección 1 incluimos algunos resultados teóricos de carácter general como lo son la estimación de la media y de la función de autocovarianza de un proceso en L^2, un listado de test de aleatoriedad y finalmente ver cómo se pueden realizar predicciones de valores futuros a partir de los datos observados para un proceso estacionario. En la sección 2 damos la definición de los procesos ARMA y vemos sus conceptos teóricos. En la sección 3, vemos cómo puede llevarse a cabo la predicción para un proceso ARMA en particular. En la sección 4 vemos cómo se estiman los parámetros en un proceso ARMA y las propiedades asintóticas que tienen los estimadores. En la sección 5 vemos criterios para la elección de los valores de p y q para ajustar a un conjunto de datos un ARMA(p,q). La sección 6 está dedicada a lo que se llama el diagnóstico del modelo (ver qué tan bien ajusta el modelo a los datos observados). En la sección 7 se encuentran los pasos a realizar para modelar una serie de datos mediante procesos ARMA. En las secciones 8 y 9 se presentan los modelos ARIMA y SARIMA respectivamente y en la sección 10 se encuentran los ejercicios del capítulo.

1 Preliminares

Antes de comenzar con los procesos ARMA que es la principal herramienta de este capítulo veremos algunos resultados de carácter general para procesos estocásticos estacionarios que son de gran utilidad para realizar inferencia.

1.1 Estimación de la media y de la función de autocovarianza

El siguiente teorema vale para cualquier serie de tiempo estacionaria y sirve para estimar, de manera consistente la media.

Teorema 1 *Si $X_1, X_2, ..., X_n$ son las variables correspondientes al proceso estacionario $\{X_t\}_{t\in\mathbb{Z}}$ en L^2, con media μ y función de autocovarianza $\gamma(h)$ tal que $\gamma(h) \to 0$ cuando $|h| \to +\infty$. Entonces*

1. $\lim_{n\to+\infty} \mathbb{V}\left(\overline{X}_n\right) = 0.$
2. *Si* $\sum_{h=-\infty}^{+\infty} |\gamma(h)| < +\infty$ *entonces* $\lim_{n\to+\infty} n\mathbb{V}\left(\overline{X}_n\right) = \sum_{h=-\infty}^{+\infty} \gamma(h).$

Observación 2 *Observamos que la condición 1, nos dice que*

$$\mathbb{V}\left(\overline{X}_n\right) = \mathbb{E}\left(\left(\overline{X}_n - \mu\right)^2\right) \to 0,$$

lo que significa que $\overline{X}_n$ *converge a* μ *en media cuadrática, por lo que* $\overline{X}_n$ *es un estimador consistente (en probabilidad) de* μ.

Observación 3 *La condición 1 nos dice que la ley de los grandes números para un proceso estacionario es válida.*

Al igual que la ley de los grandes números recién comentada, hay una versión del teorema central del límite para $\overline{X}_n$ en el caso de procesos estacionarios que enunciamos a continuación.

Teorema 4 *Teorema central del límite para procesos estacionarios en* L^2.

Si $X_1, X_2, ..., X_n$ *son las variables correspondientes al proceso estacionario* $\{X_t\}_{t\in\mathbb{Z}}$ *en* L^2, *con media* μ *y función de autocovarianza* $\gamma(h)$ *tal que*

$\sum_{h=-\infty}^{+\infty} |\gamma(h)| < +\infty$, *entonces*

$$\frac{\sqrt{n}}{\sigma_n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} N(0,1) \text{ cuando } n \to +\infty$$

siendo $\sigma_n^2 = \sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma(h)$.

Observación 5 *El resultado anterior permite confeccionar intervalos de confianza aproximados para el valor de* μ. *Por ejemplo, un intervalo de confianza de nivel aproximado al* 95% *para* μ *sería*

$$\left(\overline{X}_n - \frac{1.96\widehat{\sigma}_n}{\sqrt{n}}, \overline{X}_n + \frac{1.96\widehat{\sigma}_n}{\sqrt{n}}\right)$$

siendo $\widehat{\sigma}_n = \sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\widehat{\gamma}(h)$ *el estimador natural de* σ_n.

1.2 Tests de aleatoriedad

Dada un muestra de observaciones $X_1, X_2, ..., X_n$, ¿cuándo estamos ante presencia de observaciones i.i.d.? Existen un conjunto de test de hipótesis para verificar esta pregunta. Si bien no existe ningún test universalmente consistente para responder a esta pregunta, existen un conjunto de tests que al menos detectan posibles casos de dependencia. Enunciamos algunos de ellos en las próximas subsecciones. En todos ellos nos planteamos el test $H_0 : X_1, X_2, ..., X_n$ son i.i.d. vs $H_1 : H_0$ no es cierto,

1.2.1 Test de Box–Pierce(1970)

Si $X_1, X_2, ..., X_n$ son i.i.d. en L^2 y le llamamos $\widehat{\rho}(j)$ al coeficiente de correlación muestral de rezago (lag) j, entonces $\sqrt{n}\widehat{\rho}(j)$ es aproximadamente (para n grande) $N(0,1)$ y son independientes cuando variamos j) y por lo tanto el

estadístico $Q = n\sum_{j=1}^{H}\widehat{\rho}^2(j)$ tendrá distribución aproximada $\chi^2(H)$. Este sencillo resultado nos permite plantear como región crítica asintótica de nivel α al conjunto

$$R = \left\{Q > \chi^2_{1-\alpha}(H)\right\}$$

siendo $\chi^2_{1-\alpha}(H) = F^{-1}(1-\alpha)$ y F la función de distribución de una variable $\chi^2(H)$. En R, $\chi^2_{1-\alpha}(H)$ se calcula mediante qchisq$(1-\alpha, H)$. La función en R Box.test(x,lag=H) siendo x un vector de datos lleva a cabo este test.

1.2.2 Test de Ljung–Box (1978)

Es prácticamente el mismo test anterior, con la diferencia en que se utiliza una mejor aproximación a la variable $\chi^2(H)$ utilizando como estadístico de prueba $Q_{LB} = n(n+2)\sum_{j=1}^{H}\frac{\widehat{\rho}^2(j)}{n-j}$. En la prueba de Ljung-Box la región crítica es la misma que la de Box–Pierce, cambiando Q por Q_{LB}.

La función en R Box.test(x,lag=H,type="Ljung-Box") siendo x un vector de datos lleva a cabo este test.

1.2.3 Test de McLeod & Li (1983)

Plantean utilizar en lugar de Q, el estadístico $Q_{ML} = n(n+2)\sum_{j=1}^{H}\frac{\widehat{\rho}^2_{WW}(j)}{n-j}$ siendo $\widehat{\rho}_{WW}(j)$ el coeficiente de correlación muestral de rezago j entre las observaciones al cuadrado. Esto es

$$\widehat{\rho}_{WW}(j) = \frac{1}{n}\sum_{i=1}^{n-j} X_i^2 X_{i+j}^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} X_i^2\right)^2.$$

Observación 6 *Observar que en los test de Box-Pierce, Ljung-Box y McLeod-Li, se debe elegir el valor de H para realizar el test. El valor de H sería bueno que sea relativamente alto, pero pequeño en relación con n por dos motivos: en primer lugar porque la convergencia a la χ^2 se obtiene para $n \to +\infty$ con H fijo, y en segundo lugar, tener en cuenta que en la medida en que crece el valor de j, menos datos se tienen para estimar $\rho(j)$ por lo que la estimación pierde calidad.*

1.2.4 Test sobre el punto de retorno

Si $X_1, X_2, ..., X_n$ son observaciones, se dice que hay un punto de retorno en el tiempo $i = 2, 3, ..., n-1$, si se cumple que $X_i > X_{i-1}, X_{i+1}$ o si se cumple que $X_i < X_{i-1}, X_{i+1}$.

Si le llamamos $T_n =$ cantidad de puntos de retorno entre las n observaciones, se puede probar que si las observaciones son i.i.d. entonces $\mu_n = \mathbb{E}(T_n) = 2(n-2)/3$ y $\sigma_n^2 = \mathbb{V}(T_n) = (16n-29)/90$ y además

$\frac{T_n-\mu_n}{\sigma_n} \xrightarrow{d} N(0,1)$. La idea en la cual se basa el test, es que si T_n es significativamente superior a lo que sería su valor esperado en el caso en el que las variables sean i.i.d. entonces las observaciones fluctúan más que en el caso

i.i.d. (que es $2(n-2)/3$) mientras que si T_n es significativamente menor que $2(n-2)/3$, entonces hay una correlación positiva entre valores cercanos entre sí.

Por lo tanto se toma como región crítica asintótica

$$R = \left\{ \left| \frac{T_n - \mu_n}{\sigma_n} \right| > \phi^{-1}(1-\alpha/2) \right\}$$

siendo ϕ la función de distribución de una variable $N(0,1)$. En R, el valor de $\phi^{-1}(1-\alpha/2)$ se obtiene mediante qnorm$(1-\alpha)$.

1.2.5 Test de rangos

El test de rangos es particularmente útil para detectar dependencias lineales entre los datos.

Si $X_1, X_2, ..., X_n$ son observaciones (correspondientes a variables continuas de forma que $P(X_i = X_j) = 0$ para todo $i \neq j$), definimos P_n =cantidad de pares (i,j) tales que $i < j$ y además $X_i < X_j$. En total hay $\binom{n}{2} = \frac{n(n-1)}{2}$ pares (i,j) tales que $i < j$. Además si las variables fueran i.i.d., entonces $P(X_i < X_j) = \frac{1}{2}$, por lo tanto $\mu_n = \mathbb{E}(P_n) = \frac{n(n-1)}{4}$. Se puede probar además que en el caso i.i.d. la varianza es $\sigma_n^2 = \mathbb{V}(P_n) = n(n-1)(2n+5)/72$ y que $\frac{P_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1)$. En este caso, la idea en la cual está basado el test es de que si P_n es significativamente mayor (menor) que $\frac{n(n-1)}{4}$ (que sería el valor esperado si las observaciones fueran i.i.d.) entonces sería indicador de una tendencia creciente (decreciente) en el conjunto de observaciones.

Por lo tanto se plantea como región crítica asintótica para la prueba

$$R = \left\{ \left| \frac{P_n - \mu_n}{\sigma_n} \right| > \phi^{-1}(1-\alpha/2) \right\}.$$

1.2.6 Test del signo de la diferencia

Si $X_1, X_2, ..., X_n$ son observaciones, definimos S_n = cantidad de valores de i para los cuales $X_i - X_{i-1} > 0$ para $i = 2, 3, ..., n$. Se prueba en este caso que $\mu_n = \mathbb{E}(S_n) = \frac{n-1}{2}$, $\sigma_n^2 = \mathbb{V}(S_n) = \frac{n+1}{2}$ y que $\frac{S_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1)$. Se plantea la región crítica asintótica en la forma

$$R = \left\{ \left| \frac{P_n - \mu_n}{\sigma_n} \right| > \phi^{-1}(1-\alpha/2) \right\}.$$

Ahora sí, con todas las herramientas que hemos introducido hasta ahora en el capítulo anterior y éste, estamos en condiciones de definir los procesos ARMA y ver cómo modelar adecuadamente (o no) una serie observada de datos por dichos modelos.

1.3 Predicción

En esta sección veremos cómo se puede obtener una predicción lineal de un valor futuro válido para cualquier proceso estacionario en L^2 y a partir del mismo, utilizando las estimaciones de las componentes de tendencia y estacional, ver cómo realizar predicciones para una serie de tiempo cualesquiera.

1.3.1 Predicción lineal para un proceso estacionario

Consideramos $X_1, X_2, ..., X_n$ como las primeras n observaciones de un proceso estacionario y centrado $\{X_t\}_{t\in\mathbb{Z}}$ en L^2. Se puede hacer exactamente el mismo cálculo que haremos ahora si el proceso no fuera centrado.

Planteamos el problema de encontrar el mejor predictor del valor futuro X_{n+h} como una función lineal de los valores $X_1, X_2, ..., X_n$. Le llamaremos a dicho predictor $\widehat{X}_{n+h}$ y como será lineal de los valores pasados, tendrá la forma $\widehat{X}_{n+h} = a_1X_1 + a_2X_2 + ... + a_nX_n$. Queremos hallar los valores de $a_1, a_2, ..., a_n$ de modo que la esperanza $\mathbb{E}\left(\left(\widehat{X}_{n+h} - X_{n+h}\right)^2\right)$ sea mínima.

Para encontrarlo, planteamos

$$\mathbb{E}\left(\left(\widehat{X}_{n+h} - X_{n+h}\right)^2\right) =$$

$$\mathbb{E}\left(\left(a_1X_1 + a_2X_2 + ... + a_nX_n - X_{n+h}\right)^2\right) = \varphi\left(a_1, ..., a_n\right).$$

Para buscar el mínimo de φ hallamos las derivadas parciales e igualamos a cero, por lo que nos queda:

$$2\mathbb{E}\left(\left(a_1X_1 + a_2X_2 + ... + a_nX_n - X_{n+h}\right)X_i\right) = 0 \text{ para } i = 1, 2, ..., n.$$

O sea que, usando que el proceso es estacionario, por lo que todas las variables tienen esperanza $\mu = 0$, las ecuaciones quedan en la forma

$$a_1\mathbb{E}\left(X_1X_i\right) + a_2\mathbb{E}\left(X_2X_i\right) + ... + a_n\mathbb{E}\left(X_nX_i\right) - \mathbb{E}\left(X_{n+h}X_i\right) = 0$$

que equivale a

$$a_1\gamma\left(|1-i|\right)+a_2\gamma\left(|2-i|\right)+...+a_n\gamma\left(|n-i|\right) = \gamma\left(|n+h-i|\right) \text{ para } i = 1, 2, ..., n.$$

Estas n ecuaciones lineales se escriben matricialmente como

$$\Gamma_n a = \gamma_n\left(h\right)$$

siendo

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \Gamma_n = \begin{pmatrix} \gamma\left(0\right) & \gamma\left(1\right) & \cdots & \gamma\left(n-1\right) \\ \gamma\left(1\right) & \gamma\left(0\right) & \cdots & \gamma\left(n-2\right) \\ \vdots & & & \vdots \\ \gamma\left(n-1\right) & \gamma\left(n-2\right) & \cdots & \gamma\left(0\right) \end{pmatrix}$$

$$\text{y } \gamma_n(h) = \begin{pmatrix} \gamma(n+h-1) \\ \gamma(n+h-2) \\ \vdots \\ \gamma(h) \end{pmatrix}.$$

Si la matriz Γ_n es invertible, obtenemos la solución

$$\widehat{a} = \Gamma_n^{-1}\gamma_n(h).$$

Si no lo es, es posible demostrar que si existe solución es única, lo cual queda como ejercicio.

Se puede probar que si $\gamma(0) > 0$ y $\gamma(h) \to 0$ cuando $h \to +\infty$, entonces Γ_n es invertible para todo n.

1.3.2 Propiedades de $\widehat{X}_{n+h}$

Las siguientes propiedades valen en el caso en que el proceso no sea necesariamente centrado, con demostraciones análogas, adaptadas al caso no centrado.

1. $\mathbb{E}\left(\left(X_{n+h} - \widehat{X}_{n+h}\right)^2\right) = \gamma(0) - \widehat{a}^T\gamma_n(h)$.
2. $\mathbb{E}\left(X_{n+h} - \widehat{X}_{n+h}\right) = 0$.
3. $\mathbb{E}\left(\left(X_{n+h} - \widehat{X}_{n+h}\right)X_i\right) = 0$ para todo $i = 1, 2, ..., n$.
4. Si $\gamma_n(h) \to 0$ cuando $h \to +\infty$, entonces
   $\lim_{h\to+\infty} \mathbb{E}\left(\left(\widehat{X}_{n+h} - \mu\right)^2\right) = 0$ para todo n.

Observación 7 *La propiedad 2, nos dice que* $\mathbb{E}\left(\widehat{X}_{n+h}\right) = \mathbb{E}(X_{n+h})$.

Observación 8 *La propiedad 3 junto con la 2, nos dicen que no hay correlación entre el error de predicción y cualquiera de las variables predictoras.*

Observación 9 *La propiedad 4, implica que* $\widehat{X}_{t+h} \xrightarrow{P} \mu$ *cuando* $h \to +\infty$ *para todo* t. *Esta propiedad resulta intuitiva, ya que si tenemos las observaciones* $X_1, X_2, ..., X_t$ *y queremos predecir un valor futuro muy lejano (* X_{t+h} *con* h *grande), dado que las covarianzas son cercanas a cero entre* X_{t+h} *y cada una de las variables observadas* $X_1, X_2, ..., X_t$, *entonces la información que me dan las observaciones* $X_1, X_2, ..., X_t$ *son despreciables y estimaríamos el valor de* X_{t+h} *simplemente con el promedio del proceso.*

1.3.3 Algoritmos de Durbin–Levinson y de innovación

Si tenemos una serie de datos muy grande (n grande), la resolución del sistema $\Gamma_n a = \gamma_n(h)$, puede ser muy costoso computacionalmente. Los algortimos de Durbin–Levinson y de innovación son dos algoritmos iterativos para llegar a la solución del sistema. Los mismos utilizan $\hat{X}_{n+1}$ para obtener $\hat{X}_{n+2}$ y así sucesivamente. Se terminan obteniendo ecuaciones en recurrencias que son más rápidas de implementar computacionalmente.
Los detalles de estos dos algoritmos se pueden encontrar en el libro de Brockwell y Davis.
Los paquetes en R para predecir valores en un modelo ARMA utillizan estos métodos para el cálculo predictivo.

1.3.4 Predicción de valores futuros para procesos no necesariamente estacionarios

Dada una serie de observaciones correspondientes a una serie de tiempo, y supongamos que

estamos ante un modelo aditivo de la forma $X_t = T_t + E_t + I_t$ (como habíamos dicho en el capítulo 1, suponemos que no hay componente cíclica). En el capítulo anterior vimos cómo estimar la tendencia ($\widehat{T}_t$) por diversos métodos y la componente estacional ($\widehat{E}_t$). Luego las quitamos a los datos y obtenemos un nuevo conjunto de observaciones $Y_t = X_t - \widehat{T}_t - \widehat{E}_t$. A éstas observaciones les aplicamos los test de estacionariedad y si los pasan modelaríamos con un proceso estacionario como por ejemplo los ARMA. Como veremos en próximas secciones de este capítulo, los modelos ARMA permiten realizar predicciones de valores futuros, por ejemplo predecir Y_{t+h} para determinado valor de h, conociendo las observaciones $Y_1, Y_2, ..., Y_t$. A partir de esto, resulta muy fácil predecir el valor de X_{t+h} a partir de nuestra serie observada $X_1, X_2, ..., X_t$ ya que como sabemos que $Y_t = X_t - \widehat{T}_t - \widehat{E}_t$, entonces si $\widehat{Y}_{t+h}$ es la predicción de Y_{t+h}, predecimos el valor de X_{t+h} mediante la fórmula

$$\widehat{X}_{t+h} = \widehat{Y}_{t+h} + \widehat{T}_{t+h} + \widehat{E}_{t+h}.$$

Por ejemplo, supongamos que tenemos datos trimestrales de modo que la componente estacional se reduce a cuatro valores numéricos $\widehat{E}_1, \widehat{E}_2, \widehat{E}_3$ y $\widehat{E}_4$ y supongamos que la tendencia es lineal y la estimamos con la recta de regresión $T = \widehat{a}t + \widehat{b}$. Supongamos que tenemos datos desde el primer semestre de 2000 hasta el cuarto trimestre de 2019 ($X_1, X_2, ..., X_{80}$) y queremos predecir el primer trimestre de 2022. Queremos obtener una predicción del valor de X_{89}. Entonces desestacionalizamos los datos (con lo que obtenemos las $Y_t = X_t - \widehat{T}_t - \widehat{E}_t$), le ajustamos un modelo a las Y_t y luego calculamos la predicción de X_{89} mediante

$$\widehat{X}_{89} = \widehat{Y}_{89} + 89\widehat{a} + \widehat{b} + \widehat{E}_1.$$

Observamos que es más útil en estos casos estimar la tendencia mediante una recta u otro tipo de función y no aplicar promedios móviles dado que con promedios móviles, no podría estimar la tendencia en el instante $t = 89$.

2 Procesos ARMA. Definiciones y conceptos teóricos

Definición 10 *Procesos autorregresivos de promedio móvil: ARMA(p,q)*

Si $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso estocástico estacionario y centrado, se dice que es un proceso ARMA(p,q) si y sólo si existen constantes $\phi_1,\phi_2,...,\phi_p,\theta_1,\theta_2,...,\theta_q$ con $\phi_p\neq 0$ y $\theta_q\neq 0$, y un proceso $\{\varepsilon_t\}$ donde se verifica

$$X_t=\phi_1X_{t-1}+\phi_2X_{t-2}+...+\phi_pX_{t-p}+$$
$$\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+...+\theta_q\varepsilon_{t-q}+\varepsilon_t$$

donde $\{\varepsilon_t\}$ es ruido blanco y los polinomios ϕ y θ no tienen factores en común.

Definición 11 *Se dice que un proceso $\{X_t\}_{t\in\mathbb{Z}}$ es un ARMA(p,q) con media μ si y sólo si $\{X_t-\mu\}_{t\in\mathbb{Z}}$ es un ARMA(p,q).*

Observación 12 *La expresión anterior se escribe también como*

$$\phi(B)X_t=\theta(B)\varepsilon_t$$

siendo B es el operador shift, definido en el capítulo anterior, es decir, $BX_t=X_{t-1}$,

$$\phi(B)=1-\phi_1B-\phi_2B^2-...-\phi_pB^p,$$
$$\theta(B)=1+\theta_1B+\theta_2B^2+...+\theta_qB^q.$$

Los procesos autorregresivos y los de promedios móviles, son casos particulares de los procesos ARMA(p,q) y los definimos a continuación.

Definición 13 *Procesos autorregresivos de orden p : AR(p) .*

Si $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso estocástico estacionario y centrado, se dice que es un proceso AR(p) si y sólo si existen constantes $\phi_1,\phi_2,...,\phi_p$ con $\phi_p\neq 0$ y un proceso $\{\varepsilon_t\}$ donde se verifica

$$X_t=\phi_1X_{t-1}+\phi_2X_{t-2}+...+\phi_pX_{t-p}+\varepsilon_t$$

donde $\{\varepsilon_t\}$ es ruido blanco (variables no correlacionadas e idénticamente distribuídas con media cero).

La expresión anterior se escribe también como

$$\phi(B)X_t=\varepsilon_t.$$

Definición 14 *Procesos de medias móviles de orden q : MA(q)*

Si $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso estocástico estacionario y centrado, se dice que es un proceso MA(q) si y sólo si existen constantes $\theta_1,\theta_2,...,\theta_q$ con $\theta_q\neq 0$, y un proceso $\{\varepsilon_t\}$ donde se verifica

$$X_t=\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+...+\theta_q\varepsilon_{t-q}+\varepsilon_t$$

donde $\{\varepsilon_t\}$ es ruido blanco.

Observación 15 *La expresión anterior se escribe también como*

$$X_t=\theta(B)\varepsilon_t.$$

2.1 Modelos autorregresivos, AR(p)

La idea de estos modelos es muy intuitiva. Frecuentemente, una observación al instante t de una serie de tiempo, depende de las observaciones anteriores. Entonces se plantea un modelo en el cual la observación X_t dependa linealmente de las p observaciones anteriores: $X_{t-1}, X_{t-2}, ..., X_{t-p}$. Se le incorpora un ruido que engloba toda otra dependencia de X_t, sea con otras variables, o sea con las mismas en una dependencia de otro tipo de función no lineal.

Vamos a calcular la función autocovarianzas de un proceso AR(1) para el caso en el cual $|\phi| < 1$.

Como la igualdad que nos da la definición del proceso $AR(1)$

$$X_t = \phi X_{t-1} + \varepsilon_t$$

vale para todo t, la vamos aplicando recursivamente para $X_{t-1}, X_{t-2}, ...$ Entonces

$$X_t = \phi\left(\phi X_{t-2} + \varepsilon_{t-1}\right) + \varepsilon_t = \phi^2 X_{t-2} + \phi\varepsilon_{t-1} + \varepsilon_t =$$
$$\phi^2\left(\phi X_{t-3} + \varepsilon_{t-2}\right) + \phi\varepsilon_{t-1} + \varepsilon_t = \phi^3 X_{t-3} + \phi^2\varepsilon_{t-2} + \phi\varepsilon_{t-1} + \varepsilon_t = ... =$$
$$\phi^k X_{t-k} + \sum_{i=1}^{k-1} \phi^i \varepsilon_{t-i}.$$

Entonces, cualesquiera sean ϕ, k se tiene que

$$X_t = \phi^k X_{t-k} + \sum_{i=0}^{k-1} \phi^i \varepsilon_{t-i}.$$

Si hacemos $k \to +\infty$ y usando ahora condición $|\phi| < 1$ obtenemos que

$$X_t = \sum_{i=0}^{+\infty} \phi^i \varepsilon_{t-i}.$$

Por lo tanto, en el caso en el que $|\phi| < 1$, hemos probado que todo AR(1) se puede escribir como una suma infinita de variables del ruido blanco.

Observamos que esta serie es convergente (en probabilidad) porque $|\phi| < 1$.

Observamos que a partir de la última igualdad se deduce que en el caso en el que $|\phi| < 1$, el proceso es centrado, ya que

$$\mathbb{E}(X_t) = \mathbb{E}\left(\sum_{i=0}^{+\infty} \phi^i \varepsilon_{t-i}\right) = \sum_{i=0}^{+\infty} \phi^i \mathbb{E}\left(\varepsilon_{t-i}\right) = 0,$$

es decir que en la definición del proceso AR(p) se puede evitar la condición de ser un proceso centrado de la definición.

Ahora, a partir de esta igualdad, calculamos la función de autocovarianzas.

Si $h \in \mathbb{N}$, entonces

$$\gamma(h) = \mathbb{COV}\left(X_{t+h}, X_t\right) = \mathbb{E}\left(X_{t+h} X_t\right) = \mathbb{E}\left(\sum_{i=0}^{+\infty} \phi^i \varepsilon_{t+h-i} \sum_{j=0}^{+\infty} \phi^j \varepsilon_{t-j}\right) =$$

$$\sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty}\phi^i\phi^j\mathbb{E}\left(\varepsilon_{t+h-i}\varepsilon_{t-j}\right)=\sigma^2\sum_{i=h+1}^{+\infty}\phi^i\phi^{i-h}=\frac{\sigma^2}{\phi^h}\sum_{i=h}^{+\infty}\phi^{2i}=$$

$$\frac{\sigma^2}{\phi^h}\frac{\phi^{2h}}{1-\phi^2}=\frac{\sigma^2\phi^h}{1-\phi^2}.$$

En la primer igualdad de la línea anterior se usó que los ε_t son ruido blanco con varianza σ^2 y luego la fórmula de la geométrica:

$$\sum_{i=p}^{+\infty}x^i=\frac{x^p}{1-x}\text{válido para }|x|<1$$

para $x=\phi^2$.

De la fórmula $\gamma(h)=\frac{\sigma^2\phi^h}{1-\phi^2}$ deducimos que $\gamma(h)\to 0$ cuando $h\to+\infty$ con decrecimiento exponencial. También se deduce que cuando $0<\phi<1$ las covarianzas son todas positivas, mientras que si $-1<\phi<0$, entonces $\gamma(h)>0$ para h par y $\gamma(h)<0$ para h impar. La función de autocorrelación queda entonces

$$\rho(h)=\frac{\gamma(h)}{\gamma(0)}=\phi^p.$$

En la Figura 1 vemos las primeras 100 observaciones (de un total de 1.000) de la trayectoria de un proceso AR(1) con su función de autocorrelación para $\phi=0.9$ (arriba) y $\phi=-0.9$ (abajo). Se observa que la trayectoria del proceso AR(1) con $\phi=-0.9$ es mucho más irregular que el correspondiente a $\phi=0.9$. Esto no es casualidad, se explica por la función de autocorrelación. Como vimos, las covarianzas de un AR(1) con $\phi>0$ son siempre positivas, mientras que cuando $\phi<0$, es oscilante, se pasa de una correlación positiva a una negativa, lo cual hace que la trayectoria sea mucho más irregular.

Se puede probar que no existe un proceso AR(1) con $\phi=1$, porque de la igualdad $X_t=X_{t-1}+\varepsilon_t$, se puede verificar que la función de autocovarianzas $\mathbb{E}\left(X_{t+h}X_t\right)$ quedan en función de t, por lo que el proceso no es estacionario.

Observación 16 *Si un proceso AR(1) cumple que $|\phi|>1$, de la igualdad $X_t=\phi X_{t-1}+\varepsilon_t$ pasamos a la de $X_{t-1}=\frac{1}{\phi}\left(X_t-\varepsilon_t\right)$ que expresa el valor de X_{t-1} en función del siguiente valor futuro y si aplicamos reiteradamente ésta igualdad (o sea, escribo X_t en función de X_{t+1}, éste en función de X_{t+2}, etc), se llega a la igualdad $X_t=-\sum_{i=0}^{+\infty}\left(\frac{1}{\phi}\right)^i X_{t+i}$ por lo que expresamos el valor de X_t en función del futuro. Ahora esta serie es convergente dado que al ser $|\phi|>1$ se tiene que $1/|\phi|<1$.*

Cuando un proceso se puede expresar para cada instante en función únicamente de variables aleatorias medidas en todo su pasado, se dice que el proceso es causal. De esta forma un proceso AR(1) con $|\phi|<1$ es un proceso causal.

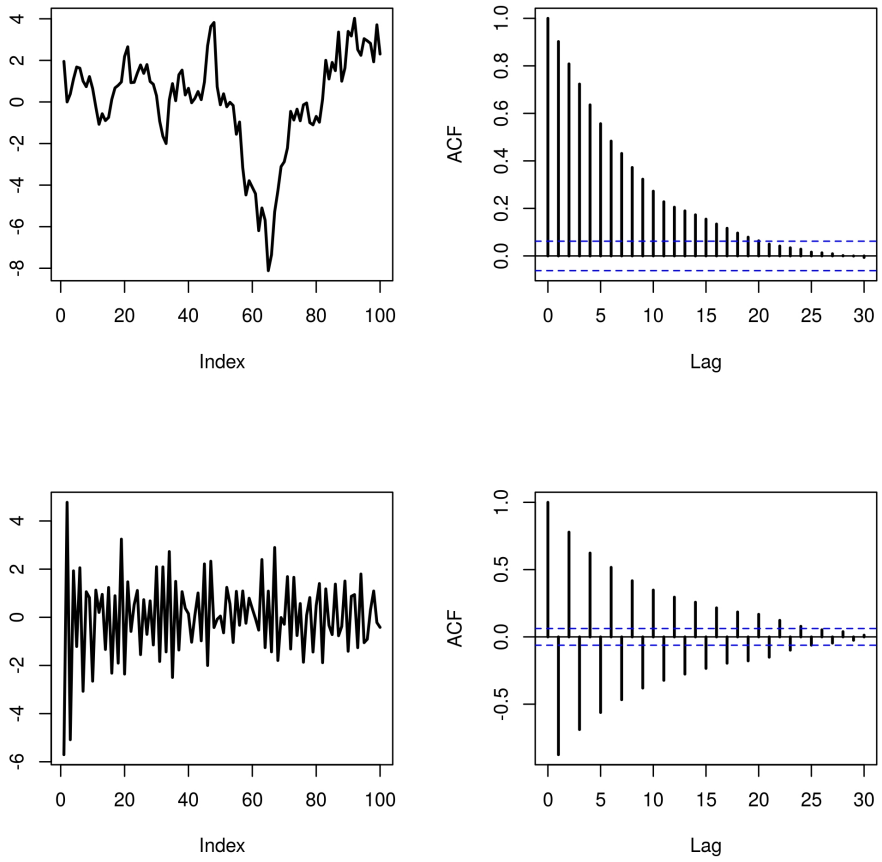

Figura 1: 100 primeras observaciones de un proceso AR(1) con $\phi = 0.9$(arriba izquierda) y su función de autocorrelación (arriba derecha). 100 primeras observaciones de un proceso AR(1) con $\phi = -0.9$ (abajo izquierda) y su función de autocorrelación (abajo derecha).

2.2 Procesos de medias móviles, MA(q)

Comenzamos calculando la función de autocovarianzas de un proceso MA(1) .

Si $X_t = \varepsilon_t + \theta\varepsilon_{t-1}$ entonces $\mathbb{E}(X_t) = 0$ por lo que el proceso es centrado. En cuanto a la función de autocovarianzas tenemos que si $h \in \mathbb{N}$, se tiene que

$$\gamma(h) = \mathbb{COV}(X_{t+h}, X_t) = \mathbb{E}(X_{t+h}X_t) = \mathbb{E}(\varepsilon_{t+h} + \theta\varepsilon_{t+h-1})(\varepsilon_t + \theta\varepsilon_{t-1}) =$$

$$\mathbb{E}(\varepsilon_{t+h}\varepsilon_t) + \theta\mathbb{E}(\varepsilon_{t+h-1}\varepsilon_t) + \theta\mathbb{E}(\varepsilon_{t+h}\varepsilon_{t-1}) + \theta^2\mathbb{E}(\varepsilon_{t+h-1}\varepsilon_{t-1}) =$$

$$\begin{cases} \sigma^2\left(1+\theta^2\right) & \text{si} \quad h=0 \\ \sigma^2\theta & \text{si} \quad |h|=1 \\ 0 & \text{si} \quad |h|>1 \end{cases}.$$

La principal observación es que la función de autocovarianza es nula cuando dos observaciones distan al menos dos unidades de tiempo. Queda dentro de los ejercicios el probar que la función de autocovarianza de un MA(q) se anula cuando $h > q$.

La función de autocorrelación queda entonces

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} 1 & \text{si} \quad h=0 \\ \frac{\theta}{1+\theta^2} & \text{si} \quad |h|=1 \\ 0 & \text{si} \quad |h|>1 \end{cases}.$$

Observación 17 *La función de autocorrelación de un proceso $MA(1)$ con parámetro θ coincide con la función de autocorrelación de un proceso $MA(1)$ con parámetro $1/\theta$.*

Observación 18 $|\rho(1)| \leq 1/2$ *(porque* $0 \leq (1-|\theta|)^2 = 1 - 2|\theta| + \theta^2$ *entonces* $2|\theta| \leq 1+\theta^2$*) por la función ρ de un proceso $MA(1)$ arranca en $\rho(0)=1$, baja abruptamente a un número menor que $1/2$ ($\rho(1)$) da cero a partir de $\rho(2)$ en adelante.*

Observación 19 *Si $X_t = \varepsilon_t + \theta\varepsilon_{t-1}$ entonces $\varepsilon_t = -\theta\varepsilon_{t-1} + X_t$ por lo que podemos pensar en que el ruido blanco es como tener un $AR(1)$ de parámetro $-\theta$ donde $\{X_t\}$ jugaría el papel del ruido blanco. Si $|\theta| < 1$, haciendo la misma cuenta que la realizada en la descomposición del $AR(1)$ como suma infinita de los valores de los ruidos en los instantes anteriores, llegamos a que*

$$\varepsilon_t = \sum_{i=0}^{+\infty} (-\theta)^i X_{t-i}.$$

En estos casos es cuando decimos que el proceso es invertible, es decir que en lugar de poder escribir el proceso $\{X_t\}$ como suma infinita en función de los valores de los ε_t en instantes anteriores a t, escribimos ε_t como suma infinita en función de los valores de X_t en instantes anteriores a t.

Observación 20 *Si $\{X_t\}_{t\in\mathbb{Z}}$ es un $MA(1)$ y los $\varepsilon_t \sim N\left(0,\sigma^2\right)$, entonces el proceso $\{X_t\}_{t\in\mathbb{Z}}$ es Gaussiano (porque la distribución conjunta de $(X_{t_1}, X_{t_2}, ..., X_{t_k})$ es normal multivariada, ya que el vector puede ser escrito como una matriz por una cantidad finita de los ε_t).*

2.3 Causalidad

Definición 21 *Causalidad*

Si $\{X_t\}_{t\in\mathbb{Z}}$ es un $ARMA(p,q)$ se dice que es causal si y sólo si el proceso admite una representación en la forma

$$X_t = \sum_{i=0}^{+\infty} \psi_i \varepsilon_{t-i} \text{ donde } \sum_{i=0}^{+\infty} |\psi_i| < +\infty,\ \psi_0 = 1.$$

El siguiente teorema nos da una condición necesaria y suficiente para ver si un proceso ARMA(p,q) es causal o no y nos indica (en caso de ser causal) cómo hallar los coeficientes ψ_i.

Recordemos que el proceso ARMA lo escribíamos en forma compacta como

$$\phi(B)X_t = \theta(B)\varepsilon_t$$

siendo B es el operador shift, definido en el capítulo anterior, es decir, $BX_t = X_{t-1}$,

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p,$$
$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q.$$

Teorema 22 *Si $\{X_t\}_{t\in\mathbb{Z}}$ es un ARMA(p,q). Entonces $\{X_t\}_{t\in\mathbb{Z}}$ es causal si y sólo si $\phi(z) \neq 0$ para todo z complejo, tal que $|z| \leq 1$. Además, en el caso de ser causal, se tiene que los coeficientes ψ_i se obtienen de la siguiente igualdad:*

$$\sum_{i=0}^{+\infty} \psi_i z^i = \frac{\theta(z)}{\phi(z)} \text{ para todo } |z| \leq 1.$$

Observación 23 *El teorema anterior nos dice que los coeficientes ψ_i se obtienen de realizar el desarrollo de Taylor alrededor de 0 de la función $\frac{\theta(z)}{\phi(z)}$. También se pueden obtener igualando coeficientes en la igualdad $\phi(z)\sum_{i=0}^{+\infty}\psi_i z^i = \theta(z)$, es decir de igualar*

$$\left(1 - \phi_1 z - \phi_2 z^2 - ... - \phi_p z^p\right)\left(1 + \psi_1 z + \psi_2 z^2 + ...\right) = 1 + \theta_1 z + \theta_2 z^2 + ... + \theta_q z^q.$$

Observación 24 *La condición $\phi(z) \neq 0$ para todo z complejo, tal que $|z| \leq 1$, significa que las raíces del polinomio ϕ están todas en la región $|z| > 1$.*

Ejemplo 25 *Un modelo AR(1) de la forma $X_t = \phi X_{t-1} + \varepsilon_t$ es causal si y sólo si $|\phi| < 1$ ya que el polinomio $\phi(z) = 1 - \phi z = 0$ si y sólo si $z = 1/\phi$ siendo $|1/\phi| > 1$. Por ejemplo $X_t = -0.8X_{t-1} + \varepsilon_t$ es causal, pero $X_t = 3X_{t-1} + \varepsilon_t$ no es causal.*

Ejemplo 26 *El modelo ARMA$(2,1)$ dado por $X_t = \frac{5}{6}X_{t-1} + \frac{1}{6}X_{t-2} + 5\varepsilon_{t-1} + \varepsilon_t$ es causal porque $\phi(z) = 1 - \frac{5}{6}z + \frac{1}{6}z^2 = 0$ si y sólo si $z = 2$, $z = 3$ ambas raíces tienen módulo mayor que uno.*

Ejemplo 27 *El modelo AR(2) dado por $X_t = \frac{1}{2}X_{t-1} + \frac{1}{2}X_{t-2} + \varepsilon_t$ no es causal porque tiene una raíz unitaria (significa con módulo 1) porque $\phi(z) = 1 - \frac{1}{2}z - \frac{1}{2}z^2 = 0$ si y sólo si $z = -2$, $z = 1$.*

2.4 Invertibilidad

El concepto de proceso ARMA invertible es análogo al de causalidad pero intercambiando los papeles entre los X_t y los ε_t.

Definición 28 *Invertibilidad*

Si $\{X_t\}_{t\in\mathbb{Z}}$ *es un ARMA*(p,q) *se dice que es invertible si y sólo si es posible escribir una representación en la forma*

$$\varepsilon_t = \sum_{i=0}^{+\infty} \pi_i X_{t-i} \text{ donde } \sum_{i=0}^{+\infty} |\pi_i| < +\infty,\ \pi_0 = 1.$$

Teorema 29 *Si* $\{X_t\}_{t\in\mathbb{Z}}$ *es un ARMA*(p,q)*. Entonces* $\{X_t\}_{t\in\mathbb{Z}}$ *es invertible si y sólo si* $\theta(z) \neq 0$ *para todo* z *complejo, tal que* $|z| \leq 1$*. Además, en el caso de ser invertible, se tiene que los coeficientes* π_i *se obtienen de la siguiente igualdad:*

$$\sum_{i=0}^{+\infty} \pi_i z^i = \frac{\phi(z)}{\theta(z)} \text{ para todo } |z| \leq 1.$$

Este teorema no requiere demostración dado que la misma surge de intervambiar los papeles de X_t con los de ε_t en el teorema anĺogo sobre causalidad.

Ejemplo 30 *El modelo ARMA*$(2,2)$ *dado por* $X_t = \frac{5}{6}X_{t-1} + \frac{1}{6}X_{t-2} + \frac{1}{4}\varepsilon_{t-2} + \varepsilon_t$ *es causal porque* $\phi(z) = 1 - \frac{5}{6}z + \frac{1}{6}z^2 = 0$ *si y sólo si* $z = 2$, $z = 3$ *ambas raíces tienen módulo mayor que uno. También es invertible porque* $\theta(z) = \frac{1}{4}z^2 + 1 = 0$ *si y sólo si* $z = \pm 2i$ *y ambas raíces tienen módulo* $2 > 1$.

Observación 31 *Vale la pena notar que las propiedades de causalidad e invertibilidad no son únicamente del proceso* $\{X_t\}_{t\in\mathbb{Z}}$ *sino de la relación entre el proceso* $\{X_t\}_{t\in\mathbb{Z}}$ *con el del ruido blanco* $\{\varepsilon_t\}_{t\in\mathbb{Z}}$*. Un proceso ARMA podría no ser causal o no invertible con respecto a cierto ruido blanco* $\{\varepsilon_t\}_{t\in\mathbb{Z}}$*, pero sí podría ser causal e invertible respecto a otro ruido blanco* $\{\varepsilon'_t\}_{t\in\mathbb{Z}}$*. Esto vendrá justificado más formalmente en el teorema siguiente.*

Teorema 32 *Si* $\{X_t\}_{t\in\mathbb{Z}}$ *es un proceso ARMA*(p,q) *tal que* $\phi(B)X_t = \theta(B)\varepsilon_t$ *donde* $\theta(z) \neq 0$ *para* $|z| = 1$*, entonces existen polinomios* ϕ' *y* θ' *y un ruido blanco* ε' *tales que* $\phi'(B)X_t = \theta'(B)\varepsilon'_t$ *tales que* $\phi'(z) \neq 0$ *para todo* $|z| \leq 1$ *y* $\theta'(z) \neq 0$ *para todo* $|z| \leq 1$.

El teorema anterior, dice que si la condición $\theta(z) \neq 0$ para $|z| = 1$ se cumple en un proceso ARMA con determinado ruido blanco, entonces el mismo proceso es un proceso ARMA causal e invertible con otro ruido blanco. Como en un proceso ARMA, las variables que integran el ruido blanco no son variables observables, la falta de causalidad o de invertibilidad no son problema a la hora de modelar una serie de tiempo.

2.5 Raíces unitarias en series de tiempo

Cuando en un proceso ARMA alguno de los dos polinomios $\phi(z)$ o $\theta(z)$ tienen raíces en el círculo unidad (o cerca del círculo unidad) $|z| = 1$ se dice que el proceso tiene una raíz unitaria. Se puede probar que la existencia de una raíz

unitaria en una igualdad del tipo $\phi(B)X_t = \theta(B)\varepsilon_t$, hace que el proceso no sea estacionario (como comentamos en el caso del AR(1) con $|\phi| = 1$). Más explícitamente, se puede probar que una raíz cerca del círculo unidad en el polinomio autorregresivo ($\phi(z)$) indica que los datos no son estacionarios y que conviene desestacionalizarlos (quizá mediante el operador diferenciación cierta cantidad de veces) mientras que una raíz unitaria en el polinomio de los promedios móviles, puede indicar que el proceso está sobrediferenciado. El test de Dickey–Fuller por ejemplo está basado en esta idea, por eso en el capítulo 1, en el test de hipótesis se habla de la existencia o no de raíces unitarias. Como fue dicho en su momento, el test es pobre en el sentido de que sirve para modelar mediante procesos ARMA, pero quizá la serie de datos que uno disponga podría no ser bien modelada por un proceso ARMA y en esos casos el test podría inducirnos en error. De todas formas, como ya se dijo antes, existen pocas herramientas para probar mediante un test de hipótesis si una serie de datos es estacionaria o no y en la práctica este tipo de test funcionan bastante bien. En el libro de Brockell y Davies se puede encontrar una explicación más profunda de todo esto.

2.6 Función de autocovarianza de un proceso ARMA(p, q) causal

La representación de un proceso causal ARMA(p, q) como suma infinita de las variables que integran el ruido blanco nos permite obtener una fórmula relativamente sencilla para hallar las autocovarianzas del proceso.

$$\gamma(h) = \mathbb{COV}(X_{t+h}, X_t) = \mathbb{E}(X_{t+h}X_t) = \mathbb{E}\left(\sum_{i=0}^{+\infty}\psi_i\varepsilon_{t+h-i}\ \sum_{j=0}^{+\infty}\psi_j\varepsilon_{t-j}\right) =$$

$$\sum_{i=0}^{+\infty}\psi_i\sum_{j=0}^{+\infty}\psi_j\mathbb{E}(\varepsilon_{t+h-i}\ \varepsilon_{t-j}\ ) = \sigma^2\sum_{i=h}^{+\infty}\psi_i\psi_{i-h}.$$

2.7 Función de autocorrelación parcial (PACF)

Definición 33 *Función de autocorrelación parcial: PACF*

Si $\{X_t\}_{t\in\mathbb{Z}}$ *es un proceso estacionario en* L^2*, definimos para cada* $h = 1, 2, 3, ...$

$$\begin{aligned}\phi_{11} &= \rho(X_{t+1}, X_t) = \rho(1)\\ \phi_{hh} &= \rho\left(X_{t+h} - \widehat{X}_{t+h}, X_t - \widehat{X}_t\right) \text{ para } h \geq 2\end{aligned}$$

siendo

$$\widehat{X}_t = \beta_1 X_{t+1} + \beta_2 X_{t+2} + ... + \beta_{h-1}X_{t+h-1}$$

y

$$\widehat{X}_{t+h} = \beta_1 X_{t+h-1} + \beta_2 X_{t+h-2} + ... + \beta_{h-1} X_{t+1}$$

las regresiones de X_t *y* X_{t+h} *sobre* $\{X_{t+h-1}, X_{t+h-2}, ..., X_{t+1}\}$ *respectivamente. a la que le llamamos función de autocorrelación parcial del proceso.*

Observación 34 *Las regresiones* $\widehat{X}_t = \beta_1 X_{t+1} + \beta_2 X_{t+2} + ... + \beta_{h-1} X_{t+h-1}$ *y* $\widehat{X}_{t+h} = \beta_1 X_{t+h-1} + \beta_2 X_{t+h-2} + ... + \beta_{h-1} X_{t+1}$ *sobre* $\{X_{t+h-1}, X_{t+h-2}, ..., X_{t+1}\}$*, significan que se deben hallar los parámetros* $\beta_1, \beta_2, ..., \beta_{h-1}$ *de modo de minimizar* $\mathbb{E}\left(X_{t+h} - \widehat{X}_{t+h}\right)^2$ *y* $\mathbb{E}\left(X_t - \widehat{X}_t\right)^2$ *respectivamente. Observar que* $\widehat{X}_{t+h}$ *es la mejor predicción cuadrática del valor* X_{t+h} *a partir del conjunto de observaciones que se encuentran entre* X_{t+1} *y* X_{t+h-1} *(es decir conociendo las* $h-1$ *observaciones pasadas) mientras que* $\widehat{X}_t$ *es la mejor predicción cuadrática del valor* X_t *conociendo las mismas observaciones, es decir las* $h-1$ *observaciones posteriores a* X_t*.*

Observación 35 *La función de autocorrelación parcial, mide la autocorrelación entre el error cometido al estimar (predecir)* X_{t+h} *a partir de* $\widehat{X}_{t+h}$ *con la de estimar* X_t *a partir de* $\widehat{X}_t$ *cuando se conocen el conjunto de observaciones entre* X_t *y* X_{t+h}*.*

Se prueba que cuando $h > p$, entonces, utilizando la causalidad del proceso se deduce que $\widehat{X}_{t+h} = \phi_1 X_{t+h-1} + \phi_2 X_{t+h-2} + ... + \phi_p X_{t+h-p}$ y $\widehat{X}_t = \phi_1 X_{t+1} + \phi_2 X_{t+2} + ... + \phi_p X_{t+h-1}$ y por lo tanto,

$$\phi_{hh} = \rho\left(X_{t+h} - \widehat{X}_{t+h}, X_t - \widehat{X}_t\right) = \rho\left(\varepsilon_{t+h}, \varepsilon_t\right) = 0 \text{ para } h \geq p+1.$$

Observación 36 *En el caso de los procesos AR(p) causales,* $\phi_{11}, \phi_{22}, ..., \phi_{pp}$ *no tienen por qué ser* 0 *y además se puede probar que* $\phi_{pp} = \phi_p$*.*

Con argumentos similares a los del caso AR(p) , se puede probar que la PACF de un proceso MA(q) invertible nunca se anula.

Conclusión. La ACF de un AR(p) causal nunca se anula mientras que la de un MA(q) se anula a partir de $h \geq q+1$, mientras que en el sentido inverso, la PACF de un AR(p) se anula a partir de $h \geq p+1$, mientras que la de un MA(q) es decreciente pero nunca se anula.

2.7.1 Aplicación

Una posible aplicación puede ser la siguiente. Hemos visto en los ejemplos y observaciones anteriores, que la PACF de un proceso AR(p) es nula para $h \geq p+1$ y es no nula en $h = p$. Entonces, si las PACF de los datos muestrales son significativamente distintas de cero para valores de $h \leq p$ y cercanas a cero para $h \geq p+1$, para determinado valor de p, entonces puede ser razonable modelar a partir de un proceso AR(p) .

Se puede demostrar que para procesos AR(p) causales, que cuando n es suficientemente grande, $\widehat{\phi}_{hh}$ (estimación empírica de la PACF en h a partir de

la serie de datos observada) tiene una distribución aproximadamente $N(0, 1/n)$ y son aproximadamente independientes para valores de $h \geq p+1$, se puede utilizar como regla empírica la siguiente condición:

Si $\left|\widehat{\phi}_{hh}\right| \leq 1.96/\sqrt{n}$ para valores de $h \geq p+1$ y $\left|\widehat{\phi}_{hh}\right| > 1.96/\sqrt{n}$ para valores de $h \leq p$, entonces puede ser razonable ajustar la serie de datos por un modelo AR(p) . Con el mismo tipo de idea se llega a una regla similar utilizando las ACF para obtener cierta razonabilidad de modelar mediante un modelo MA(q). Es decir, si $|\widehat{\rho}(h)| \leq 1.96/\sqrt{n}$ para valores de $h \geq q+1$ y $|\widehat{\rho}(h)| > 1.96/\sqrt{n}$ para valores de $h \leq q$, entonces puede ser razonable ajustar la serie de datos por un modelo MA(q) .

En la Figura 2 se grafican las ACF y PACF para dos series de 1.000 observaciones X e Y. Para la serie X, se ve que la ACF siempre es positiva, mientras que la PACF tiene valores por fuera de las bandas de confianza hasta $h = 4$ y a partir de $h = 5$ no podemos rechazar la hipótesis de que el valor sea 0, por lo que ambos gráficos sugieren un modelo AR(4) para X. En cuanto a la serie Y, con un argumento similar, podemos decir que ambos gráficos sugieren modelar mediante un MA(1) (a pesar de que hay un rezago aislado en $h = 9$) que está por fuera de la barrera de confianza.

En R la función de autocorrelación parcial viene dado por la función pacf (pacf(x,lag.max=k)).

Nota. Si bien este criterio puede ser utilizado, en general es muy difícil distinguir entre la función de autocovarianzas de un ARMA y las de un AR o un MA, por lo tanto para decidir mejor cómo ajustar nuestra serie de datos, es más conveniente definir un criterio de medida de ajuste del modelo a los datos y optimizarlo respecto a los valores de p y q. Esto será desarrollado en la sección 5.

3 Predicción en un proceso ARMA

Como es intuitivo, se puede probar que en un proceso AR(p) : $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \varepsilon_t$, se cumple que $\widehat{X}_{t+1} = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p}$. En la práctica, como no se conocen los valores de $\phi_1, \phi_2, ..., \phi_p$, se los sustituye por sus estimaciones $\hat{\phi}_1, \hat{\phi}_2, ..., \hat{\phi}_p$.

En general, para un proceso MA(q) o ARMA(p, q) no existen fórmulas tan sencillas para el cálculo de las predicciones de valores futuros como para el caso AR(p), pero sí existen fórmulas que son recursivas basadas en los algoritmos de innovación y de Durbin–Watson, y son las que aplican los distintos paquetes estadísticos para realizar dichos cálculos.

3.1 Intervalos de predicción

Si $\{X_t\}_{t\in\mathbb{Z}}$ es un ARMA(p, q) causal donde el ruido blanco es Gaussiano, entonces el proceso queda Gaussiano, y como toda combinación lineal de variables de un proceso Gaussiano es normal con esperanza ser (por la propiedad 2 de $\hat{X}_{t+h}$, entonces $X_{t+h} - \hat{X}_{t+h} \sim N(0, \sigma_n^2(h))$. Utilizando la causalidad se puede

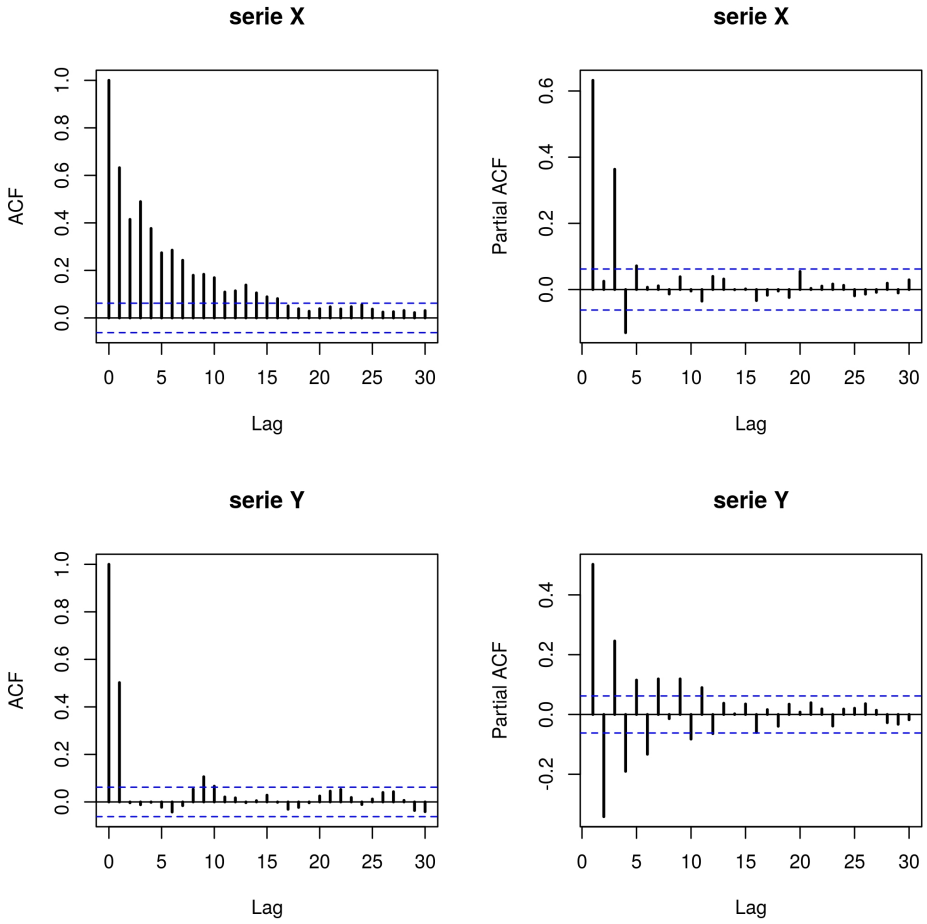

Figura 2: ACF y PACF para dos series de 1000 datos X (arriba) e Y (abajo).

calcular y obtener $\sigma_n^2(h) = \sigma^2 \sum_{i=0}^{h-1} \psi_i^2$ siendo σ^2 la varianza del ruido blanco y los coeficientes ψ_i surgen del Teorema 22. Por lo tanto el intervalo de predicción de nivel $1-\alpha$ para el valor X_{t+h} a partir de $X_1, X_2, ..., X_t$ es

$$\left(\hat{X}_{t+h} - \phi^{-1}(1-\alpha/2)\sigma_n(h), \hat{X}_{t+h} + \phi^{-1}(1-\alpha/2)\sigma_n(h)\right).$$

4 Estimación de parámetros

4.1 Estimación de parámetros por máxima verosimilitud

Dadas $X_1, X_2, ..., X_n$ observaciones correspondientes a un proceso $\{X_t\}_{t\in\mathbb{Z}}$ ARMA(p,q) donde los valores de p y q son conocidos se realiza por máxima verosimilitud.

Proposition 37 *Si $X_1, X_2, ..., X_n$ son observaciones correspondientes a un proceso $\{X_t\}_{t\in\mathbb{Z}}$ causal ARMA(p, q) donde el ruido blanco es Gaussiano (es decir que ε_t son i.i.d. $N\left(0,\sigma^2\right)$) y le llamamos $\phi = (\phi_1, \phi_2, ..., \phi_p)$ y $\theta = (\theta_1, \theta_2, ..., \theta_q)$ a los parámetros del modelo, entonces la función de verosimilitud correspondiente a las n observaciones, viene dada por*

$$L\left(\phi,\theta,\sigma^2\right) = \frac{1}{\sqrt{\left(2\pi\sigma^2\right)^n r_0 r_1 ... r_{n-1}}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i-\widehat{X}_i\right)^2/r_{i-1}}$$

siendo $\widehat{X}_i$ el predictor lineal óptimo definido a partir de $X_1, X_2, ..., X_{i-1}$ ($\widehat{X}_1 = 0$) y $r_{i-1} = \frac{1}{\sigma^2}\mathbb{E}\left(\left(X_i - \widehat{X}_i\right)^2\right)$.

La demostración de la proposición se basa en que al ser el proceso causal, se puede escribir como combinación lineal infinita de los ruidos blancos en los instantes pasados: $X_t = \sum_{i=0}^{+\infty}\psi_i\varepsilon_{t-i}$, serie que es convergente en probabilidad. Dado que los ruidos blancos son gaussianos, se tiene que el proceso $\{X_t\}_{t\in\mathbb{Z}}$ es Gaussiano. Entonces si le llamamos $X = (X_1, X_2, ..., X_n)$ al vector de observaciones, se tendrá que la densidad del mismo es de la forma

$$f_X(x) = \frac{1}{\sqrt{\left(2\pi\sigma^2\right)^n \det(\Gamma)}} e^{\frac{-1}{2}X^T\Gamma^{-1}X}$$

finalmente se puede ver (esto da trabajo) que Γ puede ser diagonalizada de modo que $X^T\Gamma^{-1}X = \frac{1}{2}\sum_{i=1}^{n}\left(X_i - \widehat{X}_i\right)^2/r_{i-1}$ y que $\det(\Gamma) = r_0 r_1 ... r_{n-1}$.

Para hallar los estimadores de los parámetros, consideramos el logaritmo de

$$\ln\left(L\left(\phi,\theta,\sigma^2\right)\right) =$$

$$-\ln\left((2\pi)^{n/2}\right) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2}\sum_{i=1}^{n}\ln\left(r_{i-1}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i - \widehat{X}_i\right)^2/r_{i-1}$$

y derivamos respecto a los parámetros. Tener en cuenta que tanto los r_i como los $\widehat{X}_i$ dependen de ϕ y θ (se puede probar que no dependen de σ^2).

La optimización se realiza de forma numérica dado que no se pueden obtener fórmulas explícitas que resuelvan. Si derivamos respecto a σ^2 e igualamos a cero, resulta la igualdad $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{X}_i(\phi,\theta)\right)^2/r_{i-1}(\phi,\theta)$ por lo tanto, una vez obtenidos $\widehat{\phi}$ y $\widehat{\theta}$, el estimador de σ^2

es

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{X}_i(\phi,\theta)\right)^2/r_{i-1}\left(\widehat{\phi},\widehat{\theta}\right).$$

Se puede probar además que $\widehat{\phi}$ y $\widehat{\theta}$ se obtienen minimizando la función $l(\phi,\theta) = \frac{1}{n}\sum_{i=1}^{n}\ln\left(r_{i-1}(\phi,\theta)\right) + \ln\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{X}_i(\phi,\theta)\right)^2/r_{i-1}(\phi,\theta)\right)$.

4.2 Estimación de parámetros por mínimos cuadrados

El método de máxima verosimilitud supone que el ruido blanco es Gaussiano, pero podríamos tener un proceso ARMA(p,q) donde el ruido blanco no sea Gaussiano. En ese caso, podemos considerar un estimador mínimo cuadrático definiendo $\widehat{\phi}$ y $\widehat{\theta}$ tales que minimices la función

$$S(\phi,\theta)=\sum_{i=1}^{n}\left(X_i-\widehat{X}_i(\phi,\theta)\right)^2/r_{i-1}(\phi,\theta)$$

. Es un estimador bastante razonable e intuitivo, dado que es natural pensar en encontrar aquellos parámetros que minimicen la suma de cuadrados entre todas las posibles predicciones. El conciente entre r_{i-1} simplemente estandariza a cada uno de sus sumandos.

Claramente este método no necesita conocer la distribución del proceso. Para estimar σ^2 aplicamos la fórmula obtenida anteriormente, es decir que

$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\widehat{X}_i\left(\widehat{\phi},\widehat{\theta}\right)\right)^2/r_{i-1}\left(\widehat{\phi},\widehat{\theta}\right)=\frac{1}{n}S\left(\widehat{\phi},\widehat{\theta}\right).$$

4.3 Consistencia asintótica de los estimadores máximo verosímiles

Teorema 38 *Si $X_1,X_2,...,X_n$ son observaciones correspondientes a un proceso $\{X_t\}_{t\in\mathbb{Z}}$ causal e invertible ARMA(p,q) y le llamamos $\beta=(\phi_1,\phi_2,...,\phi_p,\theta_1,\theta_2,...,\theta_q)$ al vector de parámetros y $\widehat{\beta}=\left(\widehat{\phi}_1,\widehat{\phi}_2,...,\widehat{\phi}_p,\widehat{\theta}_1,\widehat{\theta}_2,...,\widehat{\theta}_q\right)$. Entonces, cuando n tiende a infinito, la distribución del vector $\widehat{\beta}$ es aproximadamente $N_{p+q}\left(0,\frac{1}{n}V(\beta)\right)$ en donde la matriz $V(\beta)$ puede ser aproximada por $2H^{-1}(\beta)$ siendo $H(\beta)$ la matriz Hessiana de la función l, es decir que en el lugar i,j está definida como $\frac{\partial^2 l(\beta)}{\partial\beta_i\partial\beta_j}$ siendo $l(\phi,\theta)=\frac{1}{n}\sum_{i=1}^{n}\ln(r_{i-1})+\ln\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\widehat{X}_i\right)^2/r_{i-1}\right)$.*

Observación 39 *En la práctica β no es conocido pero si n es grande, se aproxima $H(\beta)$ por $H\left(\widehat{\beta}\right)$.*

Observación 40 *El teorema, es válido cualquiera sea el ruido blanco, es decir no tiene por qué ser Gaussiano, y en ese caso, se le está llamando estimadores máximo verosímiles a los obtenidos mediante las fórmulas para el caso Gaussiano.*

5 Elección de los valores de p y q

Un punto crucial es el de la elección de los valores de p y q para luego ajustarles un modelo ARMA(p,q). Recordar que cuando $p=0$ obtenemos los MA(q) y cuando $q=0$ obtenemos los AR(p). Por supuesto muchas veces es conveniente ajustarle a una serie de datos luego de desestacionalizada y centrada modelos

ARMA(p,q) para distintos valores de p y q, luego comparar distintas medidas, por ejemplo las medidas de diagnóstico y quedarse con la mejor elección en base a algún criterio.

Es bueno tener siempre presente el principio de parsimonia, que básicamente sugiere trabajar con la menor cantidad de parámetros como sea posible, ya que tener demasiados parámetros en el modelo suele traer el problema del sobreajuste del modelo a los datos observados además de varianzas grandes en los estimadores lo cual no es deseable.

De todas formas, a continuación mencionamos los criterios mayormente utilizados para la selección de los valores de p y q a ajustar.

5.1 Criterio de información de Akaike (AIC)

Supongamos que las observaciones $X_1, X_2, ..., X_n$ son observaciones de un proceso ARMA(p,q) Gaussiano (lo cual quiere decir que el ruido blanco es Gaussiano lo que lleva a que el proceso sea Gaussiano). Le llamamos $\left(\widehat{\phi}, \widehat{\theta}, \widehat{\sigma}^2\right)$ al estimador máximo verosímil de los parámetros vista en la sección de estimación de parámetros.

El principio de máxima verosimilitud está basado en elegir los valores de ϕ, θ y σ^2 que hagan la verosimilitud de la muestra esperada. El vector de parámetros $\left(\widehat{\phi}, \widehat{\theta}, \widehat{\sigma}^2\right)$ es $p+q+1-$dimensional en el caso de un modelo ARMA(p,q), pero si ajustamos un ARMA(p',q') con más parámetros, la verosimilitud aumentará (porque habrá más cantidad de parámetros para ajustar). Una manera de contrarrestar el efecto de que la verosimilitud sea creciente al ir agregando parámetros es agregarle. Eso es lo que hace el criterio de minimizar el AIC.

Definición 41 *Dado un proceso ARMA(p,q) con p y q fijos, definimos*

$$AIC\,(p,q) = -2\ln\left(L\left(\widehat{\phi}, \widehat{\theta}, \widehat{\sigma}^2\right)\right) + 2(p+q+1)$$

donde $\left(\widehat{\phi}, \widehat{\theta}, \widehat{\sigma}^2\right)$ es el estimador máximo verosímil de $\left(\phi, \theta, \sigma^2\right)$.

Criterio de minimización del AIC.
Ajusto un modelo ARMA$(\widehat{p}, \widehat{q})$ siendo

$$(\widehat{p}, \widehat{q}) = \arg\min AIC\,(p,q)\,.$$

Es decir que el criterio dice de calcular todos los $AIC\,(p,q)$ posibles (variando p y q) y elegir aquel par de valores p y q donde se minimice $AIC\,(p,q)$.

Observación 42 *Minimizar $AIC\,(p,q)$ equivale a maximizar $2\ln\left(L\left(\widehat{\phi}, \widehat{\theta}, \widehat{\sigma}^2\right)\right) - 2(p+q+1)$ que es una especie de maximización por máxima verosimilitud pero "'penalizada" por el factor, dado que si agregamos parámetros la expresión $2\ln\left(L\left(\widehat{\phi}, \widehat{\theta}, \widehat{\sigma}^2\right)\right)$ crece mientras que la expresión que penaliza $2\,(p+q+1)$ también crece. Dicho de otra forma, cuanto más parámetros incluyamos en el modelo, más crece la verosimilitud a la vez que más penalizamos.*

¿Por qué penalizar con la expresión $2(p+q+1)$ y no otra? Por supuesto, se puede ensayar infinitas funciones de penalización distintas a esta, pero se puede probar que $-2\ln\left(L\left(\widehat{\phi},\widehat{\theta},\widehat{\sigma}^2\right)\right)+\frac{2(p+q+1)n}{n-p-q-2}$ es un estimador insesgado del valor esperado del índice de discrepancia entre modelos de Kullback–Leibler que se define de la siguiente forma: Si $X=(X_1,X_2,...,X_n)$ es un vector aleatorio $n-$dimensional y su función de densidad es $f_X(x,\theta)$ donde $\theta\in A$, entonces la discrepancia entre el modelo para θ' y el modelo para θ'' es

$$d(\theta''|\theta')=\mathbb{E}_{\theta'}\left(-2\ln\left(f_X(x,\theta'')\right)\right)-\mathbb{E}_{\theta'}\left(-2\ln\left(f_X(x,\theta')\right)\right)=$$

$$\mathbb{E}_{\theta'}\left(-2\ln\left(f_X(x,\theta'')\right)\right)+2\ln\left(f_X(x,\theta')\right)$$

o sea que es el valor esperado suponiendo que el modelo correcto es θ' de la diferencia entre los logaritmos de las verosimilitudes del modelo para θ'' y del modelo para θ' multiplicadas por 2.

Observamos que $\frac{2(p+q+1)n}{n-p-q-2}\to 2(p+q+1)$ cuando $n\to+\infty$ motivo por el cual se resta $2(p+q+1)$ simplemente por ser una expresión más sencilla y similar (al menos cuando n es grande).

5.2 Criterio de información bayesiano (BIC)

Definición 43 *Dado un proceso ARMA(p,q) con p y q fijos, definimos*

$$BIC(p,q)=-2\ln\left(L\left(\widehat{\phi},\widehat{\theta},\widehat{\sigma}^2\right)\right)+\ln(n)(p+q+1)$$

donde $\left(\widehat{\phi},\widehat{\theta},\widehat{\sigma}^2\right)$ es el estimador máximo verosímil de $\left(\phi,\theta,\sigma^2\right)$.

Criterio de minimización del BIC
Ajusto un modelo ARMA$(\widehat{p},\widehat{q})$ siendo

$$(\widehat{p},\widehat{q})=\arg\min BIC(p,q).$$

Como se observa, las definiciones de AIC y BIC son similares. Mientras en el AIC el factor de penalización es dos veces la cantidad de parámetros del modelo, en el BIC el factor de penalización es $\ln(n)$ veces la cantidad de parámetros.

¿Por qué penalizar de esta forma? Se puede demostrar que minimizar BIC equivale a maximizar la probabilidad del modelo a posteriori de los datos.

6 Diagnóstico

Una vez que elegimos el modelo a ajustar, y estimamos sus parámetros es conveniente observar si el modelo ajusta bien a la serie de tiempo observada. Se le llama diagnóstico del modelo a una vez ajustado, hacerle una serie de chequeos que nos "aseguren" (aunque en realidad no aseguran nada, sólo sirven para detectar que ciertas falencias que podrían aparecer, nuestro modelo no las tenga)

que el modelo planteado y estimado puede ser razonable para la serie de datos observada. Para ello, comenzaremos dando la definición de lo que le llamaremos los "residuos" del modelo, que intuitivamente uno los piensa como la diferencia entre los valores realmente observados y los que predeciría el modelo (aunque en rigor son dichas diferencias estandarizadas de una manera particular) los cuales es esperable que se comporten como los ε_t, es decir como ruido blanco.

Definición 44 *Residuos del modelo*
Si $\{X_t\}_{t\in\mathbb{Z}}$ *es un proceso ARMA*(p,q) $\Phi(B)X_t = \Theta(B)\varepsilon_t$ *definimos*

$$\varepsilon_t(\phi,\theta) = \frac{X_t - \widehat{X}_t(\phi,\theta)}{\sqrt{r_{t-1}(\phi,\theta)}}$$

siendo $r_{t-1} = \sigma^2\mathbb{E}\left(X_t - \hat{X}_t\right)^2$ *que les llamaremos los residuos del modelo y el estimador de los mismos que llamaremos* $\widehat{\varepsilon}_t$ *y que definimos*

$$\widehat{\varepsilon}_t = \varepsilon_t\left(\widehat{\phi},\widehat{\theta}\right) = \frac{X_t - \widehat{X}_t\left(\widehat{\phi},\widehat{\theta}\right)}{\sqrt{r_{t-1}\left(\widehat{\phi},\widehat{\theta}\right)}}$$

siendo $\widehat{\phi},\widehat{\theta}$ *los estimadores máximo verosímiles de* ϕ,θ.

La motivación de la definición de los $\widehat{\varepsilon}_t$ como los residuos del modelo se justifican por el siguiente teorema.

Teorema 45 *Si* $\{X_t\}_{t\in\mathbb{Z}}$ *es un proceso ARMA*(p,q) $\Phi(B)X_t = \Theta(B)\varepsilon_t$ *causal e invertible, entonces*

$$\mathbb{E}\left(\left(\varepsilon_t(\phi,\theta) - \varepsilon_t\right)^2\right) \to 0 \text{ cuando } t \to +\infty.$$

El teorema anterior nos dice que si $t \to +\infty$ en un proceso ARMA (donde conocemos sus parámetros) entonces $\varepsilon_t(\phi,\theta)$ son aproximaciones de ε_t. En la práctica no conocemos ϕ ni θ, pero el teorema de consistencia del estimador máximo verosímil nos dice que $\left(\widehat{\phi},\widehat{\theta}\right)$ converge a (ϕ,θ) por lo que es razonable pensar que si sustituimos (ϕ,θ) por$\left(\widehat{\phi},\widehat{\theta}\right)$, tendremos que los valores de $\widehat{\varepsilon}_t = \varepsilon_t\left(\widehat{\phi},\widehat{\theta}\right)$ son aproximaciones de ε_t para valores grandes de t.

Dado que en los procesos ARMA los valores que observamos son los X_t y no los ε_t, pero tenemos una forma de estimar los ε_t mediante los residuos que nos arroja el modelo (los $\widehat{\varepsilon}_t$), para diagnosticar el buen ajuste (o no) del modelo, podemos chequear si los $\widehat{\varepsilon}_t$ tienen un comportamiento como el de un ruido blanco. Para ello podemos graficar los $\widehat{\varepsilon}_t$ la ACF de los $\widehat{\varepsilon}_t$ y realizar los test de hipótesis para datos i.i.d. vistos en la sección anterior.

6.1 Gráfico de los residuos

Los mismos deberían no mostrar tendencias, periodicidades ni ningún tipo de patrón o estructura. Deberían mostrar una varianza constante en el tiempo y la media de los mismos debería ser cercana a cero. La media de todos los residuos se puede calcular y es de esperar que resulte un valor muy pequeño, cuanto más cercano a cero mejor.

6.2 Gráfico de las ACF y PACF de los residuos

El gráfico no debería presentar ningún tipo de patrones, ni periodicidades de ningún tipo y es de esperar que la gran mayoría de los valores queden dentro de las bandas de confianza de $\pm 1.96/\sqrt{n}$.

6.3 Test de aleatoriedad de los residuos

Los gráficos muchas veces son engañosos por lo que no hay que fiarse sólo en ellos. Es bueno hacerlo para descartar al menos patrones que puedan observarse en los gráficos. De modo que se pueden aplicar los test vistos en la sección sobre test de aleatoriedad a las observaciones $\widehat{\varepsilon}_t$.

Observación 46 *Una vez sorteados todos los puntos anteriores, se puede investigar sobre la distribución de los errores, por lo que se pueden aplicar los clásicos test de normalidad como el de Shapiro–Wilks por ejemplo que en R simplemente es utilizar el comando shapiro.test(x) siendo x el vector de datos a investigar normalidad, en nuestro caso el vector de x sería el de los residuos $\widehat{\varepsilon}_t$. También se puede graficar la densidad estimada de los $\widehat{\varepsilon}_t$ con el comando plot(density(x)) siendo x el vector de datos, en nuestro caso los $\widehat{\varepsilon}_t$. También pueden ser aplicados otros tests de normalidad como por ejemplo el de Cramér-von Mises truncado.*

Nota. Se le llama test de normalidad a todo test de hipótesis tal que a partir de una muestra de datos i.i.d se plantea H_0 la distribución de los datos es normal, versus H_1 la distribución de los datos no es normal.

7 Modelación de series de tiempo mediante procesos ARMA

La modelación de una serie de tiempo mediante procesos ARMA y luego predicción de valores futuros se suele llevar a cabo en los siguientes pasos:

1. Desestacionalizo los datos.
   (a) En primer lugar, dado que los procesos ARMA son por definición procesos centrados, es conveniente restarles la media muestral de los datos, para que queden centrados.

(b) Conviene mirar el gráfico de los datos y de la acf para detectar y luego eliminar tanto tendencia como componente estacional.

(c) Luego de quitar la tendencia y componente estacional es importante chequear la estacionariedad de la nueva serie obtenida mediante los test de hipótesis sobre existencia de raíces unitarias, así como el de la varianza constante en el tiempo.

2. Identificación del modelo.
   Una vez realizado el punto anterior, elegimos en base a algún criterio de optimización los valores de p y q para luego ajustar un ARMA(p, q) por ejemplo minimizando AIC o BIC.

3. Estimación de los parámetros del modelo.
   Una vez determinado el modelo con el cual ajustar los datos, procedemos a la estimación de los parámetros.

4. Diagnóstico del modelo.
   Estimados los parámetros, tenemos el modelo estimado en su totalidad. Con el mismo, calculo los residuos generados por el modelo. Si el modelo es correcto dichos errores deberían comportarse como ruido blanco.

   (a) Grafico la ACF y la PACF y chequeo "visualmente" que el comportamiento sea como el de ruido blanco (o sea que no debe verse ningún patrón) en los mismos y deben ser todos pequeños, dentro de las bandas de confianza.

   (b) Realizo el test de hipótesis (Test de Ljung–Box o Box–Pierce) sobre si los datos se comportan como ruido blanco.

5. Predicción.
   Una vez superados todos los puntos anteriores, hemos encontrado un modelo que ajusta bien a los datos observados, por lo que podemos realizar predicciones de valores futuros (así como estimaciones de valores pasados no observados) con las fórmulas vistas en el teórico.

8 Modelos ARIMA

Los modelos ARIMA(p, d, q) son modelos que en principio pueden ser utilizados a series de datos con componente estacional, ya que incluyen el parámetro d que es un número natural e indica el número de veces que aplicamos el operador ∇ para desestacionalizar los datos. El modelo "desestacionaliza" los datos, aplicando d veces el operador ∇ a los datos para luego ajustarles un modelo ARMA(p, q). Cuando $d = 0$ queda directamente el modelo ARMA(p, q). Al igual que con p y q en el modelo ARMA, hay que elegir de antemano los valores de p, q y d.

Vamos a la definición.

Definición 47 *El proceso* $\{X_t\}_{t\in\mathbb{Z}}$ *estacionario y centrado se dice* $ARIMA(p,d,q)$ *si y sólo existe un proceso* $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ *ruido blanco tal que*

$$\Phi(B)(1-B)^d X_t = \Theta(B)\varepsilon_t.$$

Observación 48 *Observamos que si le llamamos* $Y_t = (1-B)^d X_t$ *nos queda* $\Phi(B)Y_t = \Theta(B)\varepsilon_t$ *es decir que* $\{X_t\}_{t\in\mathbb{Z}}$ *es* $ARIMA(p,d,q)$ *si y sólo si* $\{Y_t\}_{t\in\mathbb{Z}}$ *es* $ARMA(p,q)$.

Observación 49 $(1-B)^d = \nabla^d = \nabla\nabla...\nabla$ *d veces, por lo tanto a la serie observada de las* X_t *se les aplica el operador* ∇^d *con el cual se desestacionaliza la serie y quedan los* Y_t *a los cuales se los modela mediante un* $ARMA(p,q)$.

Observación 50 *Si bien los modelos* $ARIMA(p,d,q)$ *son muy utilizados en la práctica, tienen como punto negativo que al desestacionaliza la serie mediante el operador de diferenciación* ∇^d*, pero ya vimos que el operador de diferenciación quita tendencias de tipo polinómico cuando la tendencia puede ser de tipo sinusoidal (combinación lineal de senos y cosenos) que no puede ser removida con el operador* ∇.

Observación 51 *Por la observación anterior, dado un conjunto de datos con tendencia, es preferible tener libertad de elección en cuanto al método de desestacionalización para luego ajustar un ARMA, que "atarse" a desestacionalizar com lo hacen los modelos ARIMA.*

9 Modelos SARIMA

Como vimos, los modelos ARMA son procesos estacionarios por lo que para modelar con ellos, dada una serie de datos es necesario quitarles primero la tendencia y la componente estacional. Los ARIMA, mediante el operador de diferenciación permiten quitar la tendencia a los datos, dado que el operador $(1-B)^d$ cuando es aplicado a los datos, en principio le está quitando la tendencia y se supone que no tienen componente estacional para luego ajustar un ARMA a los mismos.

Los modelos SARIMA (seasonal ARIMA), van un paso más allá, porque los datos pueden tener tanto tendencia como componente estacional, el propio modelo se las quita, para luego ajustar un ARMA. Por tal motivo, los modelos SARIMA tendrán un parámetro s (seasonal) que en la práctica puede ser utilizado como $s = 12$ si los datos son mensuales, $s = 4$ si los datos son trimestrales, $s = 52$ si los datos son semanales por ejemplo.

Definición 52 $\{X_t\}_{t\in\mathbb{Z}}$ *se dice* $SARIMA(p,d,q)\times(P,D,Q)_s$ *de período s si y sólo si el proceso* $\{Y_t\}_{t\in\mathbb{Z}}$ *definido por* $Y_t = (1-B)^d(1-B^s)^D X_t$ *es un proceso causal ARMA que verifica*

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)\varepsilon_t$$

siendo $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ *ruido blanco* $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - ... - \phi_p z^p$, $\Phi(z) = 1 - \Phi_1 z - \Phi_2 z^2 - ... - \Phi_P z^P$, $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + ... + \theta_q z^q$ *y* $\theta(z) = 1 + \Theta_1 z + \Theta_2 z^2 + ... + \Theta_Q z^Q$ *y siendo* d *y* D *números naturales.*

El factor $(1-B)^d (1-B^s)^D$ es el factor de diferenciación que quita la tendencia. En la práctica suele considerarse $D = 0$ o $D = 1$ y rara vez se utiliza un valor de $D \geq 2$. En el caso en el cual $D = 0$, tenemos el factor de diferenciación $(1-B)^d$ como en un ARIMA. En el caso en el cual $D = 1$, observamos que el $B^s X_t = X_{t-s}$ por lo que si tenemos datos mensuales, en general se considera $s = 12$ para que cada dato en en instante t se lo vincule con el correspondiente a su período, es decir, si los datos son mensuales, si en un instante t estamos en el mes de marzo por ejemplo, entonces al aplicarle polinomios en B^s a los datos, estamos vinculando a todos los datos que tienen que ver con marzo.

Observación 53 *Una vez desestacionalizados los datos y quedarnos con la serie* $\{Y_t\}_{t\in\mathbb{Z}}$*, vemos que si le llamamos* $Y'_t = \Phi(B^s) Y_t$ *y* $\varepsilon'_t = \Theta(B^s)\varepsilon_t$*, la condición* $\phi(B)\Phi(B^s) Y_t = \theta(B)\Theta(B^s)\varepsilon_t$ *se traduce a* $\phi(B) Y'_t = \theta(B)\varepsilon'_t$ *lo que implica que* $\{Y'_t\}_{t\in\mathbb{Z}}$ *es un proceso* $ARMA(p,q)$*. Dicho en palabras, una vez que le aplicamos el polinomio* $\Phi(B^s)$ *a* Y_t *y el polinomio* $\Theta(B^s)$ *a* ε_t*, luego ajustamos un* $ARMA(p,q)$*.*

El procedimiento de estimación por máxima verosimilitud se puede llevar a cabo de manera similar al caso de los procesos ARMA, dado que la igualdad $\phi(B)\Phi(B^s) Y_t = \theta(B)\Theta(B^s)\varepsilon_t$ también puede leerse como $\phi^*(B) Y_t = \theta^*(B)\varepsilon_t$ siendo $\phi^*(B) = \phi(B)\Phi(B^s)$ un polinomio en B de grado $p + sP$ (donde algunos coeficientes son nulos) y $\theta^*(B) = \theta(B)\Theta(B^s)$ un polinomio de grado $q + sQ$.

Pudiendo calcularse la verosimilitud (y los estimadores máximo verosímiles), se definen de igual forma los criterios de minimización del AIC y el de minimización de BIC. Queda come ejercicio escribir las fórmulas que quedarían en estos casos.

Observación 54 *Cuando se tiene una serie de observaciones que llamamos* $x_1, x_2, ..., x_t$*, decidir modelar mediante un SARIMA, tiene la importante ventaja de que nos evita la estimación tanto de la tendencia como la de la componente estacional, porque el modelo en sí mismo quita la tendencia y la componente estacional y al aplicar la función predict en R, ya nos da predicciones del modelo para nuestra serie de tiempo observada. Por lo tanto modelar mediante SARIMA nos evita estimar tanto la tendencia como la componente estacional.*

Nota. Si tenemos un conjunto de valores observados de una serie de tiempo modelar por SARIMA, si bien tiene la ventaja comentada en la observación anterior, puede no ser la mejor opción, dado que desestacionaliza de un modo particular (como vimos mediante la aplicación del operador ∇) y quita la componente estacional de manera particular también (debido a la fórmula en sí del

modelo). De modo que a pesar de esa simplicidad que nos ofrece modelar mediante SARIMA , es muy probable que desestacionalizando mediante los métodos vistos en el capítulo anterior y luego ajustando un ARMA a los datos desestacionalizados, podamos obtener mejores resultados para la modelación de una serie de tiempo.
Nota. Los 5 pasos marcados en la sección 7 para modelar una serie de tiempo mediante modelos ARMA, se realiza igual si modelamos mediante ARIMA o SARIMA. La única salvedad es sobre el paso 1, ya que si voy a ajustar mediante ARIMA, la tendencia la quita el propio ARIMA por lo que en el paso 1, sólo habra que quitar la componente estacional. Si vamos a ajustar mediante SARIMA, entonces tanto la tendencia como la componente estacional las quita el propio modelo SARIMA.

10 La función arima en R

En R, la función "arima" realiza los pasos desde el 2 al 5 de la sección, los cuales además pueden ser aplicados tanto a los modelos ARIMA como a los SARIMA. Supongamos que x es un vector de datos, o un objeto .ts (serie de tiempo). La función arima(x,order=c(p,0,q)) me ajusta un ARMA(p,q) a los datos ingresados x o $x.ts$. Por defecto, la función arima(x,order=c(p,0,q)) ajusta un ARMA(p,q) no necesariamente centrado, si a los datos los centramos restándole la media, entonces debemos aplicar arima(x,order=c(p,0,q),include.mean=FALSE).
si le llamo aju=arima(x,order=c(p.0.q)), entonces
$y\$coef$ me dará la estimación de los $p+q+1$ coeficientes.
$y\$residuals$ me dará los residuos del modelo (los que deberían comportarse como un ruido blanco y utilizamos para el diagnóstico del modelo, paso 4 de la sección 7).
$y\$aic$ me dará el AIC del modelo.
$y\$loglik$ me da el logaritmo de la verosimilitud del modelo (con el cual podemos rápidamente calcular el BIC).
$y\$var.coef$ me da la estimación de la matriz de covarianzas de los estimadores. En particular en la diagonal encontramos la estimación de las varianzas de los estimadores.
Una vez que calculamos el objeto y mediante la función arima, podemos realizar predicciones mediante la función "predict". Por ejemplo
$predict(y, n.ahead = 5)$ predice los siguientes 3 valores a los valores observados de x. La función arima, también puede ser utilizada para ajustar un modelo ARIMA(p,d,q), simplemente con arima(x,order=c(p,d,q)), es decir $d = 0$ es el ARMA(p,q), todo lo anteriormente comentado funciona igual para los ARIMA. La función arima incluye además los modelos SARIMA$(p,d,q)(P,D,Q)_s$, mediante arima(x,order=c(p,d,q),seasonal=list(order=c(P,D,Q),period=s)). Si a los datos originales x les restamos la media muestral para centrarlo, en conveniente incluir en la función arima el argumento include.mean=FALSE) para que el modelo estime menor cantidad de parámetros.
Para ver el funcionamiento de la función arima en R, pueden inventarse una se-

rie de datos, por ejemplo x=rnorm(100) y aplicar todos estos comandos. Como siempre, mediante la sentencia "?arima" en la consola de R se obtiene toda la información junto con algún ejemplo y referencias sobre el funcionamiento de la función arima.

11 Ejercicios

1. Escribir programas en R que permitan llevar a cabo los test de aleatoriedad desde el 1.2.3 hasta el 1.2.6.

2. Probar las 4 propiedades de $\widehat{X}_{n+h}$ asumiendo que el proceso es centrado. Sugerencia para probar la propiedad 4. Plantear $\mathbb{E}\left(\widehat{X}_{n+h}^2\right) = \mathbb{E}\left(\left(\widehat{X}_{n+h} - X_{n+h} + X_{n+h}\right)^2\right)$ desarrollar el cuadrado y sustituir $\widehat{X}_{n+h}$ por $a_1X_1 + a_2X_2 + ... + a_nX_n$, y luego utilizar la propiedad 1 junto con la hipótesis $\gamma_n(h) \to 0$.

3. Dar la fórmula para el intervalo de predicción a un paso del valor X_{t+1} en un proceso ARMA causal.

4. Probar que para el proceso MA(q) la función de autocovarianza es $\gamma(h) = \sigma^2 \sum_{i=0}^{q-h} \theta_i\theta_{i+h}$ para $h \leq q$ y $\gamma(h) = 0$ para $h \geq q+1$.

5. La serie de datos en R llamada AirPassengers, contiene un ejemplo con datos mensuales en cientos de pasajeros de una aerolínea entre los años 1949 y 1960.

   (a) Graficar los datos y observar si es conveniente modelar con una descomposición aditiva o multiplicativa.

   En lo que sigue, trabajaremos con la serie ln(AirPassengers) es decir con el logaritmo neperiano de los datos (en R log(AirPassengers)).

   (a) Desestacionalizar los datos estimando la tendencia mediante una recta.

   (b) Aplicar los tests de estacionariedad para la componente aleatoria.

   (c) Ajustar el mejor modelo ARMA(p, q) para valores de p, q entre 0 y 2 de acuerdo al criterio AIC y luego de acuerdo al criterio BIC.

   (d) Con el modelo elegido realizar el diagnóstico del mismo. ¿Ajusta bien? ¿Los residuos tienen distribución normal? Aplicar para esto el test de Shapiro-Wilks y el test de Cramér-von Mises truncado.

   (e) Con el modelo obtenido en la parte anterior predecir los valores de pasajeros para el año 1961.

   (f) Para el modelo obtenido en la parte (c), graficar los valores que predice el modelo a un paso para los 12 meses de 1960, junto con sus intervalos de predicción al 95% y los valores realmente observados.

(g) Repetir lo pedido en el punto anterior, para las predicciones a 2 pasos y para 3 pasos.

6. Consideramos los datos de los logaritmos neperianos de AirPassengers desde 1949 hasta 1959.

   (a) Desestacionalizar los datos mediante tendencia lineal.

   (b) Para cada uno de los valores de p, q variando entre 0 y 2, ajustar el modelo ARMA(p, q) y predecir la cantidad de pasajeros para el año 1960.

   (c) Si nos interesa de los modelos anteriores el que mejores predicciones realice para el año 1960 ¿cuál elijo?

   (d) Para el modelo obtenido en la parte anterior, realizar el diagnóstico del mismo. ¿Ajusta bien? ¿Los residuos tienen distribución normal?

7. Repetir el ejercicio anterior pero desestacionalizando la serie mediante aplicación del operador ∇. ¿Predice mejor este modelo que el elegido en la parte (c) del ejercicio anterior?

8. Escribir las fórmulas para el cálculo de AIC y de BIC para los modelos SARIMA$(p, d, q) \times (P, D, Q)_s$.

9. Para la serie de datos de AirPassengers modelar el logaritmo neperiano, mediante SARIMA$(p, d, q) \times (P, D, Q)_s$ utilizando los valores de p y q hallados en la parte (c) del ejercicio 5, y considerando los casos $D = 0$, $D = 1$, y P y Q variando entre 0 y 2, elegir el que tenga menor AIC y predecir con este modelo los valores correspondientes a los 12 meses de 1961. Graficar toda la serie de datos junto con las predicciones para los 12 meses incluyendo sus intervalos de predicción al nivel 95%.